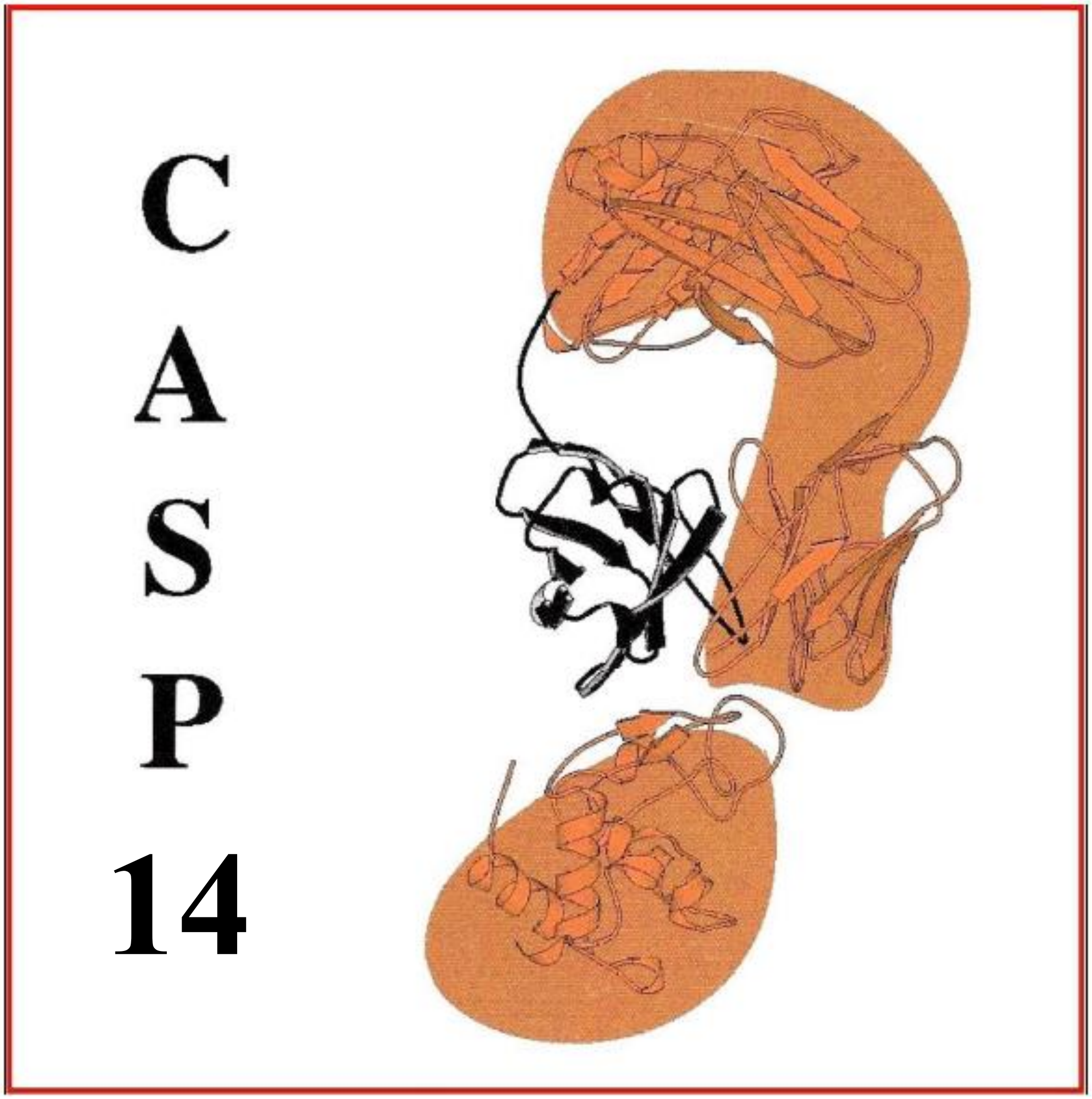
CRITICAL ASSESSMENT OF TECHNIQUES
FOR PROTEIN STRUCTURE PREDICTION



ABSTRACT BOOK

Fourteenth round
May-September 2020

TABLE OF CONTENTS

191227

iPhord: A protein structure prediction system based on deep learning

Dingyan Wang[1], Denghui Liu[2], Zhimeng Xu[2], Wenjun He[2], Chi Xu[2], Jianzhong He[2], Xinyuan Lin[2], Lei Zhang[2], Xiaopeng Zhang[2], Lingxi Xie[2], Qi Tian[2], Xi Cheng[1], Mingyue Zheng[1], Nan Qiao[2], Hualiang Jiang[1]

[1] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai, 201203, China, [2] Laboratory of Health Intelligence, Huawei Technologies Co., Ltd, Shenzhen, 518100, China.

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

The residue-residue contacts information is essential in the protein structure prediction task. A comprehensive protein structure prediction algorithm was developed by integrating contact-driven modeling, template-based modeling, and protein model assessment methods. The prediction of the residue-residue distances and orientations of contact-driven modeling was formulated into a data-driven, dense prediction problem and a deep CNN model equipped with attention modules was used for this purpose. Then, trRosetta was used to generate a set of candidate protein structures and re-ranked them using a 3D CNN model for quality assessment.

Methods

Both the template-based and template-free approaches were considered in our framework. Given a query protein sequence, CSI-BLAST[1] and SAM[2] were used to search against the protein templates database, if homologous structural templates were found, the pairwise query-template sequence alignments and the template structures would be fed into Modeller[3] to build the protein structural models. In parallel to the template-based modeling approach, the template-free approach was also proposed, which utilized a novel contact-driven deep neuron network to predict the protein structure from scratch: 1) Multiple alignments were generated for the query protein sequence by searching against different protein sequence databases (e.g. Uniclust30[4] and UniRef90[5]) through a combination of the HH-suite and HMMER programs[6]. 2) A novel deep learning model was used to take the multiple alignments as input and produce multiple contact predictions map, 3) the contact map was fed into trRosetta[7] to build structural models.

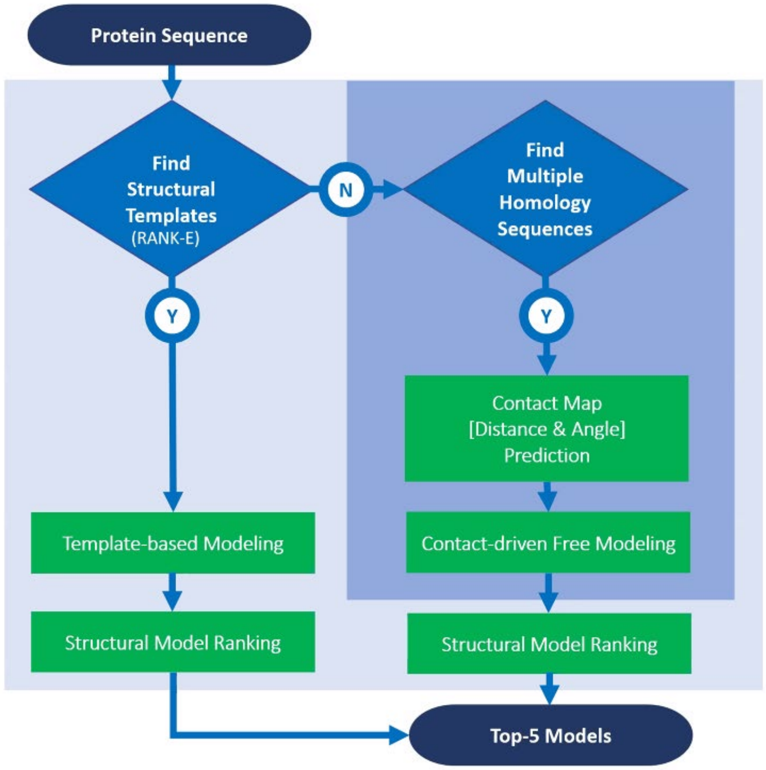

Protein Sequence
Find Structural Templates (RANK-E)
N
Find Multiple Homology Sequences
Y
Y
Contact Map [Distance & Angle] Prediction
Template-based Modeling
Contact-driven Free Modeling
Structural Model Ranking
Structural Model Ranking
Top-5 Models

Figure 1.Pipeline of iPhord

For contact map prediction, the input features derived from the multiple sequence alignment (MSA) result were cropped and then fed to the deep neural network to predict four objectives: one distance histogram and three angle histograms. A high-resolution segmentation network was used to maintain high-resolution representations by connecting high-to-low resolution convolutions in parallel. After the backbone, four independent paths were branched out with each path consisting of two convolutions to predict four objectives.

Both the template-based models and/or template-free models were added into a model pool for model ranking. To construct the ranking method, a 3D-grid box centered by the CA atom of each residue was used to extract local features. These features were fed to a 3DCNN model to predict the quality of the local structure. The 3DCNN model adopted an I3D-like[8] architecture based on Inception v1[9]. The global score was calculated by averaging the values of local scores. The top-5 scored conformations were further refined by CHARMM[10]. The refined structures and the original structures were mixed up and rescored, among which the final submitted model was selected by considering both CHARMM energy and the global score.

1. Biegert,A., Soding,J. (2009). Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A. 106, 3770-3775.
2. Hughey,R., Krogh,A. (1995). SAM: SEQUENCE ALIGNMENT AND MODELING SOFTWARE SYSTEM. University of California at Santa Cruz.

3. Webb,B., Sali,A. (2014). Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinforma. 47, 5.6.1−32.
4. Mirdita,M., Driesch,L., Galiez,C., Martin,M.J., Söding,J. & Steinegger,M. (2017). Uniclust databases of clustered and deeply annotated protein sequencesand alignments. Nucleic Acids Res. 45, D170-D176.
5. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. & Wu,C.H. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 31, 926-932.
6. Steinegger,M.; Meier,M.; Mirdita,M.; Vöhringer,H.; Haunsberger,S.J. & Söding,J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinform. 20(1), 473.
7. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A. 117, 1496-1503.
8. Joao,C., Andrew,Z. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. arXiv:1705.07750.
9. Russakovsky,O. et al. ImageNet large scale visual recognition challenge. (2015). Int J Comput Vis. 115, 211−252.
10. Brooks,B.R. et al. (2009). CHARMM: the biomolecular simulation program. J Comput Chem. 30, 1545−1614.

3DCNN_prof

Single model quality assessment using 3DCNN with profile-based features

Y. Takei[1,2], R. Sato[1] and T. Ishida[1]

[1] - Department of Computer Science, School of Computing, Tokyo Institute of Technology, Ookayama, Meguroku, Tokyo, Japan, [2] - AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL), National Institute of Advanced Industrial Science and Technology (AIST)

ishida@c.titech.ac.jp

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N*

We have previously developed a single model quality assessment method using 3DCNN1, but its performance was insufficient because it used only atom type features. Therefore, we added profile-based features, those are also used in other methods, to improve the performance.

Methods

The previous MQA method using 3DCNN[1] (Sato-3DCNN) utilized only 14 atom-type features. This study aims to improve performance adding profile-based features. As profile-based features, we used the Position-Specific Scoring Matrix (PSSM), the predicted secondary structure (predicted SS) and the predicted relative solvent accessibility (predicted RSA). We used exactly the same method as the Sato-3DCNN methods except for the addition of profile-based features. We used CASP 7-10 for training dataset, as in Sato-3DCNN.

We use PSSM as evolutionary information. PSSM was generated using PSI-BLAST[2] against Uniref90 database (downloaded April, 2019) with 2-iteration. PSSM was normalized and used as features. We also use predicted local structure as features, and actual local structure of a model structure is not used because it is considered to be observable using 3DCNN. Predicted SS and predicted RSA are used as predicted local structure. SS is predicted from the sequence profile using SSpro[3]. SSpro predict SS into 3 classes, therefore we use predicted SS in the form of 3 dimensional one-hot vector. RSA is predicted from the sequence profile using ACCpro20[3]. We normalized predicted RSA and used as a feature. PSSM, predicted SS and predicted RSA are all residue level features, but we assign them to all atoms that make up the residues.

Availability

This method is available on our website at http://www.cb.cs.titech.ac.jp/p3cmqa .

1. Sato, R. and Ishida, T. (2019). Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network. *Plos One*, **14**, e0221347.
2. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, ***25***, 3389–3402.
3. Magnan, C. N. and Baldi, P. (2014). SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics, **30**, 2592–2597.

A2I2Prot

Fusion of sequence embedding and sequence alignments for protein contact predictions

Tri Minh Nguyen[1], Hung Nguyen[1], Truyen Tran, Thin Nguyen [2]

[1] *- Deakin University, Australia*

minhtri@deakin.edu.au

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N*

Recent advances in deep convolution neural network and language modelling open up many opportunities to improve residue contacts prediction. In this work, we propose LateFuse methods which leverage the representation of sequence and alignment features. In our methods, the queried protein sequence is transformed into an embedding vector using self-supervised language modeling. The embedding vector is transformed into an embedding correlation map then combined with co-variance MSA features for contact prediction using the convolution neural network with residual architecture.

Methods

Embedding feature: In our method, the pre-trained language model provided by TAPE[1] is leveraged as an additional input feature. First, the embedding vector of a protein sequence is obtained using the pre-trained Transformer model provided by TAPE[1]. The Transformer model is trained using Pfam dataset which contains over 30 million protein sequences. The pre-train task is predicting the masked token in the input protein sequence. Then the pairwise product is performed on the embedding vector. This results in the correlation map between each position of the embedding vector. The pipeline is shown in Figure 1. The correlation map is expected to provide the correlation between tokens in protein sequence over the embedded space. Then the embedding correlation map is fed into the convolution neural network together MSA co-variance features.

The MSA- covariance feature is fed into residual blocks, which results in a 64╳L╳L feature map. Then the result feature maps are concatenated with the embedding correlation map along the channel axis. The result feature maps are fed into a batch normalization, ReLU activation, and 3╳ 3 convolution layer which transforms the feature maps dimension into 1╳L╳L. Finally, a sigmoid activation is used to predict the probability of residue-residue contact. The second approach is illustrated in Figure 2.

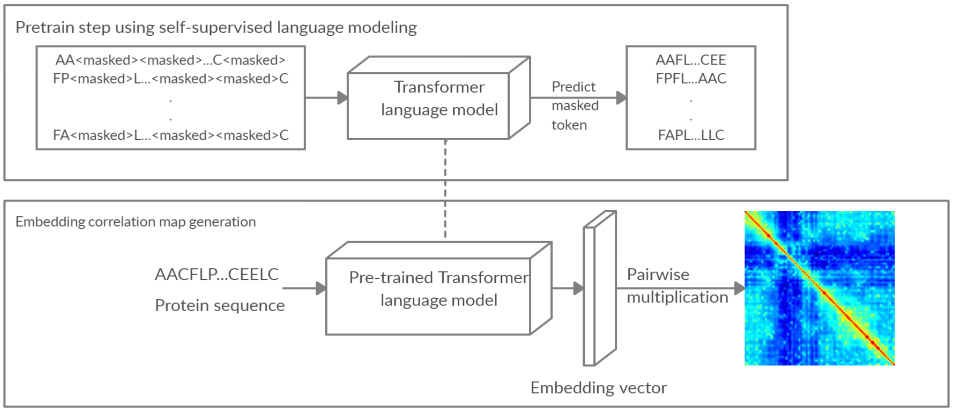

Pretrain step using self-supervised language modeling
AA<masked><masked>...C<masked>
FP<masked>L...<masked><masked>C
.
.
FA<masked>L...<masked><masked>C
Transformer language model
Predict masked token
AAFL...CEE
FPFL...AAC
.
.
FAPL...LLC
Embedding correlation map generation
AACFLP...CEELC
Protein sequence
Pre-trained Transformer language model
Pairwise multiplication
Embedding vector

Figure 1: The process of constructing the embedding correlation map from a protein sequence. First, the embedding vector with the dimension of L×768 of the protein input sequence is obtained using the pre-trained Transformer language model. L is the protein sequence length. Then the embedding vector is pairwise multiplied with its transposed vector to form the embedding correlation map with size L×L.

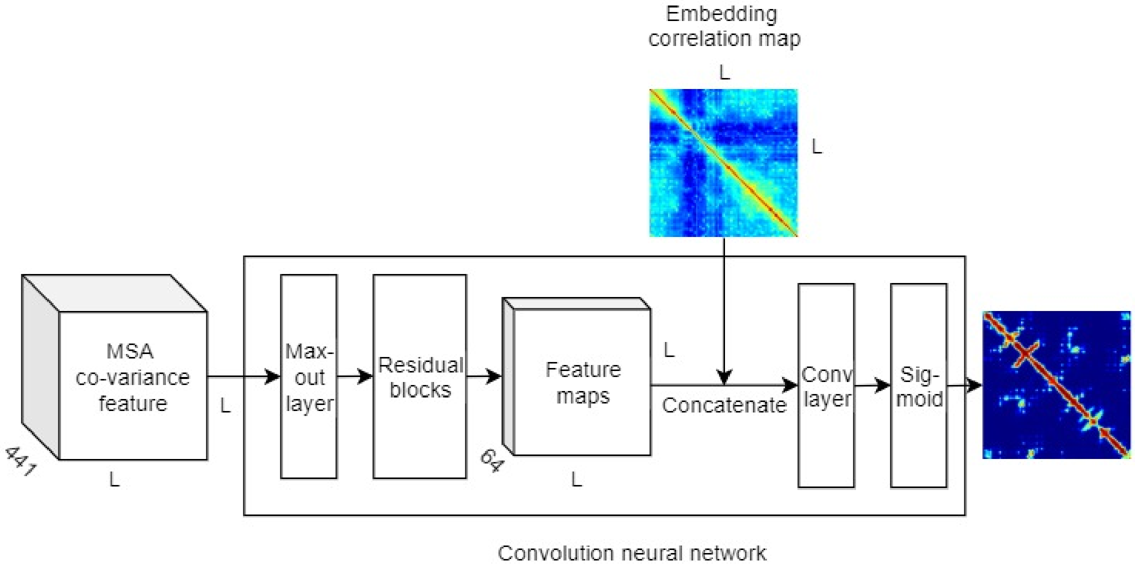


Figure 2: The model architecture of LateFuse.

1. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel,P., Song, Y.S.: Evaluating Protein Transfer Learning with TAPE. In: Advances inNeural Information Processing Systems (2019).

AILON

Protein tertiary structure prediction driven by deep neural network and cmFinder from its amino acid sequence

Wonjun Lee[1], Keyongtak Han[1], Kyungmin Cho[1] and Hyo Je Cho[1]
[1] - *AILON Inc., 148 Sagimakgol-ro, Jungwon-gu, Seongnam-si, Gyeonggi-do, 13207, Republic of Korea*
hyojec@gmail.com

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.UniRef30; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

We participated in CASP14 tertiary structure prediction as human group "AILON", with our newly developed cmFinder[1], which uses artificial intelligence-based protein structure prediction methodology via contact map-based searching for similar protein folds. The proposed cmFinder algorithm focuses on template-based modeling (TBM). Due to the difficulty in searching similar fold templates, however, the accuracy of TBM decrease when the target sequence has no fold identity within the known protein fold database. We applied another structure prediction method that resembles DISTFOLD[2], which is used for free-modeling (FM), given predicted secondary structure and distance map information.

Methods

Our tertiary structure prediction consists of the following steps (Figure 1).

Step 1: sequence analysis: Given a target protein sequence, our method first used HHblits to produce a multiple sequence alignment (MSA), and XtalPred to search homolog's templates, structural disordered regions and conserved domains.

Step 2: secondary structure and contact map prediction Secondary structure predicted by PSIPRED and SPIDER3. Contact maps (where a contact is a distance of 8 Å or less between C-beta atoms) predicted by DeepECA, SPOT-Contact, DNCON2 and MapPred.

Step 3: cmFinder or dmModeler: cmFinder, which aims to construct structural models by searching same fold in the contact map database, and dmModeler, which aims to generate structural models by CNS suite with distance restrains.

Step 4: structural model building and refinement. If the target protein has the same fold as the templates' structure, selected on the basis of cmFinder results, it can be easily built and refined with the COOT program. If not, models built by dmModeler need to be validated by structural biology and protein functional domain knowledge.

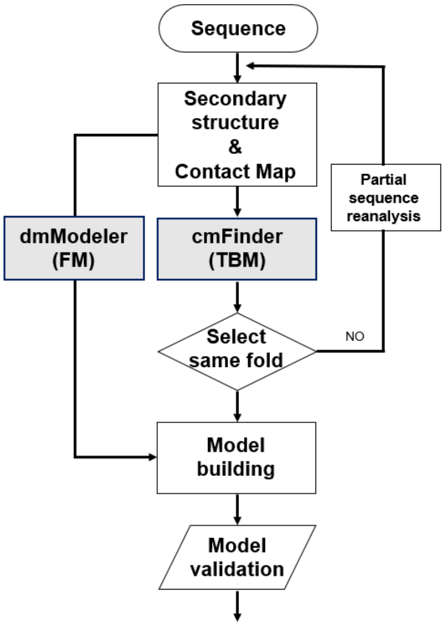


Figure 1. AILON pipeline for CASP14

Results

At the present date, several TS targets are released on PDB. Our submitted prediction models compared favorably with released structures. We calculated the TM-score and RMSD of the best of five models for the released structures (Table 1 and Figure 2).

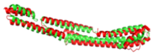
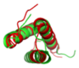
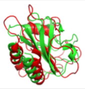
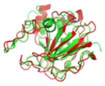
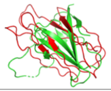
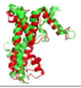
TS target ID	AILON predicted model	PDB ID released structure	TM score	RMSD (Å)	structural alignment image
T1024	model01	6T1Z	0.8407	2.715	
T1026	model02	6S44	0.6957	1.567	
T1030	model01	6POO	0.5607	5.401	
T1046S1	model03	6PX4_A	0.6918	1.847	
T1046S2	model02	6PX4_B	0.6235	3.752	
T1049	model02	6Y4F	0.6733	2.434	
T1056	model01	6YJ1	0.578	2.572	
T1064	model04	7JT1	0.2351	10.725	
T1099	model02	6YGH	0.4935	6.190	

Table 1. Structural comparision of AILON predictied structural model and released PDB structure.

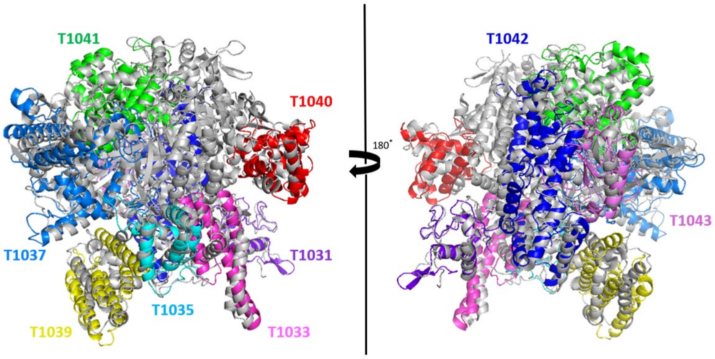


Figure 2. RNAP substructures' prediction results. Each color shown form TS1031 to TS1043 targets.

Availability

cmFinder is being prepared for publication. Upon publication, its standalone executable version would be accessible as an appended material, and the source code will be available soon.

Acknowledgements

We gratefully acknowledge the use of several tools and datasets (HHBlits, XtalPred, PSIPRED, SPIDER3, DeepECA, SPOT-Contact, DNCON2, MapPred, TM-score)

1. Cho K, Cho HJ, Lee W (2020) A protein tertiary structure prediction method using adjacent map images between amino acids. *KR 10-2020-0065123*
2. Adhikari B, Cheng J. (2018) CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC Bioinformatics. 25;19(1):22* (https://github.com/badriadhikari/DISTFOLD)

AIR: An artificial intelligence-based protocol for protein structure refinement using multi-objective particle swarm optimization

Di Wang[1,2], Cheng-Peng Zhou[1,2], and Hong-Bin-Shen[1,2]
1 - Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 2 -Key Laboratory of System, Control and Information Processing, Ministry of Education of China, Shanghai 200240, China
hbshen@sjtu.edu.cn

We have tested our updated AIR[1] to use multiple energy functions as multi-objectives to refine the protein structure model. For each target, we use 3 initial models as the input particles to AIR. We collect 20 refined models from GalaxyRefine2[2] and 10 from refineD[3]. Then, the top two models of the 20 prediction models ranked by Pcons[4] software, as well as the initial model provided by the CASP server are used as the original models. From these three initial templates, we give each of them some random perturbation, resulting in a total of 50 different particles.

In the refinement iterations, by analyzing the challenge cases in CASP13, we adopt a wake-up mechanism to modify the definition of dominance so that the high-quality models can be saved to the Pareto set during the refinement iterations. Furthermore, we also tried a decomposition-based method, which decomposes the multi-objective optimization into a set of subproblems and optimizes them in a collaborative manner. Our local experiments show that it is a promising way in dealing with complicated Pareto set shapes. After enough iteration times, we clustered the structures from the pareto set using TM-score[5] program, and the top 5 models will be chosen from the Pareto set using clustering and knee[6] algorithm.

1. Wang, D., Geng, L., Zhao, Y. J., Yang, Y., Huang, Y., Zhang, Y., & Shen, H. B. (2020). Artificial intelligence-based multi-objective optimization protocol for protein structure refinement. Bioinformatics, 36(2), 437-448.
2. Lee, G. R., Won, J., Heo, L., & Seok, C. (2019). GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure. Nucleic Acids Research, 47(W1), W451-W455.
3. Bhattacharya, D. (2019). refineD: improved protein structure refinement using machine learning based restrained relaxation. Bioinformatics, 35(18), 3320-3328.
4. Wallner, B., & Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science, 15(4), 900-913.
5. Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and bioinformatics, 57(4), 702-710.
6. Branke, J., Deb, K., Dierolf, H., & Osswald, M. (2004, September). Finding knees in multi-objective optimization. In International conference on parallel problem solving from nature (pp. 722-731). Springer, Berlin, Heidelberg.

AlphaFold2

High Accuracy Protein Structure Prediction Using Deep Learning

John Jumper[1*†], Richard Evans[1*], Alexander Pritzel[1*], Tim Green[1*], Michael Figurnov[1*], Kathryn Tunyasuvunakool[1*], Olaf Ronneberger[1*], Russ Bates[1*], Augustin Žídek[1*], Alex Bridgland[1*], Clemens Meyer[1*], Simon A A Kohl[1*], Anna Potapenko[1*], Andrew J Ballard[1*], Andrew Cowie[1*], Bernardino Romera-Paredes[1*], Stanislav Nikolov[1*], Rishub Jain[1*], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Martin Steinegger[2], Michalina Pacholska[1], David Silver[1], Oriol Vinyals[1], Andrew W Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1], Demis Hassabis[1*†]

[1]*DeepMind, London, UK,* [2]*Seoul National University, South Korea*

* Equal contribution

† Corresponding authors: John Jumper (jumper@google.com), Demis Hassabis (dhcontact@google.com)

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

In the CASP14 experiment, we deployed AlphaFold 2. This new system uses a different deep learning method than CASP13 AlphaFold, and it produces much more accurate protein structures and estimates of model accuracy. The training data for the system is publicly available and similar to that used for CASP13 AlphaFold.

Methods

Input data: Given a query sequence, we obtain related sequences by searching three databases: UniRef90[1], BFD[2,3], and MGnify clusters[4]. JackHMMER[5] is used to search UniRef90 and MGnify clusters while HHblits[6,7] is used to search BFD. Additionally, potential templates are found using HHsearch[6,7] on the PDB70 clustering of the Protein Data Bank[8] provided by the Söding lab. No server predictions are used.

Folding: The input sequence, multiple sequence alignment, and template hits are used as inputs for the deep learning-based method that produces a variety of predictions including distances, torsions, atom coordinates, and estimates of the per-residue value of the Cα-lDDT[9].

We found that existing deep-learning architectures overly favor sequence-local interactions and do not sufficiently account for global structural constraints. To remedy this, we have developed a novel, attention-based deep learning architecture to achieve self-consistent structure prediction. We also allow the deep learning algorithm to attend arbitrarily over the full MSA instead of using pairwise co-evolution features like mutual information or pseudolikelihood, allowing the algorithm to ignore irrelevant sequences as well as to extract much richer information from the MSA. The resulting algorithm shows vastly improved performance, especially for shallow MSA depths, when compared to traditional co-evolution methods.

The predicted structures are ranked according to the predicted value of the Cα-lDDT. All deep learning models were trained using publicly-available structures in the PDB.

Refinement: Each prediction is relaxed using restrained gradient descent on the Amber ff99SB force field[10] using OpenMM[11]. Empirically, the RMSD of the structure change during relaxation is small.

Manual interventions: Domains arising from H1044: We first folded four subsequences individually using crops of the full chain MSAs then re-folded the full chain using these structures as templates. The submitted domains were cropped out of this full-chain folding. We improved our models during the competition so that we can now fold 2000+ amino acid chains accurately without manual intervention.

T1064: Five additional sequences were added to the MSA using a manual search with NCBI's Protein BLAST tool[12] and a wider range of models was used before ranking.

Additional targets: For several targets, the five models produced were very similar, and we sometimes used older or differently-trained models in positions 3, 4, or 5 to increase diversity. E.g. on target T1024, templates were clustered into 3 classes to provide more diverse predictions in the last three positions.

Acknowledgements

We would like to thank Alban Rrustemi, Albert Gu, Alexey Guseynov, Charlie Beattie, Craig Donner, Emilio Parisotto, Erich Elsen, Florentina Popovici, Hector Maclean, Jacob Menick, James Kirkpatrick, Jeff Stanway, Karen Simonyan, Laurent Sifre, Sam Blackwell, Shaobo Hou, Stephan Gouws, Steven Wheelwright, Tamas Berghammer, and Zachary Wu for their contributions; Martin Steinegger and Milot Mirdita for their help with datasets; Ellen Clancy, Mildred Piovesan-Forster, and Alexander Nelson for their help managing the project; and our colleagues at DeepMind for their support.

We would also like to thank the CASP14 organisers and the experimentalists whose structures enabled the assessment.

1. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506-D515.
2. Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, **16**(7), 603-606.
3. Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, **9**(1), 1-8.
4. Mitchell, A. L. et al. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, **48**(D1), D570-D578.
5. Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**(1), 431.
6. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, **9**(2), 173-175.

7. Steinegger, M. et al. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**(1), 1-15.
8. wwPDB consortium. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* **47**: D520-D528.
9. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722-2728.
10. Hornak, V. et al. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, **65**(3), 712-725.
11. Eastman, P. et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, **13**(7), e1005659.
12. Altschul, S. F. et al. (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**:3389-3402.

AngleQA: protein single-model quality assessment based on torsion angles

Boling Wang[1], Jianyi Yang[1] and Jianzhao Gao[1]

[1] *School of mathematical sciences, Nankai University, Tianjin, 300071, China*
gaojz@nankai.edu.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

Protein structure quality assessment (QA) plays important role in analysis of protein structure. It is one of open problems in structural bioinformatics. Here we proposed a single-model quality assessment model angleQA. The proposed method was build using the full connected neural network technology and the model was optimized both TM-scores and GDT-TS scores. It can score the global quality of input model.

Methods

The proposed method was training on the all single domain targets from CASP7-10, and all targets from CASP11 and CASP12. The model was validated on all targets from the stage 2 of CASP13.

The new method angleQA, fused the following features, (1) The energy scores from dDFIRE[1], RWplus[2] and Sbord[3]. (2) The similarity scores of solvent accessibility and secondary structure and torsion angles. (3)The energy function scores based on the torsion angles and solvent accessibility predicted from SPOT1D[4]. (4)The evolutionary scores from the positive specific score matrix and outputs from HHblits[5]. (5) contact score, gap score and align length from mapAlign[6].

Results

We postpone the assessment of the approach until the official release of CASP14 results.

Availability

The proposed method, angleQA is depend on SPOT1D framework. The angleQA package is available at www.biomath.cn

1. Yang, Y., & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function, and Bioinformatics*, 72(2), 793-803.
2. Zhang, J., & Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one*, 5(10), e15386.
3. Karasikov, M., Pagès, G., & Grudinin, S. (2019). Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics*, 35(16), 2801-2808.
4. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14), 2403-2410.

5. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2), 173-175.
6. Yang, W., & Wang, L. (2020, December). mapAlign: An Efficient Approach for Mapping and Aligning Long Reads to Reference Genomes. *In International Symposium on Bioinformatics Research and Applications* (pp. 105-118). Springer, Cham.
7. Yang, Y., & Zhou, Y. (2008). Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins: Structure, Function, and Bioinformatics*, 72(2), 793-803.
8. Zhang, J., & Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PloS one*, 5(10), e15386.
9. Karasikov, M., Pagès, G., & Grudinin, S. (2019). Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. *Bioinformatics*, 35(16), 2801-2808.
10. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14), 2403-2410.
11. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 9(2), 173-175.
12. Yang, W., & Wang, L. (2020, December). mapAlign: An Efficient Approach for Mapping and Aligning Long Reads to Reference Genomes. *In International Symposium on Bioinformatics Research and Applications* (pp. 105-118). Springer, Cham.

AP_1

AP_1 structure predictions in CASP14

Hyung-Rae Kim[1]

[1]*School of Basic Sciences, Hannam University, 70 Hannamro, Daedeok-Gu Daejeon 34430, South Korea*

hr_kim@hnu.kr

The goal of AP_1 is single chain protein structure scoring and combines our refinement protocol. AP_1 employs several characteristics, such as database search and structure retrieval without calculating pair-wise potentials and without building a fixed form potential.

Methods

The goal of AP_1 is to accurately score not only the topology of a protein structure, but also the side-chain positions of the high-accuracy template-based models.

Our structure prediction pipeline consists of the following steps:

1. Five of the best models were picked using AP_1 from all submitted server models of CASP14.
2. Five of the best models were picked and used as the seed model for our refinement protocol.
3. Subsequently, five generated models were added to the seed models.
4. We applied AP_1 again to the above candidate models and selected the five best models to submit.

In CASP14, we submitted 390 models for 78 TS regular targets.

Availability

A new AP_1 is being prepared. Its standalone executable version would be accessible as an appended material, once published.

BAKER-experimental (Assembly)

Protein oligomer structure predictions guided by predicted inter-chain contacts

Minkyung Baek[1], Ivan Anishchenko[1], Hahnbeom Park[1], Ian Humphrey[1] and David Baker[1,2]
[1] - *Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA,*
[2] - *Howard Hughes Medical Institute, University of Washington, WA, USA*
dabaker@uw.edu

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:Y.3,9; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

In this CASP, we used three different approaches to generate oligomer structures based on predicted oligomer interactions from MSAs and templates. We took a template-based approach when targets had proper oligomer templates detected by HHsearch. If no oligomer templates were found, we took either *ab initio* docking or simultaneous fold-and-dock approach depending on the quality of predicted homo-oligomer contacts from MSAs.

Methods

Co-evolution based oligomeric inter-residue interaction prediction. For each subunit, MSAs were generated by running HHblits[1] against UniRef/Uniclust database and metagenomic sequence database. For homo-oligomer targets, both GREMLIN[2] and the in-house deep learning based homo-oligomer contact prediction method were used to predict inter-chain contacts from given MSAs. For hetero-oligomer targets, inter-chain contacts were predicted with GREMLIN and trRosetta[3] based on the paired alignments[4]. Predicted inter-chain contacts were utilized as restraint energies to guide overall sampling and to pick final models.

***Template-based approach*:** HHsearch[5] and TM-align[6] were used to detect oligomer templates based on not only sequence similarity but also structure similarity to the subunit structures predicted by trRosetta. Up to five oligomer templates were selected according to the HHsearch ranks among the hits having structures similar to the given subunit structure (TM-score > 0.5) and in the given oligomer state. Rosetta hybridization protocol[7] was used to build oligomer structures based on the given subunit structures and the detected templates. In the hybridization protocol, unreliable local regions were rebuilt by inserting fragments and recombining the secondary structure segments between templates. The overall structures were further refined using FastRelax in Rosetta. The inter-chain restraints from predicted contacts were applied during the model building process as well as the intra-chain restraints driven from trRosetta. The entire process was symmetry aware for homo-oligomer targets. Total 500 structures were sampled by running independent template-based modeling protocol, and 5 models having lowest Rosetta energy with inter-chain contact restraints were selected after clustering.

***Docking-based approach*:** When there were no proper oligomer templates and no predicted contacts with high confidence for the target protein, oligomer structures were predicted using *ab initio* docking with subunit structures predicted by trRosetta. SymDock[8] was employed to predict symmetric homo-oligomer structures, while ZDOCK[9] and RosettaDock[10] were used for hetero-oligomer targets. Top 50 models after clustering were further refined by FastRelax in Rosetta, and 5 models having lowest Rosetta energy were selected after clustering.

***Simultaneous fold-and-dock approach with direct gradient-based optimization*:** Small local inaccuracy at the interface can hinder generating correct oligomer structures with *ab initio* docking. Moreover, as proteins interact with other proteins, their lowest free-energy backbone conformations typically shift in response to their partners, and it is really hard to predict using typical docking after folding approach. To overcome these limitations, we developed simultaneous fold-and-dock approaches consisting of two stages of sampling. In the first low-resolution stage, the oligomer conformation is sampled by alternating gradient-based folding and low-resolution docking starting from a conformation with randomly assigned backbone torsion angles. During the gradient-based folding, the conformation is minimized against Rosetta centroid energy function with intra-chain restraints derived from trRosetta and inter-chain restraints derived from predicted contacts. During the low-resolution docking, Motif Dock Score[10] with inter-chain restraints is used to optimize orientation between subunits. In the second stage, side chains are built into the backbone conformations, and small rigid-body perturbations followed by all-atom relaxations are performed to further refine overall complex structures. For homo-oligomer targets, symmetry is considered during the entire process. We took this simultaneous fold-and-dock approach when there were no proper oligomer templates, but inter-chain contacts were predicted with high confidence based on MSAs.

1. Remmert,M., Biegert,A. & Hauser,A. (2012). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9, 173-175.
2. Kamisetty,H., Ovchinnikov,S. & Baker,D. (2013). Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. U.S.A. 110, 16674-16679.
3. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. 117, 1496-1503.
4. Ovchinnikov,S., Kamisetty,H. & Baker,D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife 3, e02030.
5. Sӧding,J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951–960.
6. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309.
7. Song,Y., DiMaio,F., Wang,R.Y.R., Kim,D.E., Miles,C., Brunette,T.J., Thompson,J. & Baker,D. (2013). High-resolution comparative modeling with RosettaCM. Structure 21, 1735-1742.
8. André,I., Bradley,P., Wang,C. & Baker,D. (2007). Prediction of the structure of symmetrical protein assemblies. Proc. Natl. Acad. Sci. U.S.A. 104, 17656-17661.
9. Chen,R., Li,L. & Weng,Z. (2003). ZDOCK: An Initial-stage Protein Docking Algorithm. Proteins 52, 80-87.
10. Marze,N.A., Burman,R.S.S., Sheffler,W. & Gray,J.J. (2018). Efficient flexible backbone protein–protein docking for challenging targets. Bioinformatics 34, 3461-3469.

BAKER-ROSETTASERVER, BAKER (TS)

Protein structure prediction guided by predicted inter-residue geometries

Ivan Anishchenko[1], Minkyung Baek[1], Hahnbeom Park[1], Justas Dauparas[1], Naozumi Hiranuma[1], Sanaa Mansoor[1], Ian Humphrey[1], and David Baker[1,2]

[1] - Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA,
[2] - Howard Hughes Medical Institute
dabaker@uw.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:Y; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y*

For this CASP round, we developed an automated modeling pipeline where the primary driving force for model building are residue-residue geometry constraints derived from coevolutionary data as well as from top scoring structural templates by deep learning. Human BAKER group TS submissions were additionally refined using the protocol outlined in 'BAKER, BAKER-experimental (Refinement)' abstract.

Methods

Sequence and template searches: Multiple sequence alignments (MSAs) for the target sequences were generated by several rounds of iterative *hhblits* search against the Uniclust30 database (Jan 2020 version) with gradually relaxed e-value cutoffs as outlined in[1]. For human BAKER group predictions, the resulting MSAs were manually inspected to fine-tune the e-value and coverage cutoffs and enriched with metagenomic sequences[2]. In either case, the generated MSAs were then used to search for putative structural templates in the PDB by *hhsearch*.

Predicting residue-residue geometries and model building: To predict residue-residue geometries, we employed two variants of the *trRosetta* network with the one relying on sequence data only (original *trRosetta*[1]) and the other additionally using the information on the top 25 putative structural homologs as identified by *hhsearch* (modified *trRosetta*[3]). Two corresponding pools of structure models were then generated using the *trRosetta* folding protocol. To recombine the two sets of models, we developed a new network, called *trRefine*, which takes the outputs of the above two networks as well as 2D-projected top scoring structure models from both pools and their residue-pairwise C_β-C_β distance errors predicted by DeepAccNet-MSA[4] as inputs and generates the refined predictions for residue-residue geometries. Based on these *trRefine* predictions, the new pool of structure models was created by the *trRosetta* folding protocol.

Model refinement and selection: For BAKER-ROSETTASERVER, the *trRefine*-derived models were re-scored using DeepAccNet-MSA (see 'BAKER-ROSETTASERVER, BAKER-experimental (EMA)' abstract for details), and three best scoring ones were picked for submissions 1-3. Submissions 4 and 5 were the top models from the original (MSA only) and modified (MSA+templates) *trRosetta* networks respectively. For human TS predictions, *trRefine* models were additionally refined using the standard Rosetta all-atom refinement protocol[4] complemented by DeepAccNet-MSA predictions (see 'BAKER, BAKER-experimental (Refinement)' abstract for details).

13. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations, Proc. Natl. Acad. Sci. U. S. A. 117, 1496-1503.
14. Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., Yang, J. (2020). Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics 36, 41-48.
15. Farrell, D.P., Anishchenko, I., Shakeel, S., Lauko, A., Passmore, L.A., Baker, D., DiMaio, F. (2020). Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. IUCrJ. 7, 881-892 .
16. Hiranuma, N.,Park, H., Anishchanka, I., Baek, M., Baker, D. (2020). Improved protein structure refinement guided by deep learning based accuracy estimation, bioRxiv, doi: https://doi.org/10.1101/2020.07.17.209643.

BAKER-ROSETTASERVER, BAKER-experimental (EMA)

Estimation of Model Quality via Deep Residual Learning

Naozumi Hiranuma[1,2], Minkyung Baek[1], Hahnbeom Park[1], Ivan Anishchenko[1], and David Baker[1,3]

[1] - Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA,
[2] - Paul G. Allen School of Computer Science & Engineering, University of Washington, WA, USA,
[3] - Howard Hughes Medical Institute
dabaker@uw.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Deep learning (DL) has been successfully used in numerous methods that aim to estimate accuracy of modeled protein structures. Recently, we developed a novel deep learning framework (DeepAccNet[1]) that estimates per-residue accuracy (C_β local distance difference test; C_β l-DDT) and residue-residue distance signed error (histogram of error; estogram) of modeled protein structures. In this CASP, we applied DeepAccNet and the variant of DeepAccNet (named DeepAccNet-MSA) to the EMA category. The predictions of DeepAccNet were submitted for “BAKER-experimental” group while those of DeepAccNet-MSA were submitted for “BAKER-ROSETTASERVER” group.

Methods

We sought to develop model accuracy predictors that provide both global and local information. We developed network architectures that simultaneously make the following three types of predictions given a protein structure model: local measures of structure accuracy measured by per residue C_β local distance difference test (l-DDT)[2] scores, a native C_β contact map thresholded at 15 Å (referred to as mask), and per residue-pair distributions of signed C_β-C_β distance error against corresponding native structures (referred to as estograms; histogram of errors); C_α is taken for GLY. Rather than predicting single error values for each pair of positions, we instead predict histograms of errors (analogous to the distance histograms employed in the structure prediction networks of[3,4]), which provide more detailed information about the distributions of possible structures and better represent the uncertainties inherent to error prediction.

DeepAccNet: The predictions of DeepAccNet are based on 1D, 2D, and 3D features that reflect accuracy at different levels. Defects in high resolution atomic packing are captured by 3D convolution operations performed on 3D atomic grids around each residue defined in a rotationally invariant local frame, similar to the Ornate method[5]. 2D features are defined for all residue pairs, and they include Rosetta inter-residue interaction terms, which further report on the details of the interatomic interactions, while residue-residue distance and angular orientation features provide lower resolution structural information. At the 1D per residue level, the features are the amino acid sequence, backbone torsion angles, and the Rosetta intra-residue energy terms. The network architecture is based on the ResNet architecture[1].

DeepAccNet-MSA: We also trained a predictor that additionally takes in predictions from trRosetta, which give indirect access to the information from multiple sequence alignment. The trRosettta predictions are included as additional 2D features.

Both networks were trained on approximately one million alternative structures (“decoys") with model quality ranging from 50% to 90% in GDT-TS (global distance test - tertiary structure)[6] generated by homology modeling[7], trRosetta[3], and native structure perturbation.

Availability

The code is available through github at https://github.com/hiranumn/DeepAccNet.

1. Hiranuma, N., Park, H., Anishchanka, I., Baek, M., Baker, D. (2020). Improved protein structure refinement guided by deep learning based accuracy estimation. doi:10.1101/2020.07.17.209643.
2. Mariani. V., Biasini. M., Barbato. A., Schwede., T. (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722−2728.
3. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117, 1496−1503.
4. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L,. Green, T., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature 577, 706−710.
5. Pagѐs G, Charmettant B, Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks. (2019). Bioinformatics 35, 3313−3319.
6. Zemla A. LGA: A method for finding 3D similarities in protein structures. (2003). Nucleic Acids Res 31, 3370−3374.
7. Song Y, DiMaio F, Wang RY-R, Kim D, Miles C, Brunette T, et al. (2013). High-resolution comparative modeling with RosettaCM. Structure 21, 1735−1742.

BAKER, BAKER-experimental (Refinement)

Model refinement guided by an interplay between Deep-learning and Rosetta

Hahnbeom Park[1], Minkyung Baek[1], Naozumi Hiranuma[1], Ivan Anishchenko[1], Sanaa Mansoor[1], Justas Dauparas[1] and David Baker[1,2]
[1] - *Department of Biochemistry and Institute for Protein Design, University of Washington, WA, USA;*
[2] - *Howard Hughes Medical Institute*
dabaker@uw.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:Y.3,9; Cont:N; Dist:N; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Deep learning (DL) has been successfully applied in the last CASP to infer residue-pair distances from sequence co-evolutionary information to guide *de novo* protein structure predictions. In this CASP, we sought to apply DL to protein refinement problems by guiding search using predicted errors in model structures.

Methods

We developed a deep learning framework (DeepAccNet) that estimates per-residue accuracy in l-DDT and residue-residue distance signed error in protein models and uses these predictions to guide Rosetta protein structure refinement[1]. The network uses 3D convolutions to evaluate local atomic environments followed by 2D convolutions to provide their global contexts, and outperforms other methods that similarly predict the accuracy of protein structure models without template or evolutionary information (details can be found in BAKER-EMA abstract). We made two refinement protocols integrating variants of DeepAccNet.

All-atom protocol : We integrated “DeepAccNet-MSA” into our standard Rosetta refinement protocol[2] and used it i) for the final stage refinement of trRosetta[3] models in human regular category predictions as well as ii) for refinement category predictions. DeepAccNet-MSA is a variant that takes the trRosetta network prediction as an additional input for MSA information. In both categories the models resulting from the protocol are submitted for the “BAKER” group. DeepAccNet-MSA is incorporated into every iteration in the refinement protocol at three levels. Estograms (histograms of residue-pair distance errors) were converted to residue-residue interaction potentials, which were added to the Rosetta energy function as restraints to guide sampling. Second, the per-residue l-DDT predictions were used to decide which regions to intensively sample or to recombine with other models. Third, global l-DDT prediction was used as the objective function during the selection stages of the evolutionary algorithm and to control the model diversity in the pool during iteration.

Coarse-grained protocol: We experimented with another refinement protocol with more direct DL-guided conformational search using a coarse-grained variant of DeepAccNet. This network variant, called DeepAccNet-cen, uses a coarse-grained local atomic environment (instead of all-atomic) for efficiency. The network replaces the Rosetta centroid energy function at the Monte Carlo search using fragment insertion and/or partial chunk rigid-body movements. We used this DL-guided sampler as the basic unit in a simple evolutionary algorithm in which total ~100 MC trajectories are sampled from 10 structures at every 5 iterations. The final models are further refined by a rapid all-atom refinement protocol and are submitted as models for the group “BAKER-experiment”.

Availability

The all-atom protocol is available through github (https://github.com/hiranumn/DeepAccNet) under "modeling" directory.

1. Park, H., Ovchinnikov, S., Kim, D.E., DiMaio, F., Baker, D. (2018). Protein homology model refinement by large-scale energy optimization. Proc. Natl. Acad. Sci. U. S. A. 115, 3054–3059.
2. Hiranuma, N.,Park, H., Anishchanka, I., Baek, M., Baker, D. (2020). Improved protein structure refinement guided by deep learning based accuracy estimation, bioRxiv, doi: https://doi.org/10.1101/2020.07.17.209643.
3. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations, Proc. Natl. Acad. Sci. U. S. A., 117, 1496-1503.

Bates_BMM

Protein fold construction and complex assembly by employing particle swarm optimization

Raphael.A.G. Chaleil, Tereza Gerguri and Paul.A.Bates

Biomolecular Modelling Laboratory, The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK
paul.bates@crick.ac.uk

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:Y.v; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:N*

The construction, optimization and docking of protein models remains challenging. All require extensive sampling of the high dimensional conformational space, which is intractable with methods based on exhaustive enumeration of all possible solutions. Moreover, the exact contributions of the two recognized mechanisms for protein-protein complex formation, 'conformational selection' and 'induced fit', are not known for any specific interaction. In order to address these problems, we have developed a series of heuristic methods based on Particle Swarm Optimization (PSO).

Methods

Our general methodology for protein fold construction and docking can be described as follows:

i) Fold construction using our automatic server 3D-Jigsaw-SL: The protocol first searches for homologous sequences to the query sequence using HHBlits[1] against a sequence profile database of known structures clustered at 70% sequence identity. A linear *ab initio* polypeptide corresponding to the query sequence is constructed, taking into account the bond lengths, angles and torsion angles accordingly to identified homologous fragments. All the coil regions that are not matched with a structural template are automatically adjusted in torsion angle space. The central core of the algorithm is a constricted PSO[2], which searches for a minimal Dfire[3] statistical pair potential energy. When distance information was available, either from PSICOV[4] or from discontinuous templates, a hookean force was applied as a distance restraint mechanism. Two strategies were applied for folding the structures, the first one adjusts all the torsion angles between all the fragments at once, whereas the second one adjusts the torsion of each linker region (i.e. regions between fragments from templates) one at a time, starting from the N-terminal. The latter technique is computationally more expensive; however, it achieves to generate structures with a smaller radius of gyration (i.e. the structures are more globular). This property allows to generate better, i.e. biophysically sound, models. Finally, the top 10 ranking models from 100 replicates of the algorithm at 10000 iterations (according to Dfire) are then minimized with CHARMM[5] (version 22) and the top structure, identified as having the best CHARMM energy after minimization, is selected for subsequent submission to our protein docking server, SwarmDock. For each section of a protein model different templates might have been chosen; therefore, relating models to single templates is not always possible with this methodology.

ii) Docking using SwarmDock: For the modelling of all protein complexes we used a modification to our binary protein-docking algorithm SwarmDock[6]. Our method uses the principles of PSO to search the parameter docking space. The innovations added to our automated binary server is, for homo-oligomers, to treat each particle within the swarm as an instance of a packed homo-oligomer, constrained by the appropriate symmetry operators. The objective is to optimize the particle space in order to find the most energetically favorable homo-oligomer.

Particles move through a multi-parameter space by the optimization of two sets of parameters: orientations and translations of each monomeric unit relative to the imposed symmetry and linear combinations of normal modes that adjust the conformation of each monomer, in the presence of the other monomers, in this simultaneous docking process. For hetero-oligomeric structures we employed our standard SwarmDock (https://bmm.crick.ac.uk/~svc-bmm-swarmdock) protocol[6]. This docking methodology isn't template based. Moreover, additional information, such as potential sequence conservation at the protein-protein interface, was not considered. The ranking of docked poses was obtained using our 'democratic' scoring system, as previously described[7]. To an extent, we considered both the principle of 'conformational selection' and 'induced fit' in our docking procedure. Conformational selection, by using a variety of starting protein conformations[8], obtained either by our own protein modelling server, *3D-Jigsaw-SL,* or protein models taken from the CASP14 server tar file. Induced fit, is considered too since small adjustments are made in both the backbones and side-chains of the interacting proteins upon docking via the employment of our PSO procedure.

Availability

Our automated binary protein-protein docking server, SwarmDock, can be located at:
https://bmm.crick.ac.uk/~svc-bmm-swarmdock/

1. Remmert M., Biegert A., Hauser A. & Söding J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods. 9(2),173-5.
2. Eberhart, R. C. & Kennedy, J. (1995). A new optimizer using particle swarm theory. In Proceedings of the sixth international symposium on micro machine and human science (pp. 39–43), Nagoya, Japan. Piscataway: IEEE.
3. Yang, Y. & Zhou, Y. (2008). Specific interactions for *ab initio* folding of protein terminal regions with secondary structures. *Proteins* 72, 793-803.
4. Jones D.T., Buchan D.W, Cozzetto D & Pontil M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 28(2), 184-90.
5. Brooks B.R., Brooks C.L. 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner A.R., Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor R.W., Post C.B., Pu J.Z., Schaefer M, Tidor B, Venable RM, Woodcock, H.L., Wu X, Yang W, York, D.M. & Karplus M. (2009). CHARMM: the biomolecular simulation program. J. Comput. Chem. 30(10), 1545-614.
6. Torchala M., Moal I.H., Chaleil R.A.G, Fernandez-Recio, J. & Bates P.A.(2013). SwarmDock: a server for flexible protein-protein docking. Bioinformatics. 29(6), 807-9.
7. Moal, I., Barradas-Bautista, D., Jimenez-Garcia, B., Torchala, M., van der Velde, A., Vreven, T., Weng, Z., Bates, P.A. & Fernandez-Recio., J. (2017). IRaPPA: information retrieval based integration of biophysical models for protein assembly selection. Bioinformatics, 33(12), 1806-1813.
8. Torchala M., Gerguri, T., Chaleil, R.A.G., Gordon, P., Russell, F., Keshani, M. & Bates, P.A. (2020). Enhanced sampling of protein conformational states for dynamic cross-docking within the protein-protein docking server SwarmDock. *Proteins* 88, 962-972.

Protein tertiary structure prediction by Bhattacharya group in CASP14

Md Hossain Shuvo[1], Sutanu Bhattacharya[1], Rahmatullah Roche[1], and Debswapna Bhattacharya[1,2]

[1]Department of Computer Science and Software Engineering and [2]Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA
bhattacharyad@auburn.edu

***Key:** Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

We participated in the CASP14 tertiary structure prediction experiment as a human group "Bhattacharya", which is the result of a system integration of our recently published quality estimation and refinement methods with our newly developed unpublished works in low-homology threading and *de novo* modeling.

Methods

Our pipeline exploited model selection from the CASP server pool using a combination of our newly developed distance-based deep-learning-powered single-model method QDeep[1] and our rapid multi-model structural consensus approach clustQ[2]. It also employed our newly developed unpublished modeling protocols by hybridizing distance- and contact-based hierarchical *de novo* modeling and threading. For each of the top selected models, we independently generated a pool of 100 refined models using our recently published refineD[3] method and ranked them using the method's internal scoring scheme to submit five top-ranked models.

1. Shuvo MH, Bhattacharya S, Bhattacharya D. QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB). 2020; Bioinformatics 2020; 36(S1): i285-i291.
2. Alapati R, Bhattacharya D. clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons. Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB). 2018, pp. 307–314.
3. Bhattacharya D. refineD: improved protein structure refinement using machine learning-based restrained relaxation. Bioinformatics 2019; 35(18): 3320–3328.

Bhattacharya-QDeep, Bhattacharya-QDeepU, Bhattacharya-Server

Protein model accuracy estimation by Bhattacharya groups in CASP14

Md Hossain Shuvo[1] and Debswapna Bhattacharya[1,2]

[1]Department of Computer Science and Software Engineering and [2]Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA.
bhattacharyad@auburn.edu

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

We participated in the CASP14 accuracy estimation category as a server group "Bhattacharya-QDeep" to test our newly developed distance-based deep-learning-powered single-model method QDeep[1]. We also tested a variant of the original QDeep method by separately participating as a server group "Bhattacharya-QDeepU". Additionally, we participated as a server group "Bhattacharya-Server" to test our rapid multi-model structural consensus approach clustQ[2].

Methods

QDeep method[1], tested in "Bhattacharya-QDeep" server, utilizes an ensemble of four deep residual neural network (ResNet)[4] classifiers to estimate the likelihood of residue-level C_α errors of a model at four different error thresholds of 1, 2, 4, and 8Å. Each of the four ResNet classifiers was independently trained using sequence- and structure-derived features that include distance map similarities. Ensemble averaging of the error likelihoods was then used for estimating the local and global accuracy scores. In "Bhattacharya-QDeepU", we tested a variation of the original QDeep method retrained using multiple sequence alignments generated by merging sequences from whole-genome sequence databases with metagenome database.

In "Bhattacharya-Server", we tested our multi-model structural consensus approach clustQ[2], which performs superposition-free weighted internal distance comparisons to rapidly compute the average pairwise similarity of a model with respect to other models in the model pool for estimating its global accuracy score.

Availability

QDeep is freely available at https://github.com/Bhattacharya-Lab/QDeep/.
clustQ is freely available at http://watson.cse.eng.auburn.edu/clustQ/.

1. Shuvo MH, Bhattacharya S, Bhattacharya D. QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB). 2020; Bioinformatics 2020; 36(S1): i285-i291.
2. Alapati R, Bhattacharya D. clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons. Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB). 2018, pp. 307–314.

Bhattacharya, Bhattacharya-Server

Protein structure refinement by Bhattacharya groups in CASP14

Md Hossain Shuvo[1] and Debswapna Bhattacharya[1,2]

[1]*Department of Computer Science and Software Engineering and* [2]*Department of Biological Sciences, Auburn University, Auburn, AL 36849, USA.*
bhattacharyad@auburn.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

We participated in CASP14 refinement experiment both as a human group "Bhattacharya" to test our newly developed unpublished refinement protocol, and as a server group "Bhattacharya-Server" to test our recently-published structure refinement method refineD[1].

Methods

Our newly developed unpublished refinement protocol tested in "Bhattacharya" human group starts by estimating residue-level C_α errors of the starting structure at four different error thresholds of 0.5, 1, 2, and 4Å, predicted using an ensemble of four deep residual neural network (ResNet)[2] classifiers trained on sequence- and structure-derived features[3]. These residue-level C_α errors are subsequently converted to multi-resolution restraints to be integrated with Rosetta's all-atom energy function[4] as additional scoring terms during structure refinement. A pool of 300 refined models was generated per target by iteratively employing Rosetta's FastRelax protocol[5]. We then combined the error estimation from the ensemble of deep ResNets to score the refined structures in conjunction with our rapid multi-model structural consensus approach clustQ[6] for selecting five refined models per target for submission.

In "Bhattacharya-Server", we tested our published refineD[1] protocol by generating 100 refined models per target and then selecting five refined models for submission following the above scoring strategy.

Availability

refineD is freely available at http://watson.cse.eng.auburn.edu/refineD/.

1. Bhattacharya D. refineD: improved protein structure refinement using machine learning-based restrained relaxation. *Bioinformatics 2019; 35(18): 3320–3328.*
2. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, pp. 770–778.*
3. Shuvo MH, Bhattacharya S, Bhattacharya D. QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB). 2020; Bioinformatics 2020; 36(S1): i285-i291.*
4. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput 2017; 13: 3031–3048.*
5. Khatib F, Cooper S, Tyka MD, et al. Algorithm discovery by protein folding game players.

Proc Natl Acad Sci U S A 2011; 108: 18949–18953.

6. Alapati R, Bhattacharya D. clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons. *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB). 2018, pp. 307–314*.

Bioinsilico_sbi

Three-dimensional prediction of proteins using a collection of sMotifs

A. Molina Martinez de los Reyes[1], J. Aguirre-Plans[1], N. Fernandez-Fuentes[2,3] and Baldo Oliva[1]

[1] – Strucutural Bioinformatics Lab (GRIB-IMIM), Department of Experimental and Health Science, University Pompeu Fabra, Catalonia, Spain, [2] – IBERS, Aberystwyth University, United Kingdom [3] – Univeristy of Vic, Catalonia, Spain

joaquim.aguirre@upf.edu, alexis.molina@alum.esci.upf.edu, naf4@aber.ac.uk, baldo.oliva@upf.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; Fragm:YSmotifs; Cont:N; Dist:N; Tors:N; DeepN; EMA:Y; MD:N*

The relationship between protein structure and protein function is close and brought together under the light of evolution[1]. Evolution tend to preserve energetically-favourable interactions between selected protein residues that play an important role in the function or structure (or both) of proteins. Thus, there is a certain degree of coevolution between those residues on the all members that belong to the same protein family[2] . We recently developed a novel approach named RADI, for Reduced Alphabet Direct Information, that uses a modified version of the direct-coupling analysis (DCA) algorithm and allows for fast computation of direct information values. Coevolving residues are used to drive and restraint the folding of the sequence into a three-dimensional (3D) structure. Along with this information we also used a library of super-secondary structure motifs, named sMotifs, derived from our loop structure database ArchDB14. Our algorithm ArchDBMap[3] is use to retrieve the sMotifs best matching a query sequence that will ultimately use as templated to model the 3D structure.

Methods

We used the following approach to model the structures of protein based on DI contact prediction and sMotifs:

1. We map the secondary structure predicted with SABLE4 on the sequence of the target and predict the type of super-secondary structures defined as sMotifs and classified in ArchDB14.

2. The sequence of the proteins is then used to compute the DI and select for each alphabet the top 40 pairs of residues with the higher correlation.

3. The structure of sMotifs aligned to the target sequence are used as templates for homology modelling with MODELLER5 . We add distance restraints between the pair of amino acids selected, constrain the secondary structure predicted with SABLE and generate 1000 structural models that are subsequently clustered and scored.

4. The protocol to run MODELLER is as follows:

a) we use as templates the structures of the predicted sMotifs

b) apply constraints at 8Å using a Gaussian potential on the Cβ-Cβ atoms of the selected residue-pairs with highest correlation

c) we force the type of secondary structure as mapped by the prediction of secondary structure.

5. Finally, we rank the models with DOPE6 and cluster them by similar structure, evaluate the quality of the models with Prosa20037 and select the best scored structures.

If the search performed with ArchDBMap was to throw no results, the sMotifs were replace by a single template built using the build_sequence function integrated in MODELLER.

Availability

RADI is available at: https://github.com/structuralbioinformatics/RADI
ArchDBMap is available at: https://github.com/structuralbioinformatics/archdbmap

1. Lewis, T. E., Sillitoe, I., Andreeva, A., Blundell, T. L., Buchan, D. W., Chothia, C., ... & Jones, D. T. (2015). Genome3D: exploiting structure to help users understand their sequences. Nucleic acids research, 43(D1), D382-D386.
2. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A., & Bonvin, A. M. (2018). Assessment of contact predictions in CASP12: co‐evolution and deep learning coming of age. Proteins: Structure, Function, and Bioinformatics, 86, 51-66.
3. Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marín-López, M. A., Fernandez-Fuentes, N., & Oliva, B. (2014). ArchDB 2014: structural classification of loops in proteins. Nucleic acids research, 42(D1), D315-D319.
4. Adamczak, R., Porollo, A., Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins. 59, 467-475.
5. Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. Current protocols in bioinformatics, 54(1), 5-6.
6. Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A., Rost, B., & Sali, A. (2002). Reliability of assessment of protein structure prediction methods. Structure, 10(3), 435-440.
7. Sippl, M. J. (1993). Recognition of errors in three‐dimensional structures of proteins. Proteins: Structure, Function, and Bioinformatics, 17(4), 355-362.

Bioinsilico_sbi

Protein contact predictions using a reduced alphabet and direct-coupling analysis

A. Molina Martinez de los Reyes[1], J. Aguirre-Plans[1], N. Fernandez-Fuentes[2,3] and Baldo Oliva[1]

[1] – Strucutural Bioinformatics Lab (GRIB-IMIM), Department of Experimental and Health Science, University Pompeu Fabra, Catalonia, Spain, [2] – IBERS, Aberystwyth University, United Kingdom [3] – Univeristy of Vic, Catalonia, Spain

joaquim.aguirre@upf.edu, alexis.molina@alum.esci.upf.edu, naf4@aber.ac.uk, baldo.oliva@upf.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepN; EMA:Y; MD:N*

The identification of coevolved residue pairs in protein sequences is widely used to help the prediction of three-dimensional (3D) structure in proteins[1]. Besides functional implication often pairs of coevolved residues inform of the 3D closeness and thus it can be used to guide structural prediction of proteins in the form of distance restraints[2]. Direct-coupling analysis (DCA) is used currently to identify such pairs of residues but at a high computational cost[3]. We recently developed a novel computational approach named RADI, for Reduced Alphabet Direct Information which present novel ideas to improve the speed of calculation of direct information values[4]. By using a simplified alphabet, i.e. grouping amino acids with similar physicochemical properties, RADI achieved can achieved a reduction of the computational without loss of accuracy as proved on a benchmark set. We have now applied RADI on a blind test using the sequences submitted to CASP14 under residue-residue contact prediction section. Overall, we provided prediction for 66 submitted targets.

Methods

The protocol followed to computed DI values from RADI as follow:

1. Generation of multiple-sequence alignments (MSAs): MSAs were created using the script "buildmsa.py" included in the RADI Git repository. First, the script builds a profile of the query searching for similar sequences in the uniref50 database with MMseqs2[5]. Next, it uses the query profile to find more sequence relatives in the uniref100 database. Then, the script builds a MSA of the query and the identified sequences (up to 100,000) with FAMSA[6]. Finally, it removes the columns of the MSA with insertions in the query. Note that MMseqs2 is executed with options "-s 7.5" and "--max-seq-id 1.0" for a more sensitive search.

2. Secondary structure prediction*:* The secondary structures were predicted using SABLE[7] and a 3-state alphabet, namely: helix (H), beta(E) and coil (C).

3. Calculation of DI values*.* The calculation of DI values was done using the original DCA algorithm as implemented in RADI utilizing four different alphabets, namely RA0, RA1, RA2, and RA3 (for more information on the method please refer to original publication[4].)

(i) RA0 stand for an alphabet of size $q = 21$ (i.e. 20 different amino acids plus the gap)

(ii) RA1 has a $q = 9$ represented by Positively charged: {Arg, His, Lys}. Negatively charged: {Asp, Glu}. Polars: {Ser, Thr, Asn, Gln}. Aliphatics: {Ala, Ile, Leu, Met, Val}. Aromatics: {Phe, Trp, Tyr}. Single groups: {Cys}, {Gly}, {Pro} and the gap;

(iii) RA2 has a q = 5 represented by Polar: {Arg, His, Lys, Asp, Glu, Ser, Thr, Asn, Gln, Cys}. Non-polar: {Ala, Ile, Leu, Met, Val, Phe, Trp, Tyr}. Single groups: {Gly}, {Pro} and the gap; and

(iv) RA3 has a q = 3 represented by Polar: {Arg, His, Lys, Asp, Glu, Ser, Thr, Asn, Gln, Cys, Gly}. Non-polar: {Ala, Ile, Leu, Met, Val, Phe, Trp, Tyr, Pro}. Single groups: gap

For each of the alphabet, i.e. RA{0-3} DI values are acquire for pair of amino acid belonging to two different secondary structures, i.e. pairs of residues within same secondary structure were not considered.

4. Selection and submission of top DI values. The DI values were normalized using a max-min normalization assuming 1 for the top DI value a 0 for the lowest.

Availability

RADI is available at: https://github.com/structuralbioinformatics/RADI

1. Marks, D.S., Colwell, L.J., Sheridan R., Hopf T.A., Pagnani A., Zecchina R., Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One. 6, e28766.
2. Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V., Baker, D. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. Elife. 4, e09248.
3. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A. 108, 1293-1301.
4. Anton, B., Besalu, M., Fornes, O., Bonet, J., Cuevas G. De las, Fernandez-Fuentes N., Oliva, B. (2018) RADI (Reduced Alphabet Direct Information): Improving execution time for direct-coupling analysis. bioRxiv. 406603, doi: https://doi.org/10.1101/406603
5. Steinegger, M., Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 35, 1026-1028.
6. Deorowicz, S., Debudaj-Grabysz, Gudys, A. (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Sci Rep. 6, 33964.
7. Adamczak, R., Porollo, A., Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. Proteins. 59, 467-475

Bioinsilico_sbi, Bioinsilico_sbi_PAIR

Assessing the quality of protein structural models using split-statistical potentials

A. Molina Martinez de los Reyes[1], J. Aguirre-Plans[1], N. Fernandez-Fuentes[2,3] and Baldo Oliva[1]

[1] – Strucutural Bioinformatics Lab (GRIB-IMIM), Department of Experimental and Health Science, University Pompeu Fabra, Catalonia, Spain, [2] – IBERS, Aberystwyth University, United Kingdom [3] – Univeristy of Vic, Catalonia, Spain

joaquim.aguirre@upf.edu, alexis.molina@alum.esci.upf.edu, naf4@aber.ac.uk, baldo.oliva@upf.edu

Key*: Auto:Y; CASP_serv:N; Templ:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepN; EMA:Y; MD:N*

Many scoring methods have been proposed to assess the quality of protein fold models[1-9]. Knowledge-based potentials are scoring functions derived from the analysis of empirical data[5] often used to evaluate the quality of models of a protein structure using the frequencies of residue-residue contacts per distance. Several computational methods have been implemented from knowledge-based potentials[1,2,8] . Split-Statistical Potentials (SPs) are knowledge-based potentials that consider the frequency of pairs of residues in contact and include their structural environment, such as solvent accessibility and type of secondary structure. Previously, we demonstrated that SPs can be used to: (i) identify near-native protein decoys in structure prediction[10]; and (ii) rank protein-protein docking poses[11]. The scoring of the quality of a protein structure using the Split-Statistical Potentials is available in an online server (SPserver).

Methods

Scoring: Scores are calculated using the description of a potential of mean force with the frequencies of residue-residue contacts per distance. Residue-residue contacts need to consider the amino acids type, the distance between them, and environmental features such as the type of secondary structure or the degree of exposure of the amino acids. The SPServer has 6 types of SPs available that differ on the environmental features considered for the contact definition. We use one of them defined as PAIR, which considers solely amino acid frequencies along distances[10]. The score is defined by the description of a potential of mean force (PMF). Then we define the PMF potentials as in equations 1:

$$PMF_{PAIR}(a,b) = -k_B T \log\left(\frac{P(a,b \mid d_{ab})}{P(a)\,P(b)\,P(d_{ab})}\right) \text{ (eq. 1)}$$

With k_B the Boltzmann constant, T the standard temperature (300K) and d_{ab} the distance between both residues. The terms P(·) denote the probabilities of observing interacting pairs (with or without conditions). For instance, $P(a,b|d_{ab})$ is the conditional probability that residues a,b interact at distance smaller than or equal to d_{ab}, and $P(d_{ab})$ is the probability of finding any pair of residues interacting at distance smaller than or equal to dab. The score PAIR is calculated as:

$$AIR = \sum_{a,b} PMF_{PAIR}(a,b) \text{ (eq. 2)}$$

Input: As input, users have to provide the structures of one or more proteins or protein complexes. The server input is flexible; users can provide either PDB structures, mmCIF files or compressed directories containing the structures to analyze.

Output for protein folds: For a set of protein folds, the SPServer outputs: (i) the global scores (raw and normalized) of PAIR; and (ii) the scoring profile per residue (local scores) along the protein sequence (by summing all the interactions specific for one residue). Global scores account for the overall quality of structural models, while per-residue score plots pinpoint problematic regions of the models that likely have either a wrong conformation or contacts with a wrongly modelled region. The normalization of the scores is obtained as a Z-score with respect to the scores of 1000 random sequences folded with the same structure.

Availability

The program is available in http://sbi.upf.edu/spserver/. The user can use one or several proteins as input and analyze both global and local scores.

1. Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. Adv Protein Chem Struct Biol. 2014;94:77-120.
2. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35:W407-10.
3. Melo F, Devos D, Depiereux E, Feytmans E. ANOLEA: a www server to assess protein structures. ProcIntConfIntellSystMolBiol. 1997;5:187.
4. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997;277:396.
5. Maghrabi AHA, McGuffin LJ. ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. Nucleic Acids Res. 2017;45:W416-W21.
6. Uziela K, Menendez Hurtado D, Shu N, Wallner B, Elofsson A. ProQ3D: improved model quality assessments using deep learning. Bioinformatics. 2017;33:1578-80.
7. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27:343-50.
8. Conway P, DiMaio F. Improving hybrid statistical and physical forcefields through local structure enumeration. Protein Sci. 2016;25:1525-34.
9. Olechnovic K, Venclovas C. VoroMQA web server for assessing three-dimensional structures of proteins and protein complexes. Nucleic Acids Res. 2019;47:W437-W42.
10. Aloy P, Oliva B. Splitting statistical potentials into meaningful scoring functions: testing the prediction of near-native structures from decoy conformations. BMC Struct Biol. 2009;9:71.
11. Feliu E, Aloy P, Oliva B. On the analysis of protein-protein interactions via knowledge-based potentials for the prediction of protein-protein docking. Protein Sci. 2011;20:529-41.

BrainFold

Contact Pair Prediction Using a Deep Neural Net

Susan Sullivan[1], Tomasz Religa[1], Warren Aldred[1], Ram Ganesamoorthy Kasthuri[1],
Farheen Omar[1] and Tyler Romero[1]
[1] - *Microsoft*
susansu@microsoft.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Our fully automated server uses a dense deep neural net to predict whether a residue pair in a protein sequence will be in contact in the folded protein. This classifier is run for each residue pair combination in the sequence to determine the set of contact pairs.

Methods

Our server makes use of a deep learning model that was trained on a data featurization of a subset of the atomic coordinate data available from PDB (https://www.wwpdb.org/). For a residue pair to be considered in contact (in both the training data as well as runtime predictions), the distance needs to be less than 8 angstroms between the CB atoms in the two residues, except for Glycine, where the CA atom is used.

The model is a dense DNN that's 4 layers deep and consists of about 120K input parameters. The protein files from PDB that were used to create the training data were scoped to only protein files that:

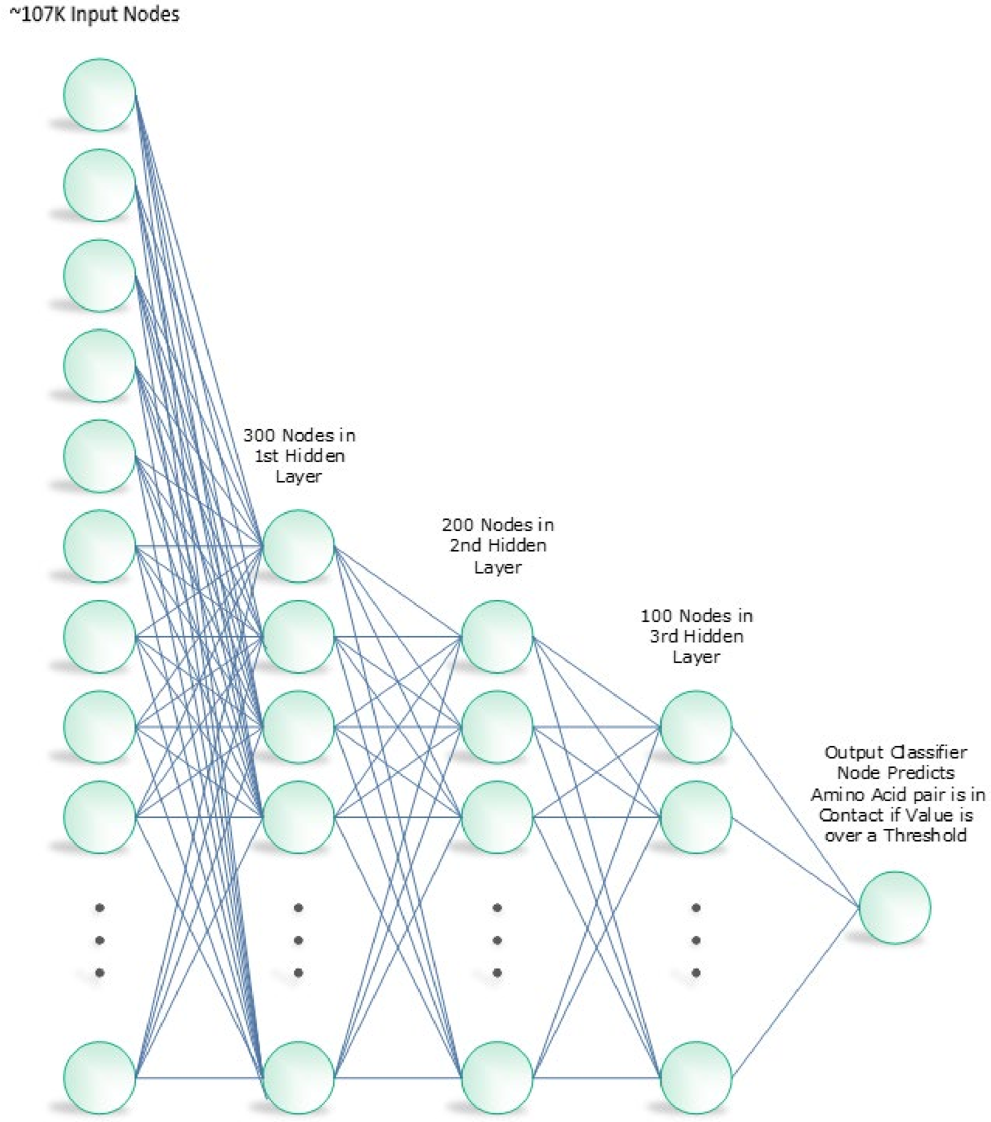
- Are internally consistent (that is, where the SEQRES and the ATOMs sections are consistent)
- Were determined by X-ray diffraction
- Were <= 1700 residues in length (since that's what the model can support)
- Don't contain nucleotides

We didn't have time to train with all such PDB protein files that met that criteria, but rather a smaller fraction of them. Model training continued during the competition, and the model version used for CASP14 predictions was updated periodically.

Predictions were ranked by the score the model assigned to each, ranging from 0 (least likely) to 1 (most likely). Contact pairs with a score >= .5 were reported as predicted contact pairs.

The architecture of the DNN is as follows:

DNN Architecture Diagram


~107K Input Nodes
300 Nodes in 1st Hidden Layer
200 Nodes in 2nd Hidden Layer
100 Nodes in 3rd Hidden Layer
Output Classifier Node Predicts Amino Acid pair is in Contact if Value is over a Threshold

The feature inputs are as follows:

1) Index of the first amino acid being examined for being in contact.

2) Index of the second amino acid being examined for being in contact.

3) Inverse of the distance between the amino acids being examined, measured in the amino acid count between them in the chain.

4-1701) (1698 categorical features) - An enum value representing which kind of amino acid is at that position. If it's 0, then that position is padded. If it's 21, that represents a break between chains. 1 through 20 represent the 20 different kinds of amino acids. If there are fewer than 1699 residues before the first residue being examined, the residues will be on the right, and the values on the left will be padded (i.e. zeros).

1702) (Categorical feature) - An enum value representing which kind of amino acid the first residue being checked for being in contact is.

1703-3400) (1698 categorical features) - An enum value representing which kind of amino acid is at that position. If it's 0, then that position is padded. If it's 21, that represents a break between chains. 1 through 20 represent the 20 different kinds of amino acids. If there are fewer than 1698 residues between the two residues being examined, the residues will be on the left, and the values on the right will be padded (i.e. zeros).
3401) (Categorical feature) - An enum value representing which kind of amino acid the second residue being checked for being in contact is.
3402-5099) (1698 categorical features) - An enum value representing which kind of amino acid is at that position. If it's 0, then that position is padded. If it's 21, that represents a break between chains. 1 through 20 represent the 20 different kinds of amino acids. If there are fewer than 1698 residues after the second residue being examined, the residues will be on the left, and the values on the right will be padded (i.e. zeros).
NOTE – Each categorical feature is input to the neural net as a one-hot vector. For padding amino acids, that one-hot vector is all zeros.

CAO-QA1 (High Accuracy Modeling)

Collaborative protein structure prediction with deep learning based de novo prediction and model selection

Mikhail Korovnik[1], Kyle Hippe[1], John C Oakley[1], Nathan Ranno[2], Chris Holland[1], Sola Gbenro[1], Ashwin Deodhar[1], Cade Lilley[1], Yajun An[3], Jie Hou[4], Dong Si[2], Dongyeon Joo[3], and Sailu Li[5], Renzhi Cao[1]

1 - Department of Computer Science, Pacific Lutheran University, 2 - Division of Computing Software Systems, University of Washington Bothell, 3 - School of Interdisciplinary Arts and Sciences, University of Washington, Tacoma, 4 - Department of Computer Science, Saint Louis University, 5 - School of Business, Pacific Lutheran University

caora@plu.edu

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:Y.1-20; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y*

We tested our new pipeline that uses manual intervention for all Human Prediction targets in CASP 14. This has been a collaborative research project since the amount of computational resources on our main server is limited and most contributors for this project are currently enrolled in undergraduate programs. We blindly tested different modeling tools and prediction ranking methods, but the general procedure of this pipeline stays roughly the same. Our pipeline utilized contact prediction and deep learning techniques for model ranking, which demonstrated powerful potential in the previous CASP [1]. In particular, the protein decoy pool is generated from both our own *de novo* prediction method as well as server predictions from CASP participants. The highest-ranking predictions are automatically refined and submitted by our server. Manual intervention is used in the process of executing scripts, reviewing and modifying final predictions, and assembly of large proteins that our servers cannot handle. In a few instances where our server could not meet expiration deadlines, we hand-picked CASP-hosted server predictions and predictions from our own *de novo* prediction method based on our prior knowledge about CASP and basic understanding of protein structures.

Methods

Step 1, the tools MetaPSICOV2 [2], CCMpred [3], and FreeContact [4] were used to make contact prediction from the protein sequence, and PSIPRED [5] was used to predict the secondary structure from the protein sequence.

Step 2, secondary structure prediction and contact prediction from the previous step were used in Unicon3D [6] for *de novo* protein structure predictions. Predictions were submitted online and collected by humans to send to the main server for quality assessment. CASP-hosted server predictions were also collected by the main server for quality assessment.

Step 3, *de novo* predictions (in-house tool) and CASP-hosted server predictions (“server” pool) were scored and ranked in their separate pools by a quality assessment tool. For the majority of CASP, the deep learning tools DeepQA [7] and QDeep [8] were used for this step. We also generated 2-5 structure predictions using DMPfold [9] and added them to our pool for model selection. DeepMSA [10], using the UniClust30 [11], UniRef90 [12], and MetaClust50 [13] databases, generated multiple sequence alignments (MSA) for input to QDeep. We have also experimented with Ornate [14], AngularQA [15], TopQA [16] and another in-house tool that uses a novel approach that we refer to as a "hierarchical structure" machine learning technique for model selection.

Step 4, after the protein model quality assessment, our pipeline automatically selects the top-3 highest scoring predictions from the server pool and the top-2 highest scoring predictions from our own in-house *de novo* predictions. For the server pool, if more than one prediction comes from the same group, then only the highest-scoring prediction from that group is chosen. The pipeline skips down the ranked list until it finds a new group to which it selects the next model from and proceeds until 3 total predictions have been selected from the server pool.

Step 5, once the final 5 predictions have been selected in the previous step, we refine those models using a tool called ModRefiner [17]. For a number of CASP 14 targets, the highest-ranking prediction from the server pool is selected as the reference structure for refining the other four predictions (the highest-ranking prediction would therefore be refined with itself as reference). We have also experimented with refining each prediction with itself as reference to see how it affects prediction accuracy and to limit the amount of changes made to the original pre-refined structure.

Step 6, the final five structure predictions are automatically submitted after the refinement step before each expiration date. If there is enough time before the submission deadline, human intervention is used by reviewing the final predictions in Chimera [18] and deciding whether any predictions need to be replaced. If so, then a different prediction from the server pool is selected, refined (like in the previous step), and submitted to replace one of predictions that were automatically submitted by the main server. Due to uncertainties in blindly testing QDeep and our in-house tool, we decided partway through CASP 14 that predictions from Zhang-Server_TS1 [19] will always replace the second-highest ranking model from our pool (thus meaning that 4 of our predictions come from CASP-hosted server predictions, and only 1 from our own pool). If Zhang-Server_TS1 was already among the top-3 in step 4, then the second-highest ranking model from our pool was used or a different prediction was hand-selected to replace it.

For large sequences containing sub-units (name starting with an "H" instead of a "T"), structures were either predicted using an in-house *ab initio* tool, or human intervention was used to assemble the protein by hand. DMPfold predictions were also sometimes used. The sub-units used for manual assembly were first selected as described in the steps above, and then put together in Chimera. Contact predictions produced by DeepMSA and predictions made by the in-house *ab initio* and DMPfold were sometimes used as reference for the manually assembled structure prediction. Predictions made by the tools listed here did not go through a refinement step, but the individual subunits selected by the QA tool were.

Availability

The software of our method is not ready for publishing yet, if you want to download the software, please contact Dr. Cao (caora@plu.edu) for the updates.

1. Hou, J., Wu, T., Cao, R. & Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins 87, 1165–1178
2. Buchan, D.W.A. & Jones, D.T. (2018). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins 86 Suppl 1, 78–83

3. Seemayer, S., Gruber, M. & Söding, J. (2014). CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 30, 3128–3130
4. Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B. (2014). FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15, 85
5. McGuffin, L.J., Bryson, K. & Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405
6. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics 32, 2791–2799
7. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 17, 495
8. Shuvo, M.H., Bhattacharya, S. & Bhattacharya, D. (2020). QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. Bioinformatics 36, i285–i291
9. Greener, J.G., Kandathil, S.M. & Jones, D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. 10, 3977
10. Zhang, C., Zheng, W., Mortuza, S.M., Li, Y. & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 36, 2105–2112
11. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J. & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170–D176
12. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H. & the UniProt Consortium (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932
13. Steinegger, M. & Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nat. Commun. 9, 2542
14. Pagès, G., Charmettant, B. & Grudinin, S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics 35, 3313–3319
15. Conover, M., Staples, M., Si, D., Sun, M. & Cao, R. (2019). AngularQA: protein model quality assessment with LSTM networks. Computational and Mathematical Biophysics 7, 1–9
16. Smith, J., Conover, M., Stephenson, N., Eickholt, J., Si, D., Sun, M. & Cao, R. (2020). TopQA: a topological representation for single-model protein quality assessment with machine learning. International Journal of Computational Biology and Drug Design 13, 144
17. Xu, D. & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys. J. 101, 2525–2534
18. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. & Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612
19. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7–8

CAO-QA1(Accuracy Estimation)

AngularQA: Protein Model Quality Assessment with LSTM Networks

Kyle Hippe[1], Mikhail Korovnik[1], Matthew Conover[1], Max Staples[1], Dong Si[2], Miao Sun[3], Renzhi Cao[1*]

1 - Pacific Lutheran University, 2 - University of Washington Bothell, 3 - JingChi Inc.

caora@plu.edu

***Key:** Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Quality Assessment (QA) plays an important role in protein structure prediction. Traditional multimodel QA method usually suffers from searching databases or comparing with other models when making predictions, which usually fail when the poor quality models dominate the model pool. We propose a novel protein single-model QA method AngularQA [1] which is built on a new representation that converts raw atom information into a series of carbon-alpha (Cα) atoms with side-chain information, defined by their dihedral angles and bond lengths to the prior residue. An LSTM network is used to predict the quality by treating each amino acid as a time-step and consider the nal value returned by the LSTM cells. To the best of our knowledge, this is the first time anyone has attempted to use an LSTM model on the QA problem; furthermore, we use a new representation which has not been studied for QA. In addition to angles, we utilize sequence properties like secondary structure parsed from protein structure at each time-step without using any database, which is different from all existing QA methods. Our experiment points out new directions for QA problems and our method could be widely used for protein structure prediction problems.

Methods

For the initial data preparation part, all data used in training our LSTM network comes from 3DRobot decoys [2] and CASP 9, 10, and 11 [3]. These have 92,535, 36,083, 15,901, and 14,193 models respectively from which we draw for training. Validation occurs on the CASP12, of which we use 6,790 models across 40 targets [3]. We begin by filtering all the models. During this process we verify the residue sequences in the predicted structures line up correctly with the native structure, and throw out any predicted models with gaps in the center. In addition, We throw out any models for which we do not have the native structure. After filtering, we are left with a total of 128,439 models with 121,875 training models and 6564 validation models.

After that, we calculate the angles and bond lengths along the backbone and side-chain as was described by UniCon3D[4]. The result is a sequence of angle and bond length information provided for each residue following along the carbon backbone. In addition, we also calculate the proximity counts, which are also calculated by counting the number of Cα atoms within a set radius of each residue's Cα atom. We perform this calculation for all radii in the discrete range [5Å, 15Å]. Moreover, the secondary structure is parsed by the program, DSSP[5], but no secondary structure prediction is used in our method, which is different from a lot of traditional QA methods [6–12]. The machine learning technique is applied to train a LSTM network on the processed feature vectors, and each LSTM cell uses a hyperbolic tangent activation with a hard sigmoid recurrent activation.

Figure 1 demonstrates the flowchart of our AngularQA method.

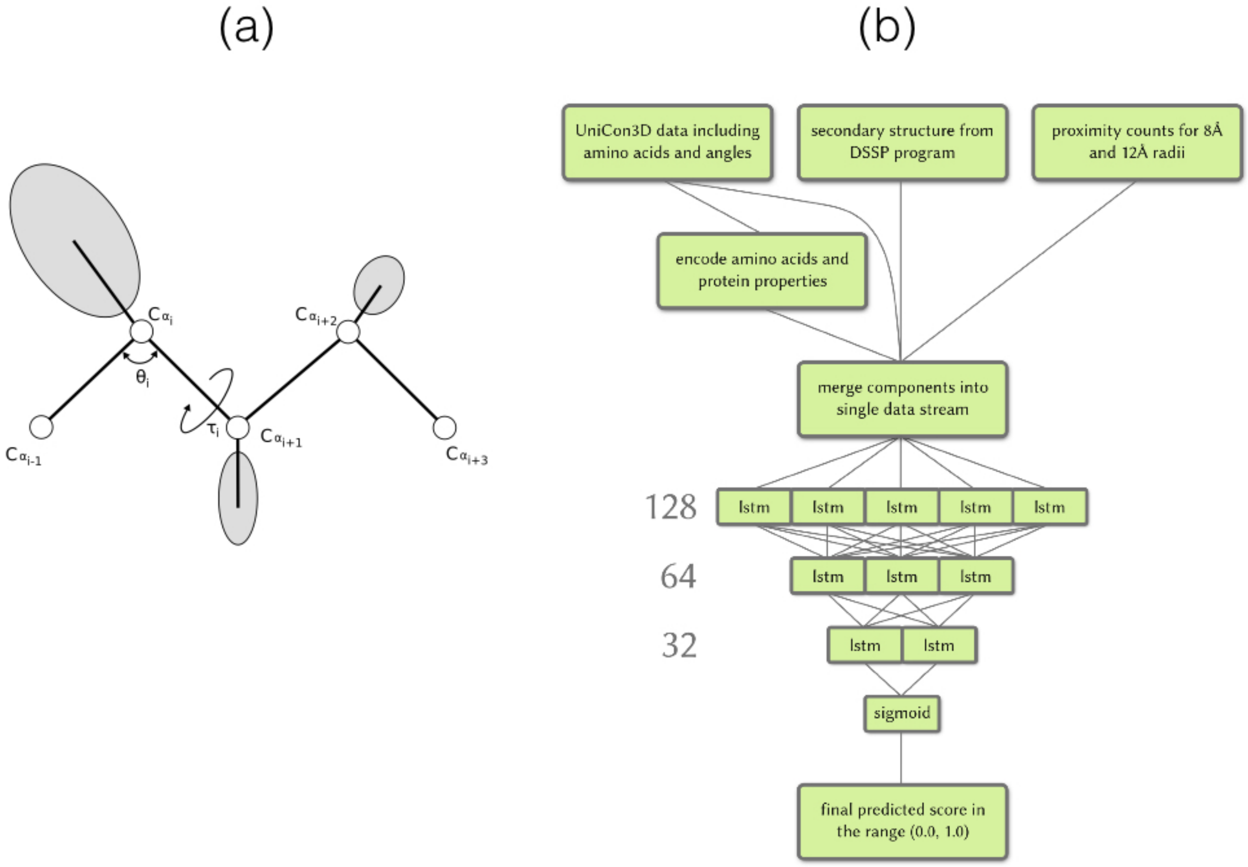

(a)
C_{α_i}
$C_{\alpha_{i+2}}$
θ_i
τ_i
$C_{\alpha_{i+1}}$
$C_{\alpha_{i-1}}$
$C_{\alpha_{i+3}}$
(b)
UniCon3D data including amino acids and angles
secondary structure from DSSP program
proximity counts for 8Å and 12Å radii
encode amino acids and protein properties
merge components into single data stream
128
64
32
lstm
sigmoid
final predicted score in the range (0.0, 1.0)

Figure 1. Flowchart of AngularQA method

Availability

The software is freely available at GitHub: https://github.com/caorenzhi/AngularQA.

1. Conover, M., Staples, M., Si, D., Sun, M. & Cao, R. (2019). AngularQA: protein model quality assessment with LSTM networks. Computational and Mathematical Biophysics 7, 1–9
2. Deng, H., Jia, Y. & Zhang, Y. (2016). 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics 32, 378–387
3. Moult, J., Pedersen, J.T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins 23, ii–v
4. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics 32, 2791–2799
5. Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C. & Vriend, G. (2011). A series of PDB related databases for everyday needs. Nucleic Acids Res. 39, D411–9

6. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. Bioinformatics 33, 1578–1580
7. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. Sci. Rep. 6, 23990
8. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 17, 495
9. Manavalan, B. & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 33, 2496–2503
10. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins 85, 1131–1145
11. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. Bioinformatics doi:10.1093/bioinformatics/bty494
12. Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 84 Suppl 1, 349–369

CAO-SERVER(Accuracy Estimation)

TopQA: a topological representation for single-model protein quality assessment with machine learning

Mikhail Korovnik[1], Kyle Hippe[1], John Smith[1], Matthew Conover[1], Natalie Stephenson[1], Jesse Eickholt[2], Dong Si[3], Miao Sun[4], Renzhi Cao[1]

[1] - Pacific Lutheran University, [2] - Central Michigan University, [3] - University of Washington Bothell,
[4] - JingChi Inc.

caora@plu.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Correctly predicting the complex three-dimensional structure of a protein from its sequence would allow for a superior understanding of the function of specific proteins with many applications. We propose a novel method, TopQA 1, which is aimed to tackle a crucial step in the protein prediction problem: assessing the quality of generated predictions. Our method, to the best of our knowledge, is the first type of method to analyse the topology of the predicted structure. We found that our new representation provided accurate information regarding the location of the protein's backbone. Using this information, we implemented a novel algorithm based on convolutional neural networks (CNN) to predict GDT_TS score for given protein models.

Methods

First, we prepared the training datasets for developing TopQA. We used a total of 176 target proteins from the CASP10 and CASP11 datasets (These can be found at: http://predictioncenter.org/download_area/), including 15,901 CASP10 models and 14,139 CASP11 models. Each protein structure model is in PDB format, and provides a standard representation for macromolecular structure data. Traditional methods 2–8 normally use the 3D structure of protein models (in PDB format) directly with help of other properties of the protein sequence, but no method has tried to modify the representation of the 3D structure model. We proposed a new representation of the 3D structure model and used that for training machine learning models.

Second, we created our new representation for each PDB file. The 3D coordinates of each carbon alpha atom were extracted, and the whole topology of this structure was kept while we scale the structure into a cube with size 1. In addition, this representation systematically mapped the mass of each carbon alpha atom in the backbone of the protein model to a three-dimensional space in the cube. This 1x1x1 cube can be scaled to any size, although for our model we generally used a 52x52x52 (see the results section for more information regarding varying dimensions). Finally, rotations were applied to this new representation to generate a robust model. With this approach, we were able to map each model numerous times, viewing the model from a slightly different angle each time. Normally, it's very costly to apply rotation to the model, but one rotation of each model in our model representation only takes a second and would be used in our final representation. Once we formatted the PDB files into this representation, we were left with a 3-dimensional matrix in which every value represented the mass of a single atom in the protein's backbone (several of these values were zero, as the matrix included the empty space of the cube surrounding the protein structure as well as the empty space encapsulated by the structure)

Finally, after transforming the pdb files into our new topologically-based representation, we trained a convolutional neural network (CNN) model. This CNN was made of two convolutional layers, a single pooling layer as well as two dense layers. The CNN was an appealing choice of machine learning method as it lends itself to images and matrices quite well [9]. We have also considered other types of machine learning methods such as an SVM, but found that CNN performed the best in our experiments. **Figure 1** shows the overall flowchart of our method.

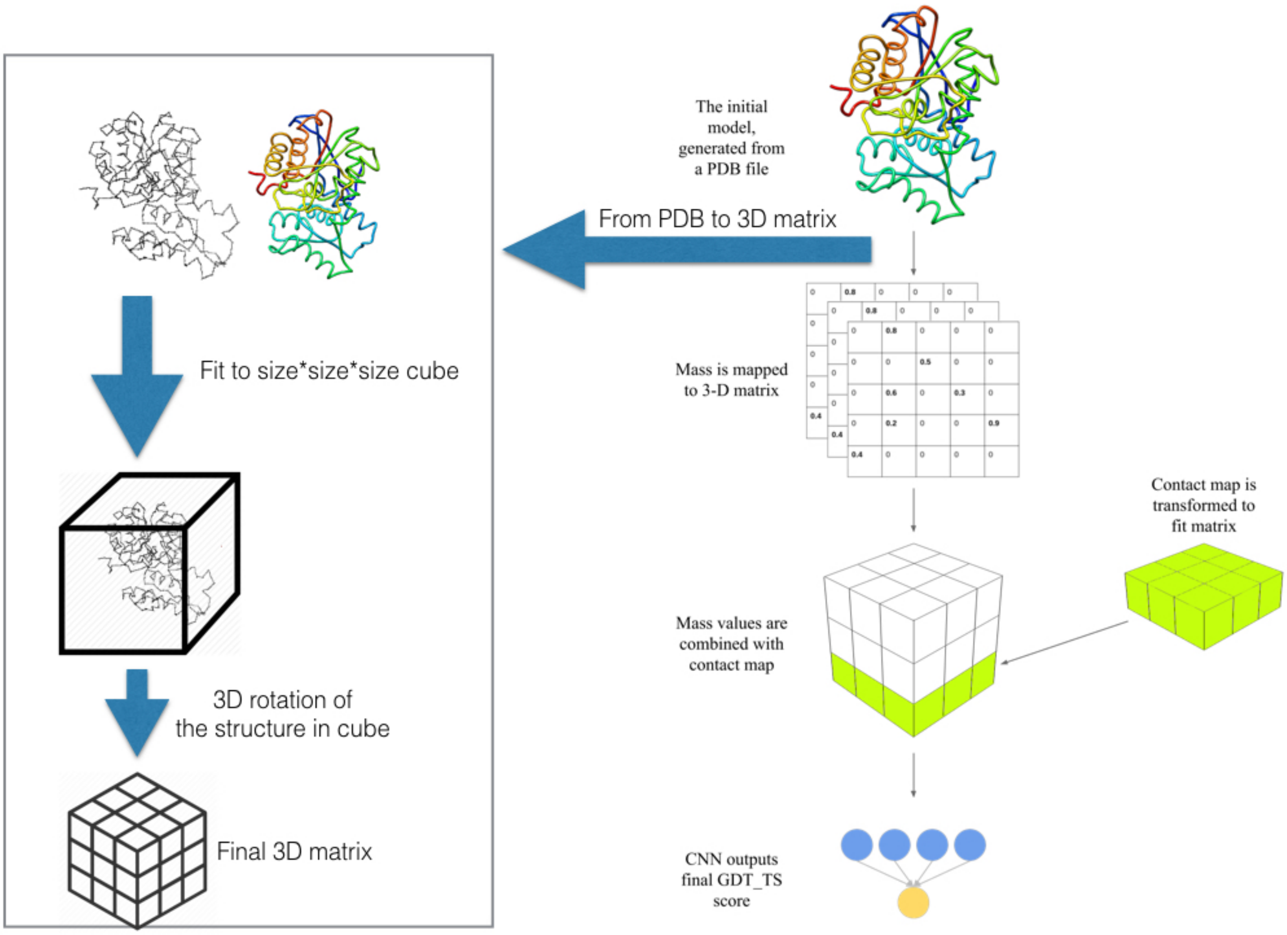

The initial model, generated from a PDB file
From PDB to 3D matrix
Fit to size*size*size cube
Mass is mapped to 3-D matrix
Contact map is transformed to fit matrix
Mass values are combined with contact map
3D rotation of the structure in cube
Final 3D matrix
CNN outputs final GDT_TS score

Figure 1. Flowchart of TopQA method

Availability

The software is freely available at GitHub: https://github.com/caorenzhi/TopQA.

1. Smith, J., Conover, M., Stephenson, N., Eickholt, J., Si, D., Sun, M. & Cao, R. (2020). TopQA: a topological representation for single-model protein quality assessment with machine learning. International Journal of Computational Biology and Drug Design 13, 144
2. Uziela, K., Menéndez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. Bioinformatics 33, 1578−1580
3. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. Sci. Rep. 6, 23990
4. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 17, 495
5. Manavalan, B. & Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 33, 2496–2503
6. Olechnovič, K. & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins 85, 1131−1145
7. Derevyanko, G., Grudinin, S., Bengio, Y. & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. Bioinformatics doi:10.1093/bioinformatics/bty494
8. Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A. (2016). Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 84 Suppl 1, 349−369
9. Si, D., Moritz, S.A., Pfab, J., Hou, J., Cao, R., Wang, L., Wu, T. & Cheng, J. (2020). Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. Sci. Rep. 10, 4282

CAO-SERVER(Topology)

de novo protein structure prediction using stepwise fragment sampling with contact prediction and model selection based on deep learning techniques

Kyle Hippe[1], Mikhail Korovnik[1], Renzhi Cao[1*]
1 - Pacific Lutheran University
caora@plu.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:Y.1-20; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP 14, we blindly tested our new de novo protein structure prediction pipeline. Instead of randomly sampling protein conformation space, this method uses stepwise fragment sampling as it is more efficient and accurate [1,2]. Contact information is also incorporated in our pipeline, as contact prediction played an important role in structure modeling in the recent CASP experiments [3–6]. Finally, deep learning techniques are used for selecting 5 models as the final prediction of our method [7].

Methods

Step 1, contact prediction is made for each protein sequence. We used the latest version of MetaPSICOV2 [3] to make contact prediction from the input protein sequence. We would like to mention that MetaPSICOV2 may fail occasionally, in this case, we use the alternative contact prediction from CCMpred and FreeContact [8,9].

Step 2, after the contact prediction was done, a request was sent to all connected computers for united-residue conformational search via stepwise and probabilistic sampling with the help of Unicon3D tool[1]. The secondary structure prediction and contact prediction from the previous step was used in Unicon3D for de novo protein structure prediction.

Step 3, compared to random sampling like Monte-Carlo search, sequential search turned out to be more efficient and accurate. Our server did sequential protein conformational search with the help of SAINT2 tool [2]. The fragment used in this step was generated by a modified version of FRAGSION tool [10], which is ultra-fast and accurate in fragment generation based on a Hidden Markov Model. Because of computational resource limitations, we only generated fragments with size 8 and 12. The contact prediction from the first step was also used to guide the protein structure prediction process.

Step 4, model selection from thousands of protein decoys is crucial in protein structure prediction. Qprob[11] is a super-fast tool to rank all decoys based on the model quality, so we selected the top 100 decoys based on Qprob's ranking. After that, we use a deep learning-based tool, DeepQA [7], with the help of clustering for diversity [12] to select 5 models as our final prediction.

Availability

The Cao-server is available at the following link:
https://www.cs.plu.edu/~caora/index.php/Cao_server/

1. Bhattacharya, D., Cao, R. & Cheng, J. (2016). UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics 32, 2791–2799
2. de Oliveira, S.H.P., Law, E.C., Shi, J. & Deane, C.M. (2018). Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. Bioinformatics 34, 1132–1140
3. Buchan, D.W.A. & Jones, D.T. (2017). Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins: Struct. Funct. Bioinf. 86, 78–83
4. Adhikari, B., Hou, J. & Cheng, J. (2018). Protein contact prediction by integrating deep multiple sequence alignments, coevolution and machine learning. Proteins 86 Suppl 1, 84–96
5. Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. & Bonvin, A.M.J.J. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins 86 Suppl 1, 51–66
6. Hou, J., Wu, T., Cao, R. & Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins 87, 1165–1178
7. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 17, 495
8. Seemayer, S., Gruber, M. & Söding, J. (2014). CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 30, 3128–3130
9. Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B. (2014). FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15, 85
10. Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2016). FRAGSION: ultra-fast protein fragment library generation by IOHMM sampling. Bioinformatics 32, 2059–2061
11. Cao, R. & Cheng, J. (2016). Protein single-model quality assessment by feature-based probability density functions. Sci. Rep. 6, 23990
12. Cao, R., Bhattacharya, D., Adhikari, B., Li, J. & Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. Bioinformatics 31, i116–i123

ClusPro

Hybrid ClusPro server in 2020 CASP/CAPRI rounds

Dima Kozakov[1], Sandor Vajda[2,3], Kathryn Porter[2], Dzmitry Padhorny[1], Israel Desta[2],
Dmitri Beglov[2], Mikhail Ignatov[1], Sergey Kotelnikov[1]
[1]Laufer Center for Physical and Quantitative Biology, Stony Brook University; Departments of [2]Biomedical Engineering [3]Chemistry, Boston University

The original ClusPro server performs rigid body docking using the PIPER program and clusters the 1000 lowest energy structures. The models are ranked according to cluster size. In order to deliver results to the user within 24 hours of submission, the current implementation of ClusPro does not include refinement beyond minimizing the energy of structures to remove steric overlaps. In spite of this limitation, the server has almost 7800 registered users, and run about 200,000 jobs in the last 3 years. In the recent years we have enhanced ClusPro with capabilities of accounting for additional information to restrain the search, including SAXS data and XL-MS cross-links.
In the latest rounds of the CASP-CAPRI experiment we have expanded the ClusPro server to use template-based information when available. Based on the target sequence we identify structures that can serve as templates for the complex, and perform homology modeling based on the biological units of the templates. If no template is available, we perform free docking as described above. The server has the option of accepting pre-selected templates as input. In addition, we explore the option of further refining and validating template-based models with free docking.

Methods.

Model preparation. Based on the sequence of the target, we automatically detect available templates using HHPred, and identify those that contain homologs of the interacting biological unit to be predicted. If no template of the complex is found, we suggest to perform free docking. Since free docking by ClusPro requires three-dimensional structures as the input, we either use the HHPRED top template or in difficult cases build an "ab initio" model of the subunit using TrRosetta. For each "easy" target most models had the same fold, with variations in loops and tails. Removal of the uncertain regions resulted in reliable "consensus" models that were used for docking.

Template based docking. If a template of the biological complex satisfying the requird stoichiometry is found then we chose the best template for each unique monomer of the complex, align multiple copies of this monomer template to the complex template and then model the whole complex using Modeller. Per rules of CAPRI we generate up to 10 models.

Free Docking. Our free docking approach consists of two steps. The first step is running PIPER, a docking program that performs systematic search of complex conformations on a grid using the fast Fourier transform (FFT) correlation approach. The scoring function includes van der Waals interaction energy, an electrostatic energy term, and desolvation contributions calculated by a pairwise potential.

The second step of the algorithm is clustering the top 1000 structures generated by PIPER using pairwise RMSD as the distance measure. The radius used in clustering is defined in terms of C_α interface RMSD. For each docked conformation we select the residues of the ligand that have any atom within 10 Å of any receptor atom, and calculate the C_α RMSD for these residues from the same residues in all other 999 ligands. Thus, clustering 1000 docked conformations involves computing a 1000 × 1000 matrix of pairwise C_α RMSD values. Based on the number of structures that a ligand has within a (default) cluster radius of 9 Å RMSD, we select the largest cluster and rank its cluster center as number one. The members of this cluster are removed from the matrix, and we select the next largest cluster and rank its center as number two, and so on. After clustering with this hierarchical approach, the ranked complexes are subjected to a straightforward (300 step and fixed backbone) van der Waals minimization using the CHARMM potential to remove potential side chain clashes. ClusPro outputs the centers of the 10 largest clusters, which were submitted as predictions.

CMH1971

Structure Prediction, Quality Assessment and Contact Prediction by EMAP_CLUST

Myong-Ho Chae[1]
[1]- *Department of Life Science, University of Sciences, Unjong-District, Pyongyang, DPR Korea*
chae1971@star-co.net.kp

EMAP_CLUST is a consensus-based QA method to predict local as well as global quality of protein models. We submitted models in three categories(TS, QA, RR) of tertiary structure prediction to CASP14.

Methods

1. QA Quality Assessment

All server models of a target protein submitted to CASP14 are ranked according to their EMAP global scores(see our EMAP abstracts), and a reference model set is constructed from top-scoring N models. Then, the pair-wise similarity score is computed between each model and all models of the reference set using TMscore[1] to produce N GDT_TS scores. The consensus-based global quality score is the EMAP-weighted mean of N GDT_TS scores. For local score, The N C_α distances (d) between the corresponding residues of a model and reference models, are computed using TMscore[1]. The distance is converted to the S-score with distance threshold d_0=3.8Å, S =$1/(1+(d/d_0)^2)$. Next, the EMAP-weighted mean (S_Weight) of N S-scores is calculated. The per residue distance deviation(Å) is calculated from S_Weight, L =min(d_0 (1/ S_Weight - 1)$^{1/2}$, 15). EMAP_CLUST was applied to the stage1 and stage2 dataset of CASP14. The size of reference model pool N, was set to 11 for stage1, and 21 for stage 2.

2. TS Regular targets for structure prediction

CASP14 stage 2 server models were evaluated by EMAP_CLUST,and the top model was selected. Scores in B-factor column were replaced by the residue CA errors from EMAP_CLUST.

3. RR Contact Prediction

Our residue-residue contact prediction method is based on the consensus of CASP14 RR contact prediction server models. First, CASP14 contact prediction server models are pre-processed(short-range contact predictions are removed and top 3L predictions are selected). Then these models are evaluated using DOOP residue-level contact pair potential[2] and 20 top-scoring models are selected. The probability scores of corresponding residue pairs in selected prediction models are summed up and rescaled.

Results

We evaluated EMAP_CLUST on CASP13 QA dataset and proved that it achieves comparable performance with the state-of- the-art QA methods.

1. Zhang,Y. & Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
2. Chae,M.H., Krull,F. & Knapp,E.W., (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction, Proteins. 83, 881−890.

CUTSP

Morphing semi-supervised protein structures predicted using distance and torsion representations with deep graph ranking

I.Drori[1], X.Ji[1], Z.Fan[1], A.G.Kharkar[1]
1- Columbia University
idrori@cs.columbia.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:Y.v; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

CUTSP CASP14 submissions were all generated by morphing predicted structures and ranking the results. We used the same methods for all predictions as well as protein docking for complexes.

Methods
We first generate multiple sequence alignments (MSAs) using HHblits[1] with UniRef[2]. Next, we use both supervised and semi-supervised approaches based on distance and torsion angle representations for predicting diverse protein structures[3,4,5]. We then morph between these structures taking into account the energy of the conformation[6]. The morphing is non-linear and allows to bypass high energy conformation barriers. We superimpose the structures[7] onto a base structure and select the top candidates.
Scoring

We rank the morphed structures using a deep neural network trained to predict quality based on previous CASPs and a graph neural network predicting quality of full-atom graph protein representations[8,9].
Docking

We perform docking of proteins with multiple chains. First, we predict the conformation of each chain and then use rigid-body protein docking[10,11] to generate a candidate set of complexes. Finally, we rank the complexes based on their energy score, and select the top candidates.

Availability
We will make our pipeline available upon publication.

1. Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2012) HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 9(2), 173–175.
2. Suzek, B., Wang, Y., Huang, H., McGarvey, P., & Wu, C. (2014) UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6), 926-932.
3. Rao, R. et al. (2019) Evaluating protein transfer learning with TAPE. In Advances in Neural Information Processing Systems, 9689–9701.
4. Yang, J. et al. (2020) Improved protein structure prediction using predicted interresidue orientations. Proceedings of the Notational Academy Sciences 117, 1496–1503.
5. Drori, I. et al. (2019) Accurate protein structure prediction by embeddings and deep learning representations, Machine Learning in Computational Biology.
6. Weiss, D. R. & Levitt, M. (2009) Can morphing methods predict intermediate structures? J. Molecular Biology 385, 665–674.

7. DeLano, Warren L et al. (2002) PyMOL: An open-source molecular graphics tool, CCP4 Newsletter on protein crystallography 40(1), 82–92.
8. Hurtado, D. M., Uziela, K. & Elofsson, A. (2018) Deep transfer learning in the assessment of the quality of protein models.
9. Sanyal, S., Anishchenko, I., Dagar, A., Baker, D. & Talukdar, P. (2020) ProteinGCN: Protein model quality assessment using graph convolutional networks.
10. Schindler, C.E.M., de Beauchêne, I.C., de Vries, S., Zacharias, M. (2017) Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI. Proteins 85(3), 391–398.
11. Eismann, S., Townshend, R., Thomas, N., Jagota, M., Jing, B., Dror, R. (2020) Hierarchical, rotation-equivariant neural networks to predict the structure of protein complexes.

Template-based Structure Prediction and Interresidue Distances and Orientations Prediction-based Structure Prediction

Tsukasa Nakamura[1], Yuya Hanazono[2]

[1] – Tohoku University, Japan, [2] - National Institutes for Quantum and Radiological Science and Technology, Japan
t.nakamura@sb.ecei.tohoku.ac.jp

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

In CASP14, we used a multiple sequence alignment (MSA) generated by our method[1] as a seed input for HHblits[2] to a perform profile–profile sequence search, and we also used the template profile database created in a similar method. To construct 3D-models, we used template-based structure prediction by MODELLER[3], interresidue distances and orientations prediction-based structure prediction by trRosetta[4], and combined them in some targets. In addition, we predicted quaternary structures that replicate experimental evidence based on a literature search.

Methods

To execute a sequence search of a target, we used SSearch[5] with MIQS[6] against the latest NCBI nr database. Then we made an MSA by using MPI-parallelized MAFFT[7,8] with homologous sequences. With the MSA as input, we used HHblits to execute an iterative profile–profile sequence search against the UniClust30[9] and BFD[10] databases.

To execute a template search and acquire profile–profile alignment between target and templates, we used HHsearch[2] against the latest PDB70 and an in-house profile database that was made by three iterations of HHblits with MSAs as input. These MSAs were made with PDB98 against NCBI nr in a similar manner for target sequences. However, we made the MSAs partly by stacking pairwise sequence alignments by SSearch instead of using MPI-parallelized MAFFT.

In our 3D-model construction step, we used MODELLER with the result of the profile–profile alignment against PDBs and trRosetta with the result of the sequence search. We intervened in the processes of trRosetta by partly substituting the input with the distances and orientations of 3D-models made by MODELLER in some targets that had good templates and made 3D-models well.

In our model selection step, we used VoroMQA[11] mainly, dDFire[12], ProQ4[13], and the rate of fit with the servers' distance predictions.

For multimeric targets, the stoichiometry of the template protein was considered to select a model. Also, experimental evidence (e.g., the number of disulfide bonds by mass spectrometry and interacting regions by pull-down assay) based on a literature search was heavily considered and we tried to replicate the evidence in 3D-models by adding restraints manually. If we needed to perform free-docking, we used Haddock[14] and ZDOCK[15]. If we considered that the target must be coiled-coil but it was hard to construct a model, we used ISAMBARD[16].

1. Nakamura,T., Oda,T., Fukasawa,Y., and Tomii,K. (2018) Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments. Proteins Struct. Funct. Bioinforma., 86, 274–282.
2. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J., and Söding,J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics, 20.
3. Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinforma., 2016, 5.6.1-5.6.37.
4. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S., and Baker,D. (2020) Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A., 117, 1496–1503.
5. Pearson,W. (2003) Finding Protein and Nucleotide Similarities with FASTA . Curr. Protoc. Bioinforma., 4.
6. Yamada,K. and Tomii,K. (2014) Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics, 30, 317–325.
7. Nakamura,T., Yamada,K.D., Tomii,K., and Katoh,K. (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics, 34, 2490–2492.
8. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol., 30, 772–780.
9. Mirdita,M., Von Den Driesch,L., Galiez,C., Martin,M.J., Soding,J., and Steinegger,M. (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res., 45, D170–D176.
10. Steinegger,M., Mirdita,M., and Söding,J. (2019) Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat. Methods, 16, 603–606.
11. Olechnovič,K. and Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins Struct. Funct. Bioinforma., 85, 1131–1145.
12. Yang,Y. and Zhou,Y. (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins-Structure Funct. Bioinforma., 72, 793–803.
13. Hurtado,D.M., Uziela,K., and Elofsson,A. (2018) Deep transfer learning in the assessment of the quality of protein models. arXiv.
14. Dominguez,C., Boelens,R., and Bonvin,A.M.J.J. (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc., 125, 1731–1737.
15. Chen,R., Li,L., and Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. Proteins Struct. Funct. Genet., 52, 80–87.
16. Wood,C.W. et al. (2017) ISAMBARD: An open-source computational environment for biomolecular analysis, modelling and design. Bioinformatics, 33, 3043–3050.

DeepML

Quality Assesment of Protein Models using Graph Convolutional Networks

Soumyadip Roy[1], Asa Ben-Hur[1]
[1] -*Colorado State University*
soumya16@colostate.edu, asa@cs.colostate.edu

Protein model quality assessment is an important problem. There have been many algorithms proposed for this task, including deep learning methods that use 3D convolution[1]. This showed the promise of deep learning architectures for this problem. We decided to go with Graph Convolutional Networks[2] (GCNs) which have not been used for this task to the best of our knowledge. Proteins can be considered as graphs with the atoms as nodes. GCNs are a very powerful neural network architecture which can produce useful feature representations of nodes in networks. Therefore, we hypothesize that GCNs can learn the features that help discriminate decoys from near native models.

Methods

In our method we considered each protein as a graph with the atoms as nodes. We then applied multiple layers of graph convolution over atom-level features, followed by a couple of dense layers.

Our aim is to predict a score for the entire protein that reflects the Global Distance Test Total Score (GDTTS) [3] of a model with respect to its native structure. In other words, we use GDTTS scores as our ground truth labels. In later work we extended this approach to predict residue level GDTTS training along with some residue level features which significantly improved performance.

Results

We trained our model on CASP 11 and CASP 12 datasets which consisted of over 200 targets in total and tested on CASP 13 datasets consisting of 143 targets. We obtained a Pearson rank correlation of 0.61. Our more advanced models, which were not ready in time for CASP 14, yielded improved accuracy with a rank correlation of 0.83.

1. Pagès, G., Charmettant, B., and Grudinin, S., Protein model quality assessment using 3D oriented convolutional neural network.
2. Fout, A., Byrd, J., Shariat, B., Ben-Hur, A., Protein Interface Prediction using Graph Convolutional Networks.
3. Yuanpeng H., Mao,B., Aramini, J., and Montelione, G., Assessment of template based protein structure predictions in CASP10.

DeepMUSICS

A novel deep learning framework for protein structure prediction

Y. Zhang[1], F. Pan[1], C. Lo[1], X. Liu[2], X. Pang[3], and J. Zhang[1, *]
[1]*Department of Statistics,* [2]*Department of Computer Science, Florida State University*
[3]*Insilicom LLC, Tallahassee, FL*
jinfeng@stat.fsu.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

In this CASP competition, we participated in the protein refinement category using a deep learning based method. We used two types of deep learning models for sampling conformations and evaluating their qualities. To sample conformations, we used a cut-regrow scheme with sequential importance sampling, where fragments of 3-10 residues are cut and regrown one torsion angle at a time. The torsion angles at each step are predicted by deep learning based torsion angle prediction models. After a fragment is regrown, the quality of the conformation is evaluated by a deep learning based energy function model. The cut-regrow is done many times on low quality regions of the conformation using Metropolis-Hastings algorithm. The resultant conformations are further refined by Molecular Dynamics simulations and then ranked by another energy function model before submission. The whole method is named DeepMUSICS (Deep learning powered MUlti-scale Sequential Importance Conformation Sampling).

Methods

We designed a set of torsion prediction models categorized by the length of the fragment to be grown and the type of torsion angles (phi/psi). The models consist of a series of residual neural network (ResNet) blocks, and each ResNet block contains two layers of three-dimensional convolutional neural networks (CNN). The input was two 3D gridded boxes, with the atomic coordinates and types, around the growing site, as well as the sequence of the fragment. The boxes, with different sizes and resolutions, were fed into the ResNet structure to capture the structural environment of the fragment to be grown, and the sequences were fed into a set of dense layers. The output layer consists of 360 SoftMax nodes representing the probability in each angle bins (1°).

Two energy function models, MODEL_E1 and MODEL_E2, were trained to predict the GDT_HA score of the fragments sampled by the cut-regrow approach. The input was the 3D conformation of the fragments captured by a series of 3D gridded boxes centered at each residue with the atomic coordinates and types. All the boxes were fed into four ResNet blocks, and each block consists of two CNN layers. The output vectors were fed into a bidirectional long short-term memory (LSTM) network to predict the GDT_HA score. The GDT_HA of the whole structure was the average score of the fragments sliding through the whole sequence.

For the refinement process, we chose two regions from the initial conformations of target proteins to refine by the cut-regrow scheme. The selections were decided by the averaged residue-wised GDT_HA score predicted by MODEL_E1 and MODEL_E2, where two regions with lowest GDT_HA scores were chosen. The lengths of regions vary from 20 to 40 residues as different targets.

The cut-regrow process took only phi/psi torsion angles as variables, which were sampled based on probability distributions predicted by the torsion prediction models. The bond lengths and bond angles were fixed, and omega torsion was sampled around 180 degrees (with only PRO has a small chance to be 0°). Analytical closure was performed when the growth length was within three residues. The start and end residues for one cut-regrow iteration were randomly chosen within the two regions above, with length up to 10. After the torsion sampling and growth, MODEL_E1 was used to select the grown conformation with the highest GDT_HA, and acceptance was determined by standard Metropolis-Hastings criterion. A simulated annealing algorithm was also applied on top of Metropolis-Hastings algorithm to increase the chance of finding structures with better energies (scores from the deep learning models).

For one single target, five independent refinement runs were carried out, with each one having up to 2000 iterations. From each run, we extracted one conformation with lowest GDT_HA score predicted by MODEL_E1 and MODEL_E2. Using AMBER18[1] package, the side chains were added and modeled by Molecular Dynamics simulations with FF99SB[2] forcefield, including minimization, heating, and equilibrium runs. Restraint was added to the backbone atoms throughout the simulations. Finally, the five resulting conformations after MD were ranked by MODEL_E2 for submission.

Availability

The source codes and models are not publicly available at the moment.

1. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, P.A. Kollman. (2018) AMBER 2018. University of California, San Francisco.
2. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. and Simmerling, C. (2006), Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins, 65: 712-725. doi:10.1002/prot.21123

DeepPotential

Learning deep statistical potentials for protein folding

Yang Li[1, 2], Chengxin Zhang[1], Wei Zheng[1], Xiaogen Zhou[1], Eric W. Bell[1], Dong-Jun Yu[2], and Yang Zhang[1]

[1] - Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109, [2] - School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094;

yangzhanglab@umich.edu

***Key:** Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y*

DeepPotential makes use of deep learning based predictions as statistical potentials for protein folding. A multi-threshold strategy is applied to those prediction terms to capture folding knowledge at different levels.

Methods

The input features contain two-dimensional and one-dimensional features extracted from Multiple Sequence Alignments (MSAs). Here, two-dimensional features are mainly raw coevolutionary features, i.e., pseudolikelihood maximation of Potts model and Mutual information matrix, and their post-processing additionally; one-dimensional features are the single-site features, including one-hot sequence representation, HMM features and single-body parameters of Potts model and Mutual information matrix. DeepPotential neural networks are trained to predict terms that are critical for protein folding, i.e., Cα-Cα and Cβ-Cβ distances, inter-residue torsion angles and H-bond related geometry descriptors.

Given a query sequence, s set of candidates MSAs are built by searching against different sequence databases (Uniclust30, UniRef90, BFD, Mgnify and IMG/M) with different searching tools (HHblits, Jackhmmer and HMMsearch). Optimal MSAs are selected by the summation of the cumulative probability under 8Å (12Å for TripletRes server group) of top 10*L predicted Cβ-Cβ distance distributions for all residue pairs. DeepPotential predicts distance distribution with multiple thresholds (from 2Å to 10Å, 13Å, 16Å and 20Å). The final contact/distance prediction of DeepPotential combines distributions for all thresholds.

The distance distribution with threshold equaling to 20Å will be considered as the base distribution. A sequential combination strategy is used by replacing specific distance regions 2-*t*Å in base distribution with the corresponding distance distributions at thresholds of 16Å, 13Å and 10Å sequentially, if P(*d*>*t*Å) < 0.5. Here *d* is the distance of a residue pair and *t* is the corresponding threshold. At each iteration over thresholds, the distance distribution will be normalized to guarantee that the summation of probabilities equals to 1. The negative log of multi-threshold distance distribution and orientation distribution will be smoothed by a cubic spline to smooth potentials so that it can be optimized by gradient-descent based methods, e.g., L-BFGS implemented by the PyRosetta package[1]. The tertiary structure construction for a query sequence starts with a random structure and is optimized by repeated L-BFGS. At each iteration, DeepPotential adds random noises in torsion angles space to the structure from the previous iteration and continue the optimization. The decoy with lower energy value will be kept. FASPR[2] and FG-MD[3] are used for side-chain packing and local structure refinement after the optimization.

The predicted terms are also used for the estimation of model accuracy (EMA) by deep residual neural networks. In addition to the distance and torsion-angle terms used for differentiable decoy scoring, DeepPotential also feeds the neural networks with H-bond geometry terms[4]. For a query decoy, the Cα-Cα and Cβ-Cβ distance maps, torsional angle maps, H-bond geometry map, and their corresponding predicted probability likelihood maps together with the negative log of probability likelihood maps are also used as input features. The neural network outputs 3 types of error estimations, i.e., residue-pair distance error estimation (2D), residue-wise alignment error estimation (1D) and GDT-TS score estimation (scalar). The two-dimensional signals are reduced to one-dimension by mean operation in multiple ranges. The one-dimensional signals are averaged along the sequence length dimension and fed into a set of fully connected layers to predict the GDT-TS score. The 3 types of error estimation are trained jointly. The EMA prediction model was trained during the CASP14 season, so it was not used in tertiary structure prediction or selection.

1. Chaudhury, Sidhartha, Sergey Lyskov, and Jeffrey J. Gray. "PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta." Bioinformatics 26, no. 5 (2010): 689-691.
2. Huang, Xiaoqiang, Robin Pearce, and Yang Zhang. "FASPR: an open-source tool for fast and accurate protein side-chain packing." Bioinformatics (2020).
3. Feig, Michael. "Local protein structure refinement via molecular dynamics simulations with locPREFMD." Journal of chemical information and modeling 56, no. 7 (2016): 1304-1312.
4. Yang, Jianyi, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. "The I-TASSER Suite: protein structure and function prediction." Nature methods 12, no. 1 (2015): 7-8.

DELCLAB

Carlos, Del Carpio M.A.[1], Kawasumi, Mizuho-cho[1]

1. Graduate School of Medical Sciences. Doctoral Program in Biodefense. Nagoya City University, Mizuho-ku, Nagoya 467-8601, Japan

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:N; MetaG:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y.*

Our group has been involved in the development of several basic algorithms for the prediction of the secondary, tertiary and quaternary structures of bio-macromolecules including oligopeptides, proteins, and RNA's[1,2]. To predict the structure of proteins in past rounds of CASP we have combined classical homology methods with our genuine method based on spectral analysis of the sequences of the amino acids represented by their physicochemical properties. The methodology resulted in high accuracy of the prediction of folding patterns namely in the so-called twilight zone of sequence homology (20&#8764;30% of similarity), where the prediction of protein 3D structure based only on sequence homology methodologies are frequently of limited success. In CASP13 this methodology proved effective for several targets, namely in the prediction of particular domains that characterized those molecules. Furthermore, to improve the structure of loops in protein structures we have developed a new automatic system based on a genuine idea about protein stability.

Protein quaternary structure has been handled using our system for the assessment of complex structures MIAX[4], the main characteristics of which consist of the prediction of binding sites and a new protocol for the evaluation of the plausibility of contact regions.

In CASP14 we have constructed a multi-platform system based on all these methodologies and treated each problem in a systematic way that has enhanced the predictability of the tertiary structure of CASP targets and their quaternary structure when required.

Methods

The multi-platform automatic system proposed starts with the selection of the best homologs for the sequence in question with orthodox methodologies. When no homologs are found for the target, the process shifts to the spectral analysis of the sequences and homologs from this point of view are output that is analyzed in a piece-wise manner with the target sequence. Then the required 3D sequence for the target structure is built by the platform. Loop and structural stability analysis is then carried out with our system for protein stability analysis. Molecular dynamics and other minimization processes are then applied to the most plausible candidate structures which are then ranked according to the energetic characteristics.

On the other hand, protein assemblies are predicted using the system MIAX[3] for protein interaction assessment, which consists of protein interaction region prediction and docking of the structures. For hetero multimer structure prediction, prediction of the binding sites was performed based on a new way to assess the order of interaction of the subunits[4].

Results

Loop flexibility analysis and the consideration of the order of interaction of complexes, extensively used in CASP14, has led to a deeper insight into the way protein folding as well as complex formation occurs and the appropriate computational methodology to deal with the problem.

1. Del Carpio, C. A. & Yoshimori, A. (2002). Fully automated protein tertiary structure prediction using Fourier transform spectral methods. Protein Structure Prediction: Bioinformatics, University of California, International University Line.
2. Del Carpio, C. A. & Carbajal, J. C. (2002). Folding pattern recognition in proteins using spectral analysis methods. Genome Inform 13, 163-72.
3. Del Carpio, C.A., Ichiishi E. (2017). Inference of Protein Multimeric Complex Dynamic Order of Formation: An Active Region Recognition Based Approach. International Journal of Genomics and Data Mining 2017, 1.
4. Del Carpio, C. A., Ichiishi, E., Yoshimori, A. & Yoshikawa, T. (2002). A new paradigm for modeling biomacromolecular interactions and complex formation in condensed pahses. Proteins: Structure, Function, and Genetics 48, 696-732.

DellaCorteLab

Refinement with Improved Restrained Molecular Dynamics

Connor J Morris, Wendy M Billings, Dennis Della Corte
Dept. of Physics and Astronomy, Brigham Young University
dennis.dellacorte@byu.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y.*

A team of undergraduate students at DellaCorte Lab built a refinement protocol based on previous MD (molecular dynamics)-based protocols, particularly the one used by Feig Lab in CASP13. We used physiological salt concentrations, additional equilibrations, and a larger radius of flat-bottom restraints to improve refinement while using fewer iterations of MD simulation.

Methods

Targets were subjected to five 100 ns MD simulations using flat-bottom harmonic restraints on the C-alpha atoms. The flat-bottom harmonic restraint allowed unrestrained movement of each C-alpha atom in a radius of 5 Å before restraints restricted further deviation from the starting conformation. RWPlus[1] was used to score frames extracted every 20 ps from all trajectories. Frames were then ranked according to RWPlus score and averaged structures were generated from the top 1%, 5%, 15%, and 40% of the trajectory frames. SCWRL4[2] was used to optimize the side chains of the averaged structures, then each was subjected to an energy minimization with harmonic restraints on all heavy atoms. Model 1 was the 15% averaged structure, followed by 5%, 40%, and 1% averaged structures. The initial target structure was submitted as model 5. Residue-wise error was estimated by calculating root mean square fluctuation (RMSF) values on C-alpha atoms in the MD trajectories.

The only deviation from this protocol was when a low GDT-HA score or other factor suggested the experimental structure deviated by > 5 Å from the start structure. In this case, either unrestrained MD or an additional iteration of restrained MD was added and averaged structures from those trajectories were included in the submitted models. This was done on only 6 of 50 targets we submitted models for.

Differences between our protocol and the Feig Lab protocol in CASP13 include: physiological salt concentrations in MD, NPT equilibration before MD simulation, larger radius before flat-bottom restraints begin (5 Å vs. 4 Å), and maintaining normal hydrogen masses. Since Heo et al[3], found after CASP13 that multiple iterations of MD and the specific choice of scoring function had little effect on results, we removed the iterative rounds of MD simulation to reduce simulation time and eliminate the need to build Markov-state models and replaced Rosetta energy scoring[4] with RWPlus.

MD was performed with OpenMM[5] using explicit solvent and physiological salt concentrations. An energy minimization and NPT equilibration preceded MD simulations. GROMACS6 was used to add hydrogen atoms to the target structure prior to MD simulations, to generate an averaged structure from the top scoring MD frames, and to calculate RMSF values from MD trajectories for residue-wise error estimation.

Acknowledgments

We thank the Office of Research Computing at Brigham Young University for compute time on the BYU supercomputer, Mary Lou.

1. Zhang J. & Zhang Y. (2010). A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. PloS one 5, e15386.
2. Krivov, G. G., Shapovalov, M. V., & Dunbrack Jr, R. L. (2009). Improved prediction of protein side&#8208;chain conformations with SCWRL4. Proteins 77, 778-795.
3. Heo, L., Arbour, C. F., & Feig, M. (2019). Driven to near experimental accuracy by refinement via molecular dynamics simulations. Proteins 87, 1263-1275.
4. Park, H., Bradley, P., Greisen Jr, P., Liu, Y., Mulligan, V. K., Kim, D. E., Baker, D. & DiMaio, F. (2016). Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J. Chem. Theory Comput. 12, 6201-6212.
5. Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L., Simmonett, A.C., Harrigan, M.P., Stern, C.D., Wiewiora, R.P., Brooks, B.R., & Pande, V.S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol., 13, e1005659.
6. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX, 1, 19-25.

DellaCorteLab

De novo structure prediction with deep learning and molecular mechanics simulation

Connor J Morris, Wendy M Billings, Dennis Della Corte
Dept. of Physics and Astronomy, Brigham Young University
dennis.dellacorte@byu.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y.*

DellaCorte Lab – a team of undergraduate students – combined deep learning inter-residue distance prediction, gradient descent protein reconstruction, and molecular mechanics-based structure refinement for all submissions of CASP14 structure prediction. Heteromeric targets were assembled with protein-protein docking tools.

Methods

All submissions followed the same protocol, independent of availability of homology models. We first generated distance predictions from multiple sequence alignments (generated with HHBlits[1]), using our ProSPr distance prediction network[2] and trRosetta[3]. After manual investigation of the distance predictions, we used pyRosetta-based structure optimization to fold 100 models of an alanine chain according to the distance predictions. We mutated the final sequence using the Dunbrack rotamer library[4] to match the target sequence and performed local optimizations with pyRosetta minmover. The final structures were ranked by the Rosetta Energy Function and the top 10 were manually investigated.

Based on the similarity and quality of the models we selected one or two of the top 3 models for molecular dynamics-based refinement simulation, according to the same protocol that we describe in the refinement abstract. The differences in the protocol used on refinement targets and the protocol employed here are a reduced number of MD simulations (3 instead of 5) to reduce simulation time and an altered model submission order (5% averaged structures, 15%, 40%, 1%, pre-MD model) to put a more aggressive trajectory average as model 1. In cases of multiple comparably scored, but structurally different (RMSD > 5 Å) reconstructions, we started additional refinement simulations from different start structures and adjusted the submission order to also contain models derived from the other trajectories.

For heteromeric targets, we deviated from the protocol based on availability of homology models or solved structures. Each protein chain was folded separately and afterwards subjected to protein-protein docking with Interevdock[5]. Multiple targets required manual intervention to achieve reasonable poses.

Acknowledgements

We thank the Office of Research Computing at Brigham Young University for compute time and the local supercomputer Marylou. We also express gratitude for helpful discussions with David Wingate, Tommy Knotts, and Joshua Price at BYU and Michael Feig at MSU. We also acknowledge the contributions of other DellaCorte Lab members not directly involved in CASP14.

1. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173-175.

2. Billings, W. M., Hedelius, B., Millecam, T., Wingate, D., & Della Corte, D. (2019). ProSPr: Democratized Implementation of AlphaFold Protein Distance Prediction Network. BioRxiv 830273.
3. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. USA 117, 1496-1503.
4. Shapovalov, M. V., & Dunbrack Jr, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19, 844-858.
5. Quignot, C., Rey, J., Yu, J., Tufféry, P., Guerois, R., & Andreani, J. (2018). InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. Nucleic Acids Res. 46, W408-W416.

ProSPr: Protein Structure Prediction via Inter-Residue Distances

Connor J Morris, Wendy M Billings, Dennis Della Corte
Dept. of Physics and Astronomy, Brigham Young University
dennis.dellacorte@byu.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N.*

A team of undergraduates in the Della Corte lab implemented and trained a deep convolutional neural network (ProSPr) to predict inter-residue distance probabilities for all residue pairs. Several data augmentation strategies were employed to increase the effective training set size, and predictions were made directly into the 10 distance bin ranges specified for CASP14.

Methods

Training data were collected from the CATH s35 database,[1] resulting in sequences and structure labels for about 27k nonredundant protein domains. Multiple sequence alignments (MSAs) were generated for each sequence using both PSIBLAST[2] on the nr database, as well as HHBlits[3,4] with the Unliclust30 database.[5] The PSIBLAST PSSM contributed 2D features to the input vector. HHBlits-aligned sequences were randomly subsampled during each training epoch, which subset was then used to calculate HHM profile 2D features and 3D pair coupling information using the inverse of the shrunk covariance matrix as described previously.[6] These MSA features were combined with layers for one-hot residue encodings as well as sequence position indicators to give the full input vector.

The bulk of the ProSPr architecture was a series of 220 ResNet[7] blocks each containing projections, batchnorms, elu activations, and a 3x3 convolution with varying dilation. The final network convolutions resulted in differently shaped predictions for five simultaneous objectives: inter-residue distance predictions over 10 bins (see CASP14 format for specifications), secondary structure predictions (9 classes), backbone phi and psi torsion angles (37 bins each), and accessible surface area (11 bins). During training the loss was weighted to give preference to the quality of distance predictions, with the others acting as auxiliaries.

To augment the 27k training domains, ProSPr was trained to predict 64x64 residue crops of full LxL distance matrices. Dividing the domains in this way resulted in over 3 million training instances per epoch; using random offsets to select crops also helped increase the variety of training instances. During inference, predictions for 10 different grids of crops covering the entire domain were assembled and averaged to give the final LxLx10 distance prediction matrix.

In following with the new CASP14 contact format RMODE 2, the distance probabilities for each i,j residue pair were reported in each of the 10 distance bins ranging from 0-20Å; training ProSPr on those same bin definitions eliminated any need to aggregate probabilities across different distance ranges. The contact probability of residue-pair CBs being within 8Å was reported as the sum of probabilities over the first three distance bins. All pairs were then ranked by contact probability and – if necessary – only the 50k most probable pairs were submitted.

All targets were processed using this procedure. However, ProSPr as described here was still being trained while the experiment progressed, so different models were used across the course of the experiment. Predictions for any given target were typically made by ensembling the most recent versions of the four independent ProSPr networks being trained in parallel. After training converged, the same ensemble of four models was used to make predictions for the remaining targets (beginning with T1087).

Acknowledgements

We thank the Office of Research Computing at Brigham Young University for compute time and the local supercomputer Marylou. We also express gratitude for helpful discussions with David Wingate, Tommy Knotts, and Joshua Price at BYU and Michael Feig at MSU. We also acknowledge the contributions of other DellaCorte Lab members not directly involved in CASP14.

Availability

A previous version of ProSPr has been made available both as source code and a Docker container [https://github.com/dellacortelab/prospr]. An updated version in alignment with this description (reflecting significant changes) will be made available shortly.

1. Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., ... & Orengo, C. A. (2010). Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res. 39, D420-D426.
2. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
3. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173-175.
4. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinform. 20, 1-15.
5. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J., & Steinegger, M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170-D176.
6. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. USA 117, 1496-1503.
7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition.In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (pp. 770-778).

DESTINI

Deep-learning the protein folding code for structure prediction and sequence comparison

Mu Gao, Hongyi Zhou and Jeffrey Skolnick
Center for the Study of Systems Biology, Georgia Institute of Technology, Atlanta, USA
mu.gao@gatech.edu, hongyi.zhou@biology.gatech.edu, skolnick@gatech.edu

Key: *Auto:N; CASP_serv:Y; Templ:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP13, we introduced DESTINI[1], a contact-driven folding algorithm that takes advantage of deep convolutional neural networks designed to recognize residue-residue contact patterns. For CASP14, in DESTINI2 we extended the approach from being contact-driven to distance-matrix driven and devised a new template selection and refinement protocol. In addition, a novel deep-learning based sequence alignment algorithm, SAdLSA[2], was trained from deep-learning structural alignments to assist remote template identification and a distance-matrix alignment algorithm was used to rank templates from various sources.

Methods

The major improvements in DESTINI2 include: (1) predicting both C_α-C_α and C_β-C_β distance bins up to 20 Å using a dilated convolutional neural network composed of 40 to 50 residual blocks. (2) A new folding protocol implemented to take advantage of the distance-matrix prediction from deep-learning. (3) UniClust30[3], a large sequence library, was employed to obtain a multiple sequence alignment, which is then employed to derive input features to the deep-learning neural networks. We consider both PSI-BLAST profiles[4] and HMM profiles from HHblits[5]. Three 2D features are employed: co-evolutionary coupling scores[6], a statistical potential[7], and mutual information for pairs of residues[8]. (4) A new structural refinement component based on the TASSER^{VMT} approach[9] was adopted. Multiple models were generated by DESTINI2 with different starting templates from SP3[10], SAdLSA[2] and their 3D-jury top templates. For each set of starting templates, the top 5 models were selected based on their cluster size after SPICKER[11] clustering on the low energy trajectories from the TASSER simulation[9]. We developed a distance matrix based alignment method to align the predicted distance matrix to these models as well as models downloaded from other CASP servers serving as initial templates. All were aligned to the DESTINI2 distance matrix and selected based on their alignment scores. The selected top 10 models were subsequently refined with similar approach as in TASSER^{VMT} [9] that uses a variable number of templates to build up to 50 multiple template based models using Modeller program[12]. For submission, the GOAP energy function[7] was employed to select the top 5 models from the 50 generated models.

Human intervention was applied to multiple domain targets, which was partitioned into individual domains according to the contact prediction of the full sequence and template threading results. Each domain was then modeled separately using DESTINI2 and subsequent refinement.

Results

DESTINI2 was benchmarked on a data set composed of 362 “glass-ceiling” targets. This set is the same as the previous benchmark data set, but here we removed targets whose structures were determined by NMR. Only considering the top1 model, DESTINI2 is capable of predicting native-like folds for 69% of targets, compared to 41% by DESTINI and only 9% by the classic TASSER.

The mean TM-score is 0.52 by DESTINI2, indicating a highly likely correct fold, versus 0.39 by DESTINI. Even when there is no improvement in middle/long range contact predictions, we obtain an average TM-score improvement of 0.10, demonstrating that the distance matrix greatly improved model quality overall.

For the full, very hard "glass-ceiling" data set of 606 targets, only considering the top 1 hit, SAdLSA[2] detects a significant template with TM-score > 0.4 in a template library of about 7,000 structures for 123 targets, versus 66 by HHsearch[13]. Note that SAdLSA is not a threading algorithm because it does not use the coordinate data from the template structures for its sequence alignment; rather it is designed to predict the structural alignment of a target to a template without have the structures of either the target or template proteins.

Availability

Benchmark data sets and the DESTINI2 webserver are available at
http://sites.gatech.edu/cssb/destini.
SAdLSA is available at http://sites.gatech.edu/cssb/sadlsa.

1. Gao M, Zhou H, Skolnick J. DESTINI: A deep-learning approach to contact-driven protein structure prediction. Scientific reports. 2019;9(1):3514.
2. Gao M, Skolnick J. A novel sequence alignment algorithm based on deep learning of the protein folding code. Bioinformatics. 2020;In press.
3. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45(D1):D170-D176.
4. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-3402.
5. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods. 2011;9(2):173-175.
6. Seemayer S, Gruber M, Soding J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014;30(21):3128-3130.
7. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J. 2011;101(8):2043-2052.
8. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry. 2005;44(19):7156-7165.
9. Zhou H, Skolnick J. Template-based protein structure modeling using TASSERVMT. Proteins. 2012;80(2):352-361.
10. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins. 2005;58(2):321-328.
11. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein fold. J Comput Chem. 2004;25 865--871.
12. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234:779--815.
13. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21(7):951-960.

Tertiary structure and distance predictions with DMPfold2

S.M. Kandathil, J.G. Greener, A.M.C. Lau and D.T. Jones
Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
d.t.jones@ucl.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; MetaG:Y; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

In CASP14, we tested new versions of our DMPfold method1 for tertiary structure prediction. DMPfold2 retains many of the features of DMPfold v1, including the use of iterative restraint prediction and structure generation. Developments include using an embedding of the sequence alignment as the only input to the neural nets and a variety of methods for generating models from predicted constraints. The DMP2 group acted as an automated entry, i.e. one that could have been implemented as a server.

Methods

Multiple sequence alignments (MSAs) were built using HHblits searches against the latest UniRef30 databases available at the time of target release. Where HHblits retrieved fewer than 2000 hits, deeper MSAs were built using one iteration of our deep MSA building procedure[2] which searched the UniRef100, EBI MGnify, NCBI Transcriptome shotgun assembly (TSA), MetaEuk and IMG sequence databases, each time building a list of putative hits and using these as a custom database for a further HHblits search.

The MSA was used as input to a new generation of neural network models, which used a one-hot encoding of the MSA, and the precision matrix derived from the MSA as the only input features. The precision matrix was calculated on the fly using the fast_dca approach[3]. The DMPfold2 neural net model jointly predicts distance distributions, torsion angle probability distributions and backbone hydrogen bonds, which are used as restraints for tertiary structure building. Structural models were built using 3 methods: Distance geometry and simulated annealing (DGSA) using CNS (as implemented in DMPfold v1); DGSA using XPLOR-NIH with newer parameter sets and energy functions; and an in-house folding pipeline called Force-directed Folding (FDF) beginning from an extended chain. The DGSA-based modelling pipelines used 10 rounds of the iterated restraint generation and model building procedure, which is similar to that previously described[1].

One model was selected from each of the 3 modelling methods, and each was refined using dual-space refinement as implemented in Rosetta. Refined and unrefined models were scored and ranked using a prototype neural net operating on Cα coordinates.

Results

The new architecture of the DMPfold2 neural nets makes it considerably faster to run than those in DMPfold v1. This is because only one model has to be run to get all the predicted features, and because input feature generation takes significantly less time than in DMPfold v1. Because the neural net model uses only a precision matrix and a one-hot encoding of the MSA as input, the list of software dependencies is also greatly reduced. Initial benchmarking showed that the new approach produced significantly more native-like models than DMPfold v1 on the CASP13 FM domains.

Availability

DMPfold2 will be made available on the PSIPRED GitHub page (https://www.github.com/psipred) under a permissive licence, and also via the PSIPRED Workbench[4] (http://bioinf.cs.ucl.ac.uk/psipred).

1. Greener,J.G., Kandathil,S.M. & Jones,D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. 10, 3977.
2. Kandathil,S.M., Greener,J.G. & Jones,D.T. (2019). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins 87(12), 1092-1099.
3. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. USA. 117(3), 1496-1503.
4. Buchan,D.W.A., Jones,D.T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Research. 47, W402-W407.

E2E: Towards an end to end structure prediction pipeline.

Sirui Liu[1], Haobo Wang[1], Ivan Anishchenko[3,4], Justas Dauparas[3,4], Sergey Ovchinnikov[1,2]

[1]- FAS Division of Science, Harvard University, Cambridge, MA 02138, USA

[2]- John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA

[3]- Department of Biochemistry, University of Washington, Seattle, WA 98105, USA

[4]- Institute for Protein Design, University of Washington, Seattle, WA 98105, USA

so@fas.harvard.edu

***Key**: Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y.*

Many algorithms and pipelines have been developed that go from sequence to structure, however, a full end-to-end model remains a challenge. Here we present our initial efforts towards a fully differentiable set of modules that go from sequences to distance/dihedral matrices to 3D coordinates.

Methods

For the end to end protocol, we tested two different approaches: (A) modification of the last layer of the TrRosetta NN (neural network) model[1] to return a full alpha-carbon distance matrix and backbone dihedrals, (B) conversion of the binned distribution to a full distance matrix and dihedrals via path tracing. Finally, to recover the 3D coordinates, we experimented with two approaches: (1) decomposition of the distance matrix, (2) iterative approach that would place atoms one at a time conditioned on both distances and dihedrals. Prior work used only dihedrals to place atoms[2]. The NN models were trained and validated on the TrRosetta benchmark set. Proteins larger than 300 residues were cut into chunks of 300 residues, predicted separately, and recombined.

Most of the predictions are from the same method, some of the early predictions were based on work-in-progress NN models. For the purposes of comparison and validation, the five models are ranked by method instead of quality. For model 1, we used method A2, for model 5 we used method B2. For models 2 and 3, we tried different combinations of the experimental approaches. For a couple of targets, we submitted server models that matched our distance predictions best. As a control, for model 4, we submitted the results from the default TrRosetta protocol, using the same multiple sequence alignment as model 1 and 5. To reconstruct the sidechains, the final submitted models were relaxed with Rosetta ref2015[3], except for the first couple targets.

Results

We demonstrate this approach returns structures of comparable quality to those generated by the TrRosetta protocol that requires a very expensive minimization step. Going forward, we think these modules can be easily incorporated into any deep learning protocol for a full end-to-end training.

Availability

github.com/sokrypton/e2e

1. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences 117, 1496-1503 (2020).
2. AlQuraishi, M. End-to-end differentiable learning of protein structure. Cell systems, 8(4), 292-301 (2019).
3. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation 13, 3031-3048 (2017).

edmc_pf

Protein folding from contact maps using Euclidean distance matrix completion

A. Lafita
European Bioinformatics Institute (EMBL-EBI)
aleixlafita@ebi.ac.uk

***Key**: Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

The aim of the edmc_pf group was to improve the accuracy of contact map predictions and convert them into atomic coordinates using Euclidean distance matrices (EDMs). Protein distance matrices are low-rank EDMs, in a 3D embedding space, so their structure can be exploited together with protein geometry to complete missing entries and correct erroneous distances. Only a small number of contact (31) and tertiary structure (18) predictions were submitted due to time limitations.

Methods

The target sequence, contact map prediction and secondary structure prediction were used as input. The secondary structure of targets was predicted using the PSIPRED server[1]. Binary C-beta contact map predictions by RaptorX[2] were downloaded from the Prediction Center website, as provided by CASP. An empty distance matrix of all backbone atoms in the target was created and completed using the dissimilarity parameterization formulation (DPF) algorithm[3], using distance constraints from the contact map, secondary structure and protein geometry. More details on the matrix completion step can be found in a recent publication[4]. Very few contacts were needed as constraints to complete the matrix, so only the top k residues in contact to each residue in the target were selected. The value of k varied among targets, depending on the accuracy of the original contact matrix prediction, and was selected manually by looking at the error of the completion convergence. Distance matrix predictions were submitted in the CASP14 RR2 format using the CB-CB distances of the completed matrix including a confidence interval based on distance errors of each residue. Completed distance matrices were converted into atomic coordinates of the target protein backbone using multidimensional scaling. Mirror images were inverted manually.

Availability

Code to model protein structures using EDMs is openly available on GitHub at https://github.com/lafita/protein-edm-demo

1. Jones,DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
2. Wang,S, Sun,S, Li,Z, Zhang,R, Xu,J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Comp. Biol. 13(1): e1005324.
3. Trosset,MW (2000) Distance matrix completion by numerical optimization. Comput Optim Appl. 17(1): 11–22.
4. Lafita,A, Bateman,A (2020) Modelling structural rearrangements in proteins using Euclidean distance matrices. F1000Research, 9:728.

EDN

Protein model quality assessment solely based on the structure of individual models

S. Eismann[*,1], P. Suriana[*,1], B. Jing[1], R.L. Townshend[1], and R.O. Dror[1]
1 - Stanford University
ron.dror@stanford.edu

***Key:** Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

We participated in the quality assessment (QA) category of CASP14 with a single-model, deep-learning method. Our method only uses the 3D atomic structure to assess the quality of individual protein models.

Methods

Our method builds on a novel neural network architecture that is specifically designed to learn from 3D atomic structures[1,2]. Given just the atomic coordinates of a protein model, the network learns to predict a global quality score. Due to inherent symmetry properties of the network, the orientation in which models are provided to the network thereby does not matter.

A second aspect of our network is its hierarchical learning approach: The network first considers the local neighborhood around each atom, then aggregates this information at the level of alpha carbons, and finally outputs a global score for the entire protein model. In combination, the symmetry properties and the hierarchical approach allow the network to recognize structural motifs at different scales, and independent of spatial orientation, and also enable the network to learn end-to-end from all atoms at once.

We trained our method on candidate models submitted to CASP5-10 with the goal to predict GDT_TS for each model. For training, we relaxed each model using SCWRL[3]. We omitted this step when making our predictions for the CASP14 models. Our method uses no physics-inspired energy terms, templates or multiple-sequence alignments. We used the same method for all predictions, performed no manual intervention and did not target a specific set of proteins.

Availability

A webserver is available at http://drorlab.stanford.edu/edn.html.

1. Thomas,N., Smidt,T., Kearnes,S., Yang,L., Li,L., Kohlhoff,K., Riley,P. (2018). Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. arXiv:1802.08219.
2. Eismann,S., Townshend,R.J.L., Thomas,N., Jagota,M., Jing,B., Dror,R. (2020). Hierarchical, rotation-equivariant neural networks to predict the structure of protein complexes. arXiv: 2006:09275.
3. Krivov,G.G., Shapovalov,M.V., Dunbrack,R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins.

Template based and free docking in CASP14

C. Bassot[12*], G.Pozzati[12*], P. Kundrotas*2* [3] and A. Elofsson[12]
1-Science for Life Laboratory, Stockholm University, Solna, Sweden, 3 -Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, 12-Center for Computational Biology Department of Molecular Biosciences, The University of Kansas, Lawrence, KS, United States
arne@bioinfo.se

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

Methods

Our pipeline for multimeric CASP targets starts with the identification of the best monomeric subunit(s) from the CASP server models. The selection was carried on using Pcons1 and ProQ42. The top-scoring models were inspected manually and used for the protein-protein docking.

A search for multimeric PDB templates was performed for each multimeric target using HHsearch3. For homomeric targets, were prioritized templates matching the oligomeric state of identified templates, while for heteromeric targets hits for whom different sequences have homologous in the same PDB. All the identified targets were used to create a customized template library specific for each target.

We then run TMDOCK4 with the selected server models of the target versus the customized template library. For each run, we generated 5 models that were subsequently filtered removing the models forming backbone clashes, and relaxed using the Rosetta package5. If the relaxed protein maintained the protein-protein interaction, the target was submitted.

In the case of trimer and bigger complexes or for specific dimers, the monomeric modes were aligned manually on the selected templates using the align command of Pymol6.

When no templates were available, we obtained the docked structure using the contact prediction as constraints. By HHblits7 we generate the multiple sequence alignment that was used as input for DeepMetaPsicov8 to predict the contacts. We use as restraints the predicted contacts between the monomers or in the case of homomers between residues further than 12Å in the model. In both cases, we select only the predicted contacts with a score higher than 0.5. The contacts were finally used as restraints in Haddock9.

In one case for (T1032) we used the Swissmodel web server[10]**.**

Results

At the current date, two of the multimers we modelled have a resolved structure. Our best models are shown in Fig 1. Both targets were modelled by template-based docking.

The area of interaction was predicted correctly in both the cases but still, major differences between the models and the structures are present at the interface. These differences are reflected in the low DockQ[11] score Table 1. In T1032 different methods were used, in this case, it appears that the docking based on HHpred template search and the manual alignment with Pymol and Swissmodel perform better than the TMdock pipeline.

Target	Method	DockQ
T1032_4	**Manual HHpred/Pymol Docking**	**0.039**
T1032_1	Swissmodel	0.039
T1032_3	TMdock	0.012
T1032_2	TMdock	0.008
T1099_1 (Interface A)	**Manual HHpred/Pymol Docking**	**0.064**
T1099_1 (Interface B)	**Manual HHpred/Pymol Docking**	**0.047**

Table1 Methods and DockQ for submitted targets. In bold the models shown in Figure 1.

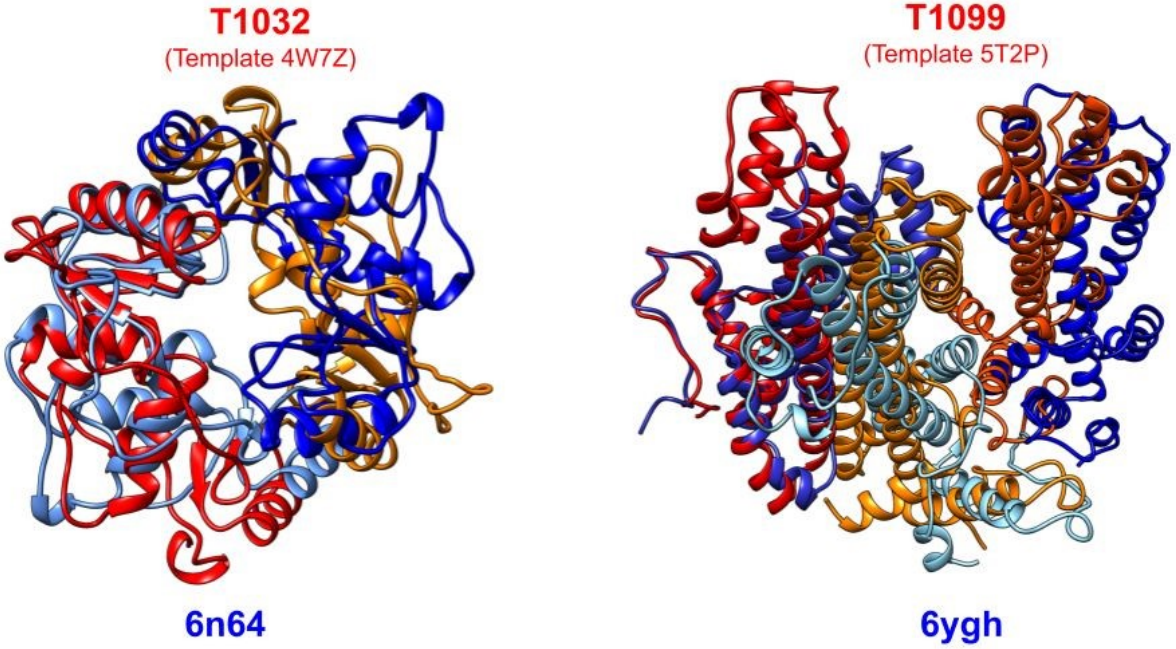


Figure 1 In red, orange, dark orange the predicted model, in blue, dark blue and light blue the corresponding resolved structure.

1. B.Wallner, P.Larsson, A.Elofsson, Pcons.net: protein structure prediction meta server, Nucleic Acids Res. 35 (2007) W369–74.
2. D.Menéndez Hurtado, K.Uziela, A.Elofsson, Deep transfer learning in the assessment of the quality of protein models, arXiv. (n.d.).
3. J.Söding, Protein homology detection by HMM-HMM comparison, Bioinformatics. 21 (2005) 951–960.
4. A.L.Lomize, I.D.Pogozheva, TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes, J. Mol. Biol. 429 (2017) 390–398.
5. A.Leaver-Fay, M.Tyka, S.M.Lewis, O.F.Lange, J.Thompson, R.Jacak, K.Kaufman, P.D.Renfrew, C.A.Smith, W.Sheffler, I.W.Davis, S.Cooper, A.Treuille, D.J.Mandell, F.Richter, Y.-E.A.Ban, S.J.Fleishman, J.E.Corn, D.E.Kim, S.Lyskov, M.Berrondo, S.Mentzer, Z.Popović, J.J.Havranek, J.Karanicolas, R.Das, J.Meiler, T.Kortemme, J.J.Gray, B.Kuhlman, D.Baker, P.Bradley, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, Methods Enzymol. 487 (2011) 545–574.
6. Schrödinger, The PyMOL Molecular Graphics System, Version 1.2r3pre, LLC. (n.d.).
7. M.Steinegger, M.Meier, M.Mirdita, H.Vöhringer, S.J.Haunsberger, J.Söding, HH-suite3 for fast remote homology detection and deep protein annotation, BMC Bioinformatics. 20 (2019) 473.
8. J.G.Greener, S.M.Kandathil, D.T.Jones, Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints, Nat. Commun. 10 (2019) 3977.

9. S.J.de Vries, M.van Dijk, A.M.J.J.Bonvin, The HADDOCK web server for data-driven biomolecular docking, Nat. Protoc. 5 (2010) 883–897.

10. A.Waterhouse, M.Bertoni, S.Bienert, G.Studer, G.Tauriello, R.Gumienny, F.T.Heer, T.A.P.de Beer , C.Rempfer, L.Bordoli, R.Lepore, T.Schwede, SWISS-MODEL: homology modelling of protein structures and complexes, Nucleic Acids Res. 46 (2018) W296–W303.

11. S.Basu, B Wallner, DockQ: A Quality Measure for Protein-Protein Docking Models, PLoS One. 11 (2016) e0161879.

Advanced Deep-Learning applications on CASP14 protein models quality assessment

F.Baldassarre[1*], G.Pozzati[23*], H.Azizpour[14] and A. Elofsson[234]

[1]-KTH - Royal Institute of Technology, Division of Robotics, Perception and Learning (RPL), Stockholm, Sweden, [2]-Science for Life Laboratory, Stockholm University, Solna, Sweden, [3]-Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden, [4]-Swedish e-Science Research Centre (SeRC)

arne@bioinfo.se, azizpour@kth.se

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:Y; EMA:N; MD:N*

Nowadays it is very important to follow up protein structure prediction methods with a quality assessment (QA) step, able to verify modelled structures' reliability. For the 14th CASP edition, we submitted quality estimates derived from two Deep Learning-based predictors, ProQ4[1] and GraphQA[2]. Here, we present a brief description of these methods, as well as a preview of such methods' performance, calculated on 14th CASP edition targets for which the crystal structures are already available.

Methods

ProQ4 is a deep learning predictor which uses as input a multiple sequence alignment (MSA), as well as a coarse representation of the protein models to be evaluated. This predictor is trained to extrapolate the Local Distance Difference Test (LDDT), a metric which allows both local and global model QA. ProQ4's neural network is composed of a complex architecture based on a comparison between pairs of protein models. The predictor ability to discriminate which one of the models in each pair is better is proven to confer a significant boost in the absolute scoring. In order to generate the input MSA, one iteration of JackHMMer has been run for each CASP target, using uniref90 as a search database. The resulting MSA in Stockholm format has been converted to fasta format using the esl-reformat tool from the HMMer package (version 3.1b2). Finally, the QA scores have been obtained by running ProQ4, after providing the fasta MSA and the list of models resulting from the different stages of each CASP target.

GraphQA estimates protein quality using a graph-based representation of protein structure and a Graph Convolutional Network. Overall, GraphQA employs input features similar to ProQ4 but achieves better performances on past CASP editions thanks to a better representation of the spatial structure, which is based on graphs rather than sequences. Specifically, the input to GraphQA is a graph whose nodes represent amino-acids and whose edges represent contacts between residues. For each node, we provide an embedding of the amino-acid type, features from an MSA computed against Uniref50, and secondary structure features from DSSP. By construction, edges are placed between nodes that are neighbours in the sequence, i.e. the corresponding residues appear close in the primary structure, or that are neighbours in space, i.e. they are within a certain distance in the tertiary structure. A single GraphQA model is trained to output many quality assessment scores, at both the residue and protein level. Namely, for each residue, LDDT and CAD scores are predicted. Also, at the protein level GraphQA predicts GDT-TS, GDT-HA, TM-score, LDDT and CAD.

Results

Currently, 18 targets from the 14th CASP edition have been linked to an available PDB structure. Comparison of the submitted QA scores with models' LDDT is summarized in Fig 1. GraphQA achieved much better performances than ProQ4, reaching on average a correlation of 0.52 (against average 0.31 correlation of ProQ4).

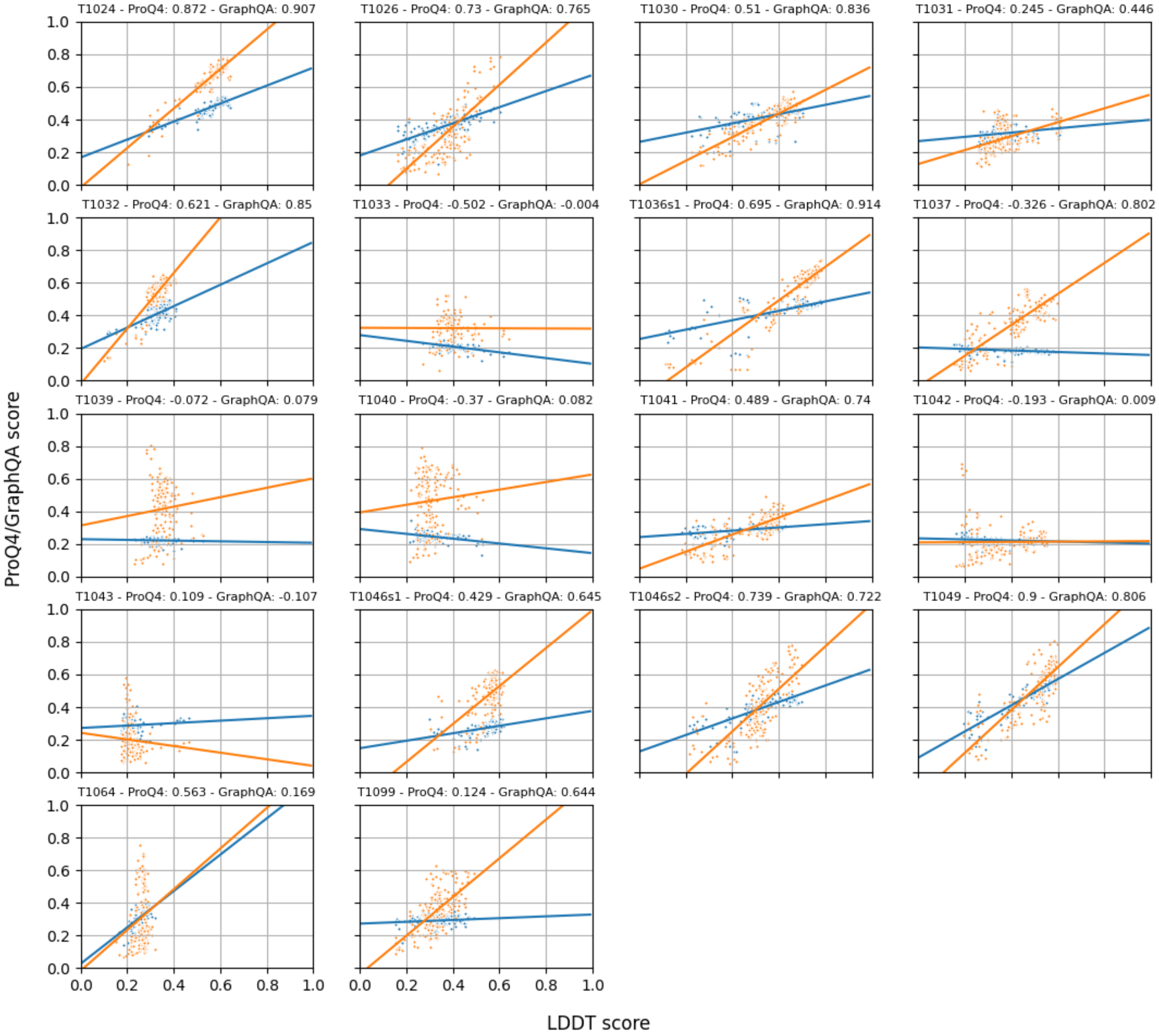


Figure 1: Comparison of ProQ4 (Blue) and GraphQA (Orange) predictions with modelled structures LDDT scores. Each subplot refers to the target reported on top of it. Single target Pearson Correlation Coefficients are also indicated for both predictors, in 14/16 targets the correlation is higher in Graph-QA than in ProQ4.

Performance on single targets displays very wide variation. In most cases, performances of the two methods are comparable, spacing from almost-perfect predictions (T1024, T1049) to completely-inaccurate estimates (T1037, T1039, T1040, T1042, T1043).

In general, GraphQA performs better than ProQ4, but there are few cases (T1049, T1064) where ProQ4 reaches higher correlation values.

1. D.Menéndez Hurtado, K.Uziela, A.Elofsson, Deep transfer learning in the assessment of the quality of protein models, arXiv. (n.d.).
2. F.Baldassarre, D.Menéndez Hurtado, A.Elofsson, H.Azizpour, GraphQA: Protein Model Quality Assessment using Graph Convolutional Networks, Bioinformatics (2020)

EMAP_CHAE

Protein model quality assessment method EMAP in CASP14

Myong-Ho Chae
Department of Life Science, University of Sciences, Unjong-District, Pyongyang, DPR Korea
chae1971@star-co.net.kp

Protein model Quality Assessment is an important topic both in protein structure prediction and in practical applications of structure models. We developed a new method EMAP to predict the residue-specific and global quality of individual protein models. The main component of EMAP is statistical potentials such as DOOP[1] and GOAP[2]. DOOP potential is a distance-dependent atomic potential based on optimization method for the protein structure prediction. To generate decoy structures for optimization of potential, the protein structures in the training set are successively broken into two rigid regions, hypothetical receptor and ligand. These pairs of receptor and ligand are docked by a docking decoy generation program to generate a large number of evenly sampled docking decoys. In EMAP we used two versions of DOOP potential, DOOP-CB which incorporates main-chain atoms and CB atoms, and DOOP-CBCG which incorporates main-chain atoms, CB, and CG atoms.

Methods

To predict residue-specific deviations of a protein model, EMAP uses the following features as input.

1. Per-residue DOOP-CB, DOOP-CBCG potentials averaged on residues within 8.5-, 12- , and 15- Å spatial sphere of a specific residue.
2. Per-residue GOAP potentials (in-house implemented) averaged on residues within 8.5-, 12-, and 15- Å spatial sphere of a specific residue.
3. Secondary structure and solvent accessibility agreements and relative accessibility within 5-, 11-residue sequence window and 12-Å spatial window of a specific residue.
4. Torsion potential, fraction of buried residue and correlation coefficient between predicted and real solvent accessibilities within 11-residue sequence window.

Three-layer perceptron was trained using above 21 features as input to predict the S-score for each residue in the model in the training set (CASP9 and CASP11 data set). The global accuracy score of a model is derived by averaging the predicted local S-scores of residues.

Results

We evaluated EMAP on CASP13 dataset and proved that it achieves the state-of- the-art performance among single-model QA methods.

1. Chae,M.H., Krull,F. & Knapp,E.W., (2015). Optimized distance-dependent atom-pair-based potential DOOP for protein structure prediction, Proteins. 83, 881−890.
2. Zhou,H. and Skolnick,J. (2011) GOAP: A Generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal, 101, 2043−2052.

FALCON-DeepFolder

CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction

Fusong Ju[1, 2], Lupeng Kong[1, 2], Xinchun Ran[1, 2] and Dongbo Bu[1, 2]

[1] - *Key lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences*

[2] - *University of Chinese Academy of Sciences*

dbu@ict.ac.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

Accurate prediction of protein tertiary structures relies heavily on understanding the fine details of the inter-residue distances. Direct coupling analysis (DCA) could identify residue co-evolution and has become the primary technique for estimating inter-residue distance. Multiple sequence alignment (MSA) contains abundance information of residue co-evolution; however, the existing DCA-based approaches exploit the co-variance matrix rather than the original multiple sequence alignment (MSA), which causes considerable information loss at the very beginning. Actually, we have observed that two proteins differ greatly in both MSAs and residue contact maps; however, the co-variance matrices derived from the two MSAs are completely identical. This clearly demonstrates the considerable information loss in converting co-variance from MSAs.

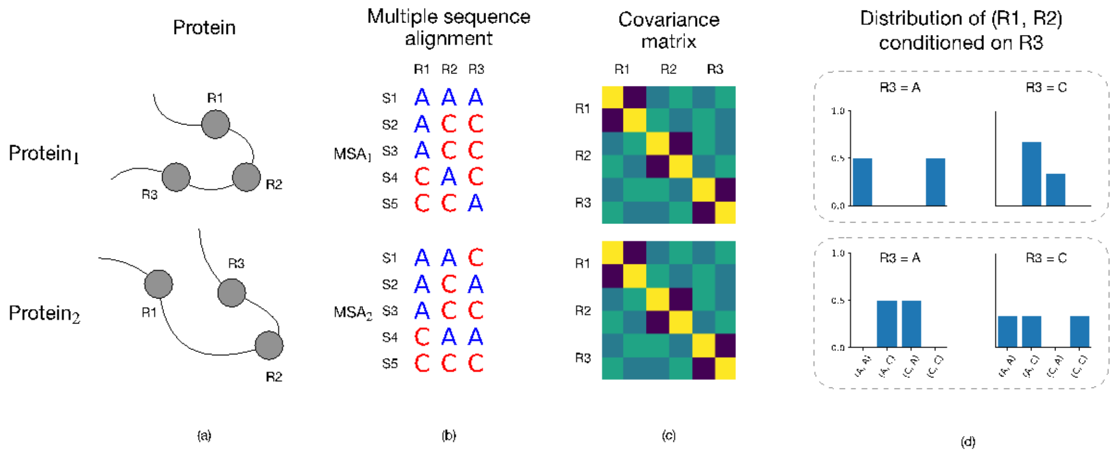


Figure 1. **The limitation of covariance-based method.** (a) Two structures with different contact pattern for residue R1 and R2. (b) Corresponding MSAs of the two proteins. (c) The two MSAs have identical covariance matrix (denote each symbol at each column as a random variable). (d) Distribution of (R1, R2) conditioned on R3. R1 and R2 have stronger direct correlation in MSA1; however, this cannot be distinguished by covariance matrix.

We have established an approach (called ProFOLD) to learn residue co-evolution directly from MSA. For this aim, we designed a novel CopulaNet architecture to model residue correlation and thereafter predict inter-residue distance. CopulaNet uses an MSA-encoder to extract context-specific mutation information for each homologous sequence independently, and then obtains high-order coevolutionary couplings by aggregating these MSA embeddings.

Methods

For a query protein, ProFOLD predicts the inter-residue distance as follows:

1. Multiple sequence alignment (MSA) generation and representation. We take multiple sequence alignment as the only input to train the neural network. For each query sequence, we first generate MSA using DeepMSA searching against uniclust30, uniref90, and metaclust.

Next, we convert the generated MSA to fixed representations and feed them into the deep neural network CopulaNet. Here, we represent MSA as a set of sequence pairs: for each aligned sequence, we construct two equal-length strings by adding gaps in aligned sequences so that matching characters are aligned in successive positions.

Finally, we encode each position with a binary vector of 41 elements, including 20 elements (corresponding to 20 amino acid types) for residue in query protein and 21 elements (corresponding to 20 amino acid types and gap) for residue in homology protein.

2. Distance distribution prediction. Our CopulaNet predicts the inter-residue distances through modeling residue correlation. The main architecture of CopulaNet consists of a deep one-dimensional convolutional residual network and a deep two-dimensional dilated convolutional residual network, which consists of 8 one-dimensional residual blocks and 72 two-dimensional residual blocks with dilated convolutions, respectively. To aggregate the residue correlation extracted from all homology proteins, we insert an average pooling layer into the above-mentioned blocks.

After residual networks, we use a fully-connected layer to predict the discretized distance between C_b atoms of the residues (or C_a for glycine). The distance range (2 to 20 angstrom) is divided equally into 36 bins. We also added an auxiliary bin to indicate residues without any contact.

3. Structure determination based on distance potential. We build the tertiary structure of query protein using the predicted inter-residue distance in a way similar to AlphaFold and trRosetta. Specifically, we first convert the predicted inter-residue distances to smooth energy potential, and then use optimization technique to build structural models with minimal energy.

Availability

https://github.com/fusong-ju/ProFOLD

1. Chengxin Zhang, Wei Zheng, SM Mortuza, Yang Li, and Yang Zhang. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics, 36(7):2105–2112, 2020.
2. Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Z´ıdek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, pages 1–5, 2020.
3. Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 2020.

FALCON-geom

Improving protein tertiary structure construction through reducing inconsistency from the predicted inter-residue distances

Xinchun Ran[1, 2], Fusong Ju[1, 2] , Lupeng Kong[1, 2] and Dongbo Bu[1, 2]

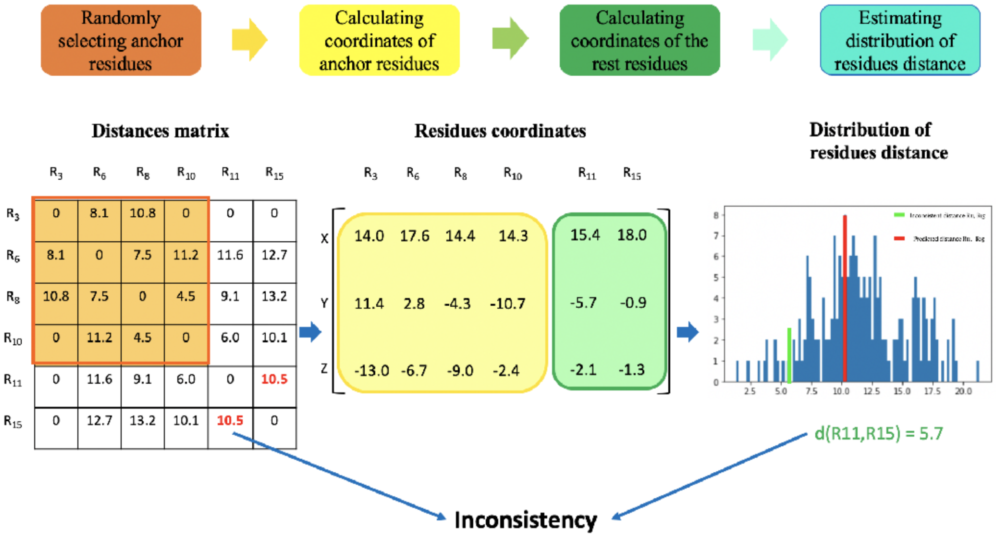
[1] - *Key lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences*

[2] - *University of Chinese Academy of Sciences*

dbu@ict.ac.cn

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

Understanding the fine details of inter-residue distances play important roles in protein structure prediction[1]. After acquiring accurate inter-residue distances, the protein tertiary structure can be easily restored using optimization technique to maximize the fitness of the structure with the distance constraints[2]. However, the predicted inter-residue distances always contain considerable inconsistency, i.e., arbitrarily selecting three residues R_i, R_j, R_k as anchors, the coordinates of residues R_a and R_b can be readily calculated by their distances to the anchors; however, the distance between calculated coordinates of R_a and R_b is usually inconsistent with the predicted distance. These inconsistencies will significantly damage the quality of the constructed protein tertiary structure. How to identify and remove these inconsistency from the predicted inter-residue distance remains a great challenge.


Randomly selecting anchor residues
Calculating coordinates of anchor residues
Calculating coordinates of the rest residues
Estimating distribution of residues distance
Distances matrix
Residues coordinates
Distribution of residues distance
Inconsistent distance Ra, Rb
Predicted distance Ra, Rb
0
1
2
3
4
5
6
7
8
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
d(R11,R15) = 5.7
Inconsistency

	R_3	R_6	R_8	R_{10}	R_{11}	R_{15}
R_3	0	8.1	10.8	0	0	0
R_6	8.1	0	7.5	11.2	11.6	12.7
R_8	10.8	7.5	0	4.5	9.1	13.2
R_{10}	0	11.2	4.5	0	6.0	10.1
R_{11}	0	11.6	9.1	6.0	0	10.5
R_{15}	0	12.7	13.2	10.1	10.5	0

	R_3	R_6	R_8	R_{10}	R_{11}	R_{15}
X	14.0	17.6	14.4	14.3	15.4	18.0
Y	11.4	2.8	-4.3	-10.7	-5.7	-0.9
Z	-13.0	-6.7	-9.0	-2.4	-2.1	-1.3

Methods

We designed a novel method for reducing inconsistency from the predicted inter-residue distance and further construct the whole protein tertiary structure. Our method works on the inter-residue distances predicted using ProFOLD (in-house work). The basic idea of our approach is sampling and optimal seeking". First, we randomly sample three residues as anchors, and then calculate coordinates of the rest residues. After repeating the sampling procedure N times (N = 200 in this study), we acquire N estimation of the distance between all residue pairs. Next, we fit these estimated distances using Gaussian mixture model (GMM), and then select the most probable distance estimation from this model. Finally, we use the alternating direction method of multipliers (ADMM)[3] technique to build protein tertiary structure that best satisfies the distance constraints.

Availability

http://protein.ict.ac.cn/FALCON-geom

1. Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1):e1005324, 2017.
2. Namrata Anand and Possu Huang. Generative modeling for protein structures. In Advances in Neural Information Processing Systems, pages 7494–7505, 2018.
3. Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

FALCON-TBM: Accurate protein threading through learning structure alignment using neural network

Lupeng Kong[1, 2], Fusong Ju[1, 2], Xinchun Ran[1, 2] and Dongbo Bu[1, 2]

[1] - *Key lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences*

[2] - *University of Chinese Academy of Sciences*

dbu@ict.ac.cn

Key: Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N

Building accurate alignment between query protein and templates (known as threading) has become the primary method for protein structure prediction. The ideal alignment of two proteins can be readily calculated when their structures are already known. That is, we superimpose (including rotation and translation) one structure onto the other, and then identify the *matched residues* with distance less than a pre-defined threshold. These matched residues form the ideal alignment between the two proteins.

The ideal alignment calculated from optimal structure superimposition usually shows a clear pattern of *dashed line*, where the dashes come from the matches of corresponding secondary structure elements. This dashed line pattern of alignment, however, have never been systematically exploited in the existing threading approaches for protein structure prediction. Deep learning has been shown to be extremely powerful in learning specific patterns for a variety of types of data especially for images. By treating superimposition matrix as an image, we train a deep neural network model to learn the dashed line pattern existing in superimposition matrix of proteins with known structures, and then apply the trained model to predict the superimposition matrix for query protein and template. By tracing within the predicted superimposition matrix, we finally construct the alignment between query protein and template.

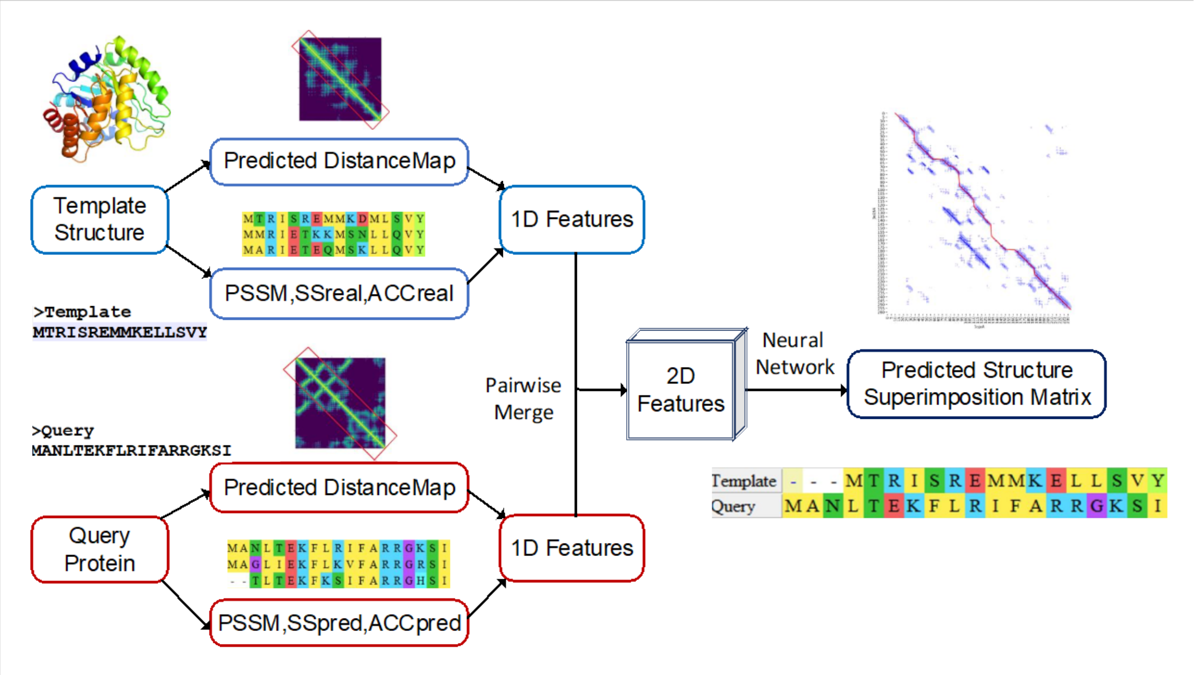
Methods

As shown in the figure (below), our approach (called ProALIGN) consists of three main steps:

1. Calculating features for query protein and template. For both query protein and template, we calculate a collection of sequence and structure features, including secondary structure, sequence profile (PSSM) and solvent accessibility. We also predict the inter-residue distance for these proteins using ProFOLD (an in-house work). Next, these features extracted from the two proteins are merged, yielding 2D features as results.

2. Predicting superimposition matrix of query protein and template. Next, we feed the merged 2D feature into the trained neural network to predict the superimposition matrix between the query protein and template. The neural network model was trained on proteins with known structures to learn the dashed line pattern of the corresponding superimposition matrix. Using the traditional Needleman-Wunsch algorithm, we trace with the predicted superimposition matrix to identify the path with the largest sum score and construct the alignment accordingly.

3. Model generating. Based on the query-template alignments quality, we extract distance restraints from selected templates. For low score region of alignments or inconsistent region of distance constraints among multiple templates, we combine the extracted distance restraints and the distance distribution predicted by ProFOLD, and then use PyRosetta[1] to build model. For several CASP14 targets, we simply run HHpred[2] and CNFpred[3] to calculate sequence-template alignment and then build models using Modeller.


Template Structure
Predicted DistanceMap
M T R I S R E M M K D M L S V Y
M M R I E T K K M S N L L Q V Y
M A R I E T E Q M S K L L Q V Y
PSSM,SSreal,ACCreal
1D Features
>Template
MTRISREMMKELLSVY
>Query
MANLTEKFLRIFARRGKSI
Pairwise
Merge
2D
Features
Neural
Network
Predicted Structure
Superimposition Matrix
Template - - - M T R I S R E M M K E L L S V Y
Query M A N L T E K F L R I F A R R G K S I
Query
Protein
Predicted DistanceMap
M A N L T E K F L R I F A R R G K S I
M A G L I E K F L K V F A R R G R S I
- - T L T E K F K S I F A R R G H S I
PSSM,SSpred,ACCpred
1D Features

Availability

http://protein.ict.ac.cn/FALCON-TBM

1. Sidhartha Chaudhury, Sergey Lyskov, and Jeffrey J Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics, 26(5):689–691, 2010.
2. Jianzhu Ma, Jian Peng, Sheng Wang, and Jinbo Xu. A conditional neural fields model for protein threading. Bioinformatics, 28(12):i59–i66, 2012.
3. Johannes S ̈oding. Protein homology detection by HMM–HMM comparison. Bioinformatics, 21(7):951–960, 2005.

FEIG and FEIG-R1,2,3 (TS)

Protein structure prediction via refinement of CASP server models

Lim Heo[1] and Michael Feig[1]

[1] – *Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*

mfeiglab@gmail.com

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

Based on the same protocol as in **FEIG-S (TS)**, we operated four automated meta servers that used structure prediction results from other selected server groups. We chose RaptorX[1], Zhang-server[2], and BAKER-ROSETTASERVER[3] for FEIG-R1, -R2, and -R3, respectively. We applied the same MD simulation-based refinement method, **FEIG-S** (refinement), except that we did not use multiple alternative initial models. For targets where we manually intervened to account for ligand binding, oligomerization, and membrane environments for the FEIG-S prediction, we used the same information during the refinement of the selected server models.

For FEIG predictions, we aggregated all of the generated MD simulation trajectories during the refinement step. We generated another set of refined models by using the trajectories. A set of structures was selected by using RWplus and RMSD to the reference structure[4,5], and it was averaged to the refined model. The initial models of the FEIG-S prediction were used as the reference structures. Also, the sampled conformations were clustered and averaged to obtain other refined models. For multiple domain targets, a subset of atoms that correspond to each domain were used to generate refined models for the corresponding domains.

1. Xu,J. & Wang,S. (2019). Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins 87, 1069-1081.
2. Zheng,W., Li,Y., Zhang,C., Pearce,R., Mortuza,S.M. & Zhang,Y. (2019). Deep-learning contact-map guided protein structure prediction in CASP13. Proteins 87, 1149-1164.
3. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117, 1496-1503.
4. Heo,L. & Feig,M. (2018). What makes it difficult to refine protein models further via molecular dynamics simulations? Proteins 86 Suppl 1, 177-188.
5. Mirjalili,V. & Feig,M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J Chem Theory Comput 9, 1294-1303.

FEIG-S (refinement)

Protein Model Refinement via Molecular Dynamics Simulations with an Improved Structure Sampling Protocol and Multiple Alternative Models

Lim Heo[1] and Michael Feig[1]

[1] – *Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*

mfeiglab@gmail.com

***Key:** Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

Protein model refinement has become one of the important steps in the endgame of protein structure prediction. Molecular dynamics simulation-based methods have shown encouraging results not only for template-based models[1], but also machine learning-based models.[2] In the previous CASP experiment, a few refinement initial models could be improved to have highly similar structures to their experimental structures with Cα-RMSD of around 1 Å or better.[3,4] We performed MD simulations with flat-bottom harmonic restraints on Cα atoms, which limited conformational sampling in the vicinity of the initial model. Conformational sampling was carried out with three iterations, and MD simulations were carried out for 2 μs in total for each target. From the post-CASP analysis, the new type of restraints for MD simulation was the most effective change for the progress, while the iterative sampling scheme merely contributed. As refinement performance highly depends on conformational sampling, efficient sampling methods are essential.

We operated an automatic refinement server, FEIG-S, during CASP14. The server is based on a new refinement protocol based on MD simulations that is augmented by template-based models. The protocol mainly consisted of two parts: (1) generation of template-based models using the original initial model and additional template structures, (2) MD simulation-based conformational sampling and followed ensemble averaging of the sampled structures. MD simulations were performed starting from the original initial model and additional multiple alternative models if they were available. All the sampled structures were considered together to create an ensemble-averaged model. locPREFMD[5] and was applied as the final step to improve stereochemical properties. Finally, residue-wise model quality was predicted by an MD-based method.[3]

Methods

The refinement protocol for FEIG-S includes two main components. First, multiple alternative initial models were generated using the original initial model and additional template structures. The template structures were searched by HHblits and HHsearch[6] and selected using structure similarity to the original initial model (TM-score > 0.6). Single template-based models were built by MODELLER[7] with sequence alignment produced by HHalign. These models were optimized further by hybridizing between them and the original initial model using Rosetta.[8] The initial pool of structures was selected using similarity to the original initial model with a cutoff, either a TM-score of 0.6 or the best TM-score among the built models minus 0.2, whichever is greater. If there were less than two models available, we did not use multiple initial models. Up to nine models were selected to construct the initial pool with the original initial model. If there were less than ten models, the selected models were replicated. After ten iterations of hybridization, the four lowest Rosetta score structures were selected as additional alternative initial models for the further conformational sampling.

Second, MD simulations were carried out with the initial models after application of locPREFMD. For each initial model, 5 independent simulations were performed for 100 ns at 360 K. MD simulations were performed with a modified CHARMM force field to facilitate barrier crossings and in the presence of explicit water molecules. We used hydrogen mass repartitioning, which re-distributed masses to make hydrogen atoms heavier (3 a.m.u.), so that a 4-fs integration time step could be used with the SHAKE algorithm. During the simulations, conformations were restrained with respect to each initial model by Cartesian and distance restraints on Cα of the proteins. Both restraints were based on a flat-bottom harmonic function. The Cartesian restraints were applied for Cartesian coordinates of every Cα atom with a force constant of 0.025 kcal/mol/Å^2 and a bottom width of 4 Å. The distance restraints were applied for distances between Cα atoms that had distances lower than 10 Å in the initial model and sequence separation was greater than 3 residues. We set 0.05 kcal/mol/Å^2 and 2 Å for the force constant and bottom width of the flat-bottom harmonic function, respectively. The restraints were gradually switched from Cartesian restraints to the distance restraints during a simulation.

All the sampled conformations were aggregated and selected to generate an ensemble averaged structure. The scheme for ensemble structure selection depended on the number of initial models. For targets where we refined only based on the original initial model, 25% of the lowest RWplus score models were selected for averaging. For targets with additional multiple initial models, the RMSD to the original initial model was additionally considered as in the scheme used during CASP12.[9,10] We applied SCWRL4[11] and locPREFMD to the ensemble averaged structure to obtain a final model. Finally, residue-wise errors were predicted by an MD-based method described earlier.[3]

The protocol was fully automated except for a few targets, which had putative binding ligands, extensive inter-protein contacts, or were assumed to be membrane proteins. Putative binding ligands were inferred from homologous structures and modeled using CGenFF parameters during the MD simulations. Externsive inter-protein contacts due to oligomerization were accounted for by simulating the oligomer species instead of monomers. The relative orientation between protein subunits was inferred from homologous structures. Refinement of membrane proteins were prepared by using CHARMM-GUI[12] with POPC lipid bilayers in order to reflect the membrane environment.

Results

When the refinement protocol was benchmarked on CASP11–13 refinement targets, it outperformed our previous protocol used during the last CASP. Model qualities were improved by 4.72 and 3.81 on average in terms of GDT-HA for protocols with multiple and single initial models, respectively, while the improvement was 2.34 for CASP13 protocol without iterations.

Among the 38 regular refinement targets, we could build additional alternative models only for 6 targets. In contrast, during the benchmark, 65 out of 103 refinement targets were available for alternative initial model building. It is probably because regular structure prediction (TS) targets were harder than before, so that most of the refinement initial models were built by contact-based methods.

1. Heo,L. & Feig,M. (2018). Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proc Natl Acad Sci U S A 115, 13276-13281.
2. Heo,L. & Feig,M. (2020). High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins 88, 637-642.
3. Heo,L., Arbour,C.F. & Feig,M. (2019). Driven to near-experimental accuracy by refinement via molecular dynamics simulations. Proteins 87, 1263-1275.
4. Read,R.J., Sammito,M.D., Kryshtafovych,A. & Croll,T.I. (2019). Evaluation of model refinement in CASP13. Proteins 87, 1249-1262.

5. Feig,M. (2016). Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. J Chem Inf Model 56, 1304-1312.
6. Steinegger,M., Meier,M., Mirdita,M., Vohringer,H., Haunsberger,S.J. & Soding,J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473.
7. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.
8. Park,H., Ovchinnikov,S., Kim,D.E., DiMaio,F. & Baker,D. (2018). Protein homology model refinement by large-scale energy optimization. Proc Natl Acad Sci U S A 115, 3054-3059.
9. Mirjalili,V. & Feig,M. (2013). Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles. J Chem Theory Comput 9, 1294-1303.
10. Heo,L. & Feig,M. (2018). What makes it difficult to refine protein models further via molecular dynamics simulations? Proteins 86 Suppl 1, 177-188.
11. Krivov,G.G., Shapovalov,M.V. & Dunbrack,R.L.,Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778-795.
12. Wu,E.L., Cheng,X., Jo,S., Rui,H., Song,K.C., Davila-Contreras,E.M., Qi,Y., Lee,J., Monje-Galvan,V., Venable,R.M., Klauda,J.B. & Im,W. (2014). CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. J Comput Chem 35, 1997-2004.

FEIG-S (TS)

High Accuracy Protein Structure Prediction via Contact Prediction and Physics-based Refinement

Lim Heo[1] and Michael Feig[1]

[1] – *Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*
mfeiglab@gmail.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:Y*

Previouisly, protein structures could be predicted at reasonable accuracy in atomistic detail via template-based modeling, followed by model refinement, e.g. via MD simulation-based methods. With recent advances in machine learning techniques and growing protein sequence databases, reliable structure prediction has become possible without explicit template structures, but based on predicted inter-residue contacts.[1,2] As in the refinement of template-based models, MD simulation-based refinement methods have also led to remarkable improvements in model qualities for machine learning-based models.[3] MD-based refinement of machine-learning models was especially effective for refining moderate-to-high accuracy machine learning-based models. There were improvements in loop, terminal regions and local structure packing.

We operated an automatic prediction server, FEIG-S, during CASP14. The server is based on a combined method of a distogram-prediction method[2], template-based modeling, and an improved refinement protocol (*see* **FEIG-S (refinement)** for details). An initial model was predicted by the distogram-based structure prediction method and template-based modeling. The model was split into domains, and each domain structure was subjected to our refinement protocol. After refinement, the models were joined by superimposing onto the initial model. Finally, locPREFMD[1] was applied to the joined model to recover correct stereochemistries at domain boundaries.

Methods

Our structure prediction method performed distogram-based structure prediction and MD simulation-based refinement, sequentially. The distogram-based structure prediction was based on trRosetta[2], but with some modifications. For a target sequence, signal peptides and expression tags at terminals were trimmed before multiple sequence alignment generation. Homologous sequences were iteratively searched against the UniClust30 database using HHblits[5] with gradually relaxing E-value cutoffs from 1e-80 to 1e-4 until enough sequences were searched. The criteria of how many sequences were considered ‘enough' was set as in the original method. If there were still not enough sequences in the MSA, sequences were searched again against UniRef100 using HMMER[6] with the MSA as an input. The searched sequences were filtered using hhfilter with a sequence identity cutoff of 90%. If there were more than 100,000 sequences in the resulting MSA, we lowered the sequence identity cutoff by 10% until the number of sequences became lower than 100,000. Inter-residue distances and orientations were predicted by trRosetta with the filtered MSA. In the meantime, template-based models were predicted for the sequence. Modeling templates were searched by using BLAST[7], HHsearch[5], and HMMER[6]. For template selection, we used an E-value cutoff of 0.001 and a sequence identity cutoff of 30%, 20%, and 20% for BLAST, HHsearch, and HMMER, respectively. For each template search method, up to ten templates were selected to build a model using MODELLER.[8] Distograms were generated from the template-based models and combined with with the trRosetta predictions using weights according to the sequence similarity of the homologs.

From the distogram prediction by trRosetta, protein domains were inferred along with a secondary structure prediction using PSIPRED.[9] Protein sequences were segmented with the secondary structure prediction into up to four consecutive residues. The segments were clustered by community detection with the predicted contacts as edges of graphs. Domains were defined based on the clustering; a sequence was split into two domains at a time. If one of the domains had less than 30 residues, we did not split. Also, if it was possible to build a template-based models, we did not split it. The domain boundary was extended for seven residues toward each terminal direction. With a sequence for the new domain definition, MSA generation and following trRosetta runs were conducted until the domain could not be split further. Predicted contacts from the iterative trRosetta runs then replaced submatrices in the contact distograms in the order of the runs. A contact map for a domain was replaced if the contact map for the subregion had a higher mean contact probability of top L pairs with sequence separation of greater or equal than twelve residues. Protein models were built by PyRosetta[10] from the replaced contacts map. We generated 16 models for a target and took the lowest Rosetta score model as an initial model for the followed refinement.

We applied our refinement protocol as described in another abstract, **FEIG-S (refinement)**, to the model. After refinement, refined models for domains were superimposed onto the initial model. Domains were joined at a residue, which had the minimum Cα deviation between domains, among the overlapped residues. Finally, locPREFMD[1] was applied to improve local bond geometry at the domain boundaries.

Results

We benchmarked the protocol on CASP13 targets. As a result, in terms of GDT-HA, the model of trRosetta showed comparable performance with 45.14 to AlphaFold[1]'s 45.19 on average for 109 human targets domains. After refinement, the refined models had a GDT-HA score of 48.52 (+3.38) on average, which was significantly better than AlphaFold (p=4.8e-4).

1. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Zidek,A., Nelson,A.W.R., Bridgland,A., Penedones,H., Petersen,S., Simonyan,K., Crossan,S., Kohli,P., Jones,D.T., Silver,D., Kavukcuoglu,K. & Hassabis,D. (2020). Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710.
2. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117, 1496-1503.
3. Heo,L. & Feig,M. (2020). High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins 88, 637-642.
4. Feig,M. (2016). Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. J Chem Inf Model 56, 1304-1312.
5. Steinegger,M., Meier,M., Mirdita,M., Vohringer,H., Haunsberger,S.J. & Soding,J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473.
6. Eddy,S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4, e1000069.

7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.
8. Sali,A. & Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.
9. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202.
10. Chaudhury,S., Lyskov,S. & Gray,J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 26, 689-691.

Assembly prediction in CASP14 with pyDock *ab initio* docking and scoring

M. Rosell[1,2], L.A. Rodríguez-Lumbreras[1,2] and J. Fernández-Recio[1,2]

[1] - Instituto de Ciencias de la Vid y del Vino (CSIC), [2] - Barcelona Supercomputing Center (BSC)
juan.fernandezrecio@icvv.es

***Key:** Auto:Y; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y*

In the past 3rd common CASP-CAPRI Assembly Prediction challenge, our modeling approach, integrating *ab initio* docking, template-based modeling, distance-based restraints, low-resolution structural data and symmetry constraints, yielded excellent performance, ranking 2nd among CAPRI predictors, and 1st among CAPRI scorers[1]. Here we describe our participation in the CASP14 Assembly category, as part of the 4th common CASP-CAPRI Assembly Prediction challenge (CAPRI Round 50). We have participated as human predictors, human scorers, and server scorers, in all the 18 proposed targets, consisting in four hetero-dimers (A1B1), six homo-dimers (A2), two homo-trimers (A3), two homo-tetramers (A4), one hetero-nonamer (A3B3C3), one homo-20mer (A20), one hetero-27mer (A6B3C12D6), and one homo-240mer (A240).

Methods

For each assembly, the models of the individual subunits were taken from the ZHANG, RaptorX, and QUARK CASP-hosted servers (only the best prediction for each server was used). In CAPRI target ID T170 (CASP target ID H1060), the experimental structures of two of the subunits were available (see more details in Results section). In two other cases (T165/H1036 and T177/H1081), there were not available models at the CASP-hosted servers for some of the subunits, so we modelled them with MODELLERv9.19.

Using the available structural models of the individual subunits as above described, we modelled all or some binary interactions in the assembly by *ab initio* docking (usually docking one pair of models from each CASP-host server). As human group we applied our pyDock[2] docking and scoring pipeline, in which we used FTDock (electrostatics on; 0.7 Å grid resolution) and ZDOCK 2.1 to generate 10,000 and 2,000 rigid-body docking poses, respectively, which were merged in a single pool for subsequent pyDock energy-based scoring. We also participated with our pyDockWEB server[3]. In homo-oligomers, docking poses not satisfying the expected symmetry (e.g. C_2 for homo-dimers, C_3 for homo-timers, etc.) were removed.

Additionally, we checked if there were available templates for all or part of the assembly interfaces. First, we used BLAST for this. In parallel, for each complex we searched for oligomeric templates from the top five released predictions from the ZHANG, QUARK, RaptorX, MULTICOM-CONSTRUCTand ROSETTA CASP-hosted servers. These monomeric models were superimposed onto the corresponding subunits of each selected template and minimized with AMBER 12.

Finally, all the generated models (either *ab initio* or template-based) were scored with pyDock, and sorted according to the summation of the binding energy of all possible interfaces. The number of available templates and their reliability determined the percentage of template-based complex models included in the top 5 and 10 submitted models. Finally, we eliminated the redundant predictions and minimized the top ten submitted models.

In the scorers experiment, we first removed models with more than 250 clashes (i.e., intermolecular pairs of atoms closer than 3 Å). Then, we applied pyDock scoring and used the same criteria to rank the docking models as in predictors (i.e. in case of reliable templates we favored models similar to such templates, we checked for symmetry, we applied *ad-hoc* distance restraints for specific targets, etc., more details in the Results section). As human scorers we introduced more human intervention than as server scorers, i.e., removing loops with non-realistic conformations, and re-scoring some of these models afterwards.

Results

We submitted models generated only by *ab initio* docking in those targets for which we could not find available templates (T169/T1054, T172/H1066, T173/H1069, T174/T1070, T178/T1083, T179/T1087, and T181/H1103). In the case of T174/T1070, all the template-based models we built had clashes, so as predictors only *ab initio* docking models were submitted, but as scorers we favored models similar to the available templates. In the case of T181/H1103, additional restraints were applied to remove poses clashing with the membrane regions.

But for the majority of targets, we could find potentially suitable templates for all or some of the predicted interfaces. In many cases, we generated models by *ab initio* docking and by template-based modeling independently, and the final proportion of models derived from these two approaches was determined by pyDock scoring and/or by the reliability of the available templates. Thus, *ab initio* docking was favored in targets T164/T1032 and T176/T1078, while template-based modeling was favored in targets T166/H1045, T167/T1050, and T168/T1052.

On the other side, in target T180/T1099, consisting in the assembly of a virus capsid with icosahedral symmetry, only template-based modeling was used.

In the remaining targets, in order to build the full assembly we combined template-based docking for some interfaces and *ab initio* docking for the other ones. This is the case of T177/H1081 (A20), in which the homo-decamer (A10) was modelled based on available templates followed by pyDock scoring, and the final assembly was built by docking two decamers. Similarly, in homo-tetrameric targets T171/T1063 and T175/T1073 (A4), one of the dimeric interfaces was modelled based on available templates and then the modelled dimers were docked to built the full assembly. In the homo-nonameric target T165/H1036 (A3B3C3), the first homo-trimer (A3) and one hetero-dimer (BC) were built based on available templates. Then, they were docked to form partial complexes (A3BC) and the full assembly was finally built by symmetry.

In the same line, target T170/H1060 was a challenging hetero-27mer, in which we applied an *ad-hoc* modeling procedure, also combining *ab initio* docking and template-based modeling. This assembly was formed by three rings with different composition and stoichiometry. The first ring was a homo-hexamer arranged as a dimer of trimers (2xA3) and was modelled by fitting two copies of the homo-trimeric x-ray structure (PDB 5NGJ) to available Cryo-EM data (EMDB ID: EMD-3689), followed by minimization. The second ring was formed by three subunits of one protein and twelve subunits of a second protein (B3C12) and was modelled using available monomeric models from CASP-host servers. Basically, the homo-trimer (B3) was modelled by building dimers (B2) with *ab initio* docking and generating the trimer by symmetry, followed by minimization and re-scoring by pyDock. The homo-trimer (B3) was docked to monomeric models of the second protein (C) obtained from CASP-hosted servers, and then B3C3 models were built by symmetry. In parallel, tetramers of the second protein (C4) were modelled by building dimers

(C2) with *ab initio* docking and superimposing them on a tetrameric template (PDB 4BEG). The hetero-15meric ring (B3C12) was finally built by superimposing the homo-tetrameric models of the second protein (C4) onto each C subunit in the docking models (B3C3). The third homo-hexameric ring (D6) was modelled by superimposing the x-ray structure of the monomer (PDB 4JMQ) on available templates (PDB 4DIV and 2X8K), followed by minimization and pyDock scoring. The final assembly of the modelled rings was done with the help of *ab initio* docking, selecting only models in which the symmetry axes of the rings were aligned. The same criteria was used in the scorers experiment.

Availability

The pyDock 3.0 program is available for academic use as a GNU/Linux binary and as a web server (https://life.bsc.es/pid/pydock/).

1. Lensink,M.F., Brysbaert,G., Nadzirin,N., et al. (2019). Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. Proteins. 87, 1200-1221.
2. Cheng,T.M.-K., Blundell.T.L. & Fernandez-Recio,J. (2007) pyDock: electrostatics and desolvation for effective rigid-body protein-protein docking. Proteins. 68, 503-515.
3. Jimenez-Garcia,B., Pons,C. & Fernandez-Recio,J. (2013) pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics. 29, 1698-1699.

Frishman_Group

DeepHelicon: Accurate prediction of inter-helical residue contacts in transmembrane proteins by residual neural networks

Jianfeng Sun, Dmitrij Frishman
Department of Bioinformatics, Wissenschaftzentrum Weihenstephan, Technische Universität München, 85354, Freising, Germany
d.frishman@wzw.tum.de

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N.*

DeepHelicon is specialized for predicting inter-helical residue contacts in transmembrane proteins in CASP14. DeepHelicon only takes as input a protein sequence in FASTA format. Residues located in the transmembrane regions are detected by the TMHMM2.0 algorithm[1].

Methods

Accurate prediction of amino acid residue contacts is an important prerequisite for generating high-quality 3D models of transmembrane (TM) proteins[2-3]. While a large number of compositional, evolutionary, and structural properties of proteins can be used to train contact prediction methods, recent research suggests that coevolution between residues provides the strongest indication of their spatial proximity[5]. We have developed a deep learning approach, DeepHelicon[4], to predict inter-helical residue contacts in TM proteins by considering only coevolutionary features. DeepHelicon comprises a two-stage supervised learning process by residual neural networks[6] for a gradual refinement of contact maps, followed by variance reduction by an ensemble of models[4].

Results

We present a benchmark study of 12 contact predictors and conclude that DeepHelicon together with the two other state-of-the-art methods DeepMetaPSICOV[6] and Membrain2[2] outperforms the 10 remaining algorithms on all datasets and at all settings[4]. On a set of 44 TM proteins with an average length of 388 residues DeepHelicon achieves the best performance among all benchmarked methods in predicting the top $L/5$ and $L/2$ inter-helical contacts, with the mean precision of 87.42% and 77.84%, respectively. On a set of 57 relatively small TM proteins with an average length of 298 residues DeepHelicon ranks second best after DeepMetaPSICOV. DeepHelicon produces the most accurate predictions for large proteins with more than 10 transmembrane helices. Coevolutionary features alone allow to predict inter-helical residue contacts with an accuracy sufficient for generating acceptable 3D models for up to 30% of proteins using a fully automated modeling method such as CONFOLD2[7]. DeepHelicon is specialized for transmembrane proteins in CASP14. The multiple sequence alignments (MSAs) of transmembrane proteins were generated using HHblits[8].

Availability

The standalone DeepHelicon software is available at https://github.com/2003100127/deephelicon.

1. Krogh,A., Larsson,B., Heijne,G. & Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J. Mol. Biol. 305, 567-580.
2. Yang,J. & Shen,H.B. (2018) MemBrain-contact 2.0: a new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. Bioinformatics 34, 230-238.
3. Hönigschmid,P. & Frishman,D. (2016) Accurate prediction of helix interactions and residue contacts in membrane proteins. J. Struct. Biol. 194, 112-123.
4. Sun,J. & Frishman,D. (2020) DeepHelicon: Accurate prediction of inter-helical residue contacts in transmembrane proteins by residual neural networks. J. Struct. Biol. 212, 107574.
5. Jones,D.T. & Kandathil,S.M. (2018) High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 34, 3308-3315.
6. Kandathil,S.M., Greener,J.G. & Jones.D.T. (2019) Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins. 87, 1092-1099.
7. Adhikari,B. & Cheng,J. (2018) CONFOLD2: improved contact-driven ab initio protein structure modeling. Proteins. 19, 22.
8. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173-175.

GAPF_LNCC_SERVER: an automated template based de novo protein structure prediction method with a multiple minima genetic algorithm

F.L. Custódio and L.E. Dardenne
Laboratório Nacional de Computação Científica-LNCC/MCTI, Petrópolis-RJ, Brasil
flc@lncc.br

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:Y.v; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

This is exactly the same method as GAPF_LNCC_SERVER, but some targets required longer conformation runs and could not be completed in the allotted server time. This occurred specially at the beginning of the prediction season.

We built a fully automated template based *de novo* PSP method that can be easily integrated into a web server, (GAPF_LNCC_SERVER workflow). At the time of CASP14 we had limited hardware to host a server, therefore some targets, of the server-only category, could not be completed in time. The workflow is based on three ideas: (i) the use of experimental information available, in the form of templates, residue-residue contact prediction, distances histogram prediction, secondary structure prediction, and fragments; (ii) a multiple minima genetic algorithm for conformational search; and a (iii) knowledge based/physics based scoring function.

We employ a coarse-grained representation where all backbone atoms are explicit, with the side chains modeled as a single superatom. The scoring function combines some physically realistic potential with knowledge-based terms to promote hydrogen bonding, secondary structure organization and inter-residues distance restrictions. Global optimization is carried out by the multiple-minima genetic algorithm (GA) and no further refinement is performed. Selection of the models is then done by means of structural redundancy filtering and energy pruning. The GAPF_LNCC_SERVER workflow was applied to all targets. Because we were limited to a desktop with a single GPU to host the server, inter-residues distance histograms were used only on targets with less than 260 residues. A "template-based *de novo*" strategy was used when suitable templates were found.

Methods

All accessory programs and tools are run locally on a desktop PC. The conformational search step runs at the Santos Dumont cluster (https://sdumont.lncc.br). Our workflow starts with (1) secondary structure prediction by PSIPRED[1] followed by (2) templates search with HHBlits[2]. (3) Residue-residue (RR) contacts prediction, and distances histogram, are made by an pytorch alphafold implementation and for targets larger than 260 residues DeepCov[4] is executed locally for RR contacts prediction. (4) Fragment libraries are created with Profrager[5] (https://www.lncc.br/sinapad/Profrager/), and fragments are selected using the secondary structure prediction, in addition to the local sequence similarities from a culled database of 34,750 chains from experimental structures.

The (5) conformational search carried out by GAPF[6] employs a genetic algorithm (GA) with seven genetic operators including Ramachandran based mutations[7] and fragment insertion. The GA methodology uses a scoring function with a proper dihedral, steric repulsion, hydrophobic compaction, hydrogen bonding formation[8], cooperative hydrogen bonding[9], RR contacts[10] and distance histograms, when available. GAPF employs a phenotype-based crowding mechanism for the maintenance of useful diversity within the populations, which has been shown to result in increased performance and to grant the algorithm multiple solution capabilities. For each target, at most 100 independent runs of the GA ware performed (dependant on time restraints) and each population contains 200 individuals, resulting in 20,000 structures. These results undergo a (6) structural redundancy filter and the overall top five structures, ranked by energy, proceeded to the next steps. (7) Side chains of the select structures are reconstructed using SCWRL4[11]. And finally, the files are (9) formatted according to CASP guidelines, including (10) filling the temperature column of the PDB files with the confidence in the prediction (0-1, where 0 is the worst). Templates were sought using HHblits[12] and those found with probabilities larger than 70% are used to seed the initial populations of the genetic algorithm.

Availability

All tools are freely available from their authors. Some steps in the protocol were carried out with experimental versions of our software that will be made available by contacting the authors. The fully functional web server should open by mid-2021. Acknowledgment: FAPERJ grant E-26/210.935/2019, Santos Dumont supercomputer (http://www.sdumont.lncc.br/). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

1. McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. Bioinformatics, 16(4), 404-405.
2. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2), 173–175.
3. Wang, S., Sun, S., & Xu, J. (2018). Analysis of deep learning methods for blind protein contact prediction in CASP12. Proteins: Structure, Function, and Bioinformatics, 86, 67–77.
4. Jones, D. T., & Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics. http://doi.org/10.1093/bioinformatics/bty341
5. Trevizani, R., Custódio, F. L., dos Santos, K. B., & Dardenne, L. E. (2017). Critical features of fragment libraries for protein structure prediction. PloS one, 12(1), e0170131.
6. Custódio, F. L., Barbosa, H. J., & Dardenne, L. E. (2014). A multiple minima genetic algorithm for protein structure prediction.Applied Soft Computing, 15, 88-99.
7. Santos, K. B., Custódio, F. L., Barbosa, H. J., & Dardenne, L. E. (2015, August). Genetic operators based on backbone constraint angles for protein structure prediction. InComputational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on (pp. 1-8). IEEE.
8. Rocha, G. K., Custódio, F. L., Barbosa, H. J. C., & Dardenne, L. E. (2015, August). A multiobjective approach for protein structure prediction using a steady-state genetic algorithm with phenotypic crowding. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on (pp. 1-8). IEEE.
9. Levy-Moonshine, A., Amir, E. A. D., & Keasar, C. (2009). Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics, 25(20), 2639-2645.
10. Santos, K. B., Rocha, G. K., Custodio, F. L., Barbosa, H. J. C., & Dardenne, L. E. (2017). Improving de novo protein structure prediction using contact maps information. In 2017 IEEE

Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1–6). IEEE. http://doi.org/10.1109/CIBCB.2017.8058535

11. Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function, and Bioinformatics, 77(4), 778-795.

GLoSA

Refinement Using Molecular Dynamics with Restraints Derived from Binding Site Templates

H. Guterres and W. Im
Department of Biological Sciences, Lehigh University
hsg218@lehigh.edu

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:N.MetaG; Fragm:Y.v; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

Molecular dynamics simulations are well-known as a reliable method for protein structure refinements. Here, we shift the focus of protein structure refinement on specific regions of the protein that are likely to have variable structures as compared to their native structure, namely the ligand binding sites. We have shown previously that by performing MD guided refinement on protein's binding site, the overall protein structures were also refined.[1] Motivated by the structure refinement results in that paper, we apply the same method to this CASP14 blind refinement targets.

Methods

We performed physics-based all-atom MD simulations in explicit solvents for the target proteins with restraints derived from binding site templates. We used our computational tool, G-LoSA (https://compbio.lehigh.edu/GLoSA/toolkit.html). Using G-LoSA we align all available binding site templates from the PDB library onto each target protein. G-LoSA search aligns local structures onto a target protein in a sequence-independent manner and calculate their similarity using GA-score. Top templates with high GA-scores are selected. From the selected templates, we identified aligned residues on the target protein as the equivalent residues. We then calculate the distance matrix between C-alpha atoms of the selected residues and derive a harmonic distance restraint potential. For our MD simulations, we apply force constants of 1.5 kcal/(molÅ^2) for distance restraints and additionally we apply a weak positional restraint of 0.1 kcal/(molÅ^2) to all the remaining alpha carbons, based on the same protocol we applied in our previous refinement work.[1] For each target we perform 3 x 50 ns production run, each started from the same initial structure but using different initial velocity random seeds. Simulations are carried out using OpenMM and CHARMM36m force fields. The refined structure is the average of the three final conformations from the simulations. For targets that allow extended simulation runs, we extended simulation time to 1 microseconds each.

Availability

G-LoSA (Graph-based Local Structure Alignment) is a computational tool for binding site predictions and similarity measurement that is freely available on https://compbio.lehigh.edu/GLoSA/toolkit.html.

1. Guterres,G., Lee,H., Im,W. (2019). Ligand-binding-site structure refinement using molecular dynamics with restraints derived from predicted binding site templates. J Chem Theory Comput. 15, 6524-6535.

graph-sh

Spherical convolutions on molecular graphs for protein model quality assessment

I. Igashov[1,2], N. Pavlichenko[2], and S. Grudinin[1]

[1] - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, [2] - Moscow Institute of Physics and Technology, 141701 Dolgoprudniy, Russia

Sergei.Grudinin@inria.fr

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

The graph-sh server is built upon the spherical graph convolutional network (S-GCN)[1]. S-GCN is a single-model QA method based on a deep neural network that processes protein molecules represented as unordered graphs. S-GCN operates on geometric information retrieved from 3D *Voronoi tessellation* of a protein model. The key idea of the proposed method is the ability to construct *rotation-equivariant* local coordinate systems associated with each residue in a protein.

Methods

Our method operates on three-dimensional protein graphs. In these graphs, the nodes correspond to the protein residues, and the edges correspond to the contact surface areas between the residues, which are computed using the Voronota[2] framework. Each node is associated with several geometric features, such as the type of the residue, the volume of the corresponding Voronoi cell, the solvent-accessible surface area of the corresponding residue, and the buriedness, which is the graph distance to the nearest solvent-accessible residue. Also, for each node **u** and all of its neighbours **v**, we computed spherical coordinates of a projection of the neighbour **v** onto a unit sphere with the center at **u**. These spherical coordinates are calculated in a local coordinate system associated with the node **u**. We unambiguously construct these local coordinate systems for all the residues in a protein model using the topology of the protein backbone, following our previous model Ornate[3].

To process such graphs, we constructed a graph neural network and trained it on local CAD-scores[4]. Our network consists of spherical convolution layers, batch normalization, and dropout layers. The spherical convolution layer performs a convolution operation in each node over all its neighbours. A convolution operation is constructed using precomputed relative coordinates and spherical harmonics as the basis functions. The essential component of the spherical convolution layer is a trainable filter, a spherical function that is represented as a combination of spherical harmonics up to the 5th order and trainable matrices (expansion coefficients). To obtain a prediction of a model's global CAD-score, we average local scores predicted by the network.

We trained S-GCN on the CASP$^{8\text{-}12}$ datasets and validated it on the CASP13 dataset. For training, the data from CASP$^{8\text{-}12}$ was preliminarily refined: we removed excessive models' parts and filtered out targets of low quality (based on VoroMQA[5] predictions). To enrich the data with more near-native examples, we generated additional near-native conformations using the NOLB[6] library.

Results

For the CASP14 predictions, we implemented the graph-sh method as an automatic server with the following workflow. First of all, a graph is built using the Voronota framework. Then, the spherical harmonics up to the 5th order are computed for all the residues and their neighbours using an in-house C++ library and the code derived from our previous method Ornate. Finally, the pre-trained PyTorch model is applied, which predicts local CAD-scores.

In our paper[1] we show that the S-GCN model outperforms the current state-of-the-art single-model methods on the CASP13 dataset in several metrics. We also mention that S-GCN uses only geometric information, and does not take into account any biological, chemo-physical, and evolution descriptors. We also compared S-GCN with a baseline message-passing graph neural network without spherical filters and observed a huge performance gap that can not be eliminated via fine-tuning. This fact can be considered as a proof of concept for a spherical convolution filter we propose in our work.

Availability

More details about S-GCN can be found at https://team.inria.fr/nano-d/software/sgcn/.

1. Igashov,I., Pavlichenko,N., & Grudinin,S. (2020). Spherical convolutions on molecular graphs for protein model quality assessment. Submitted.
2. Olechnovič,K., & Venclovas,Č. (2014). Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry, 35(8), 672-681.
3. Pagès,G., Charmettant,B., & Grudinin,S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics, 35(18), 3313-3319.
4. Olechnovič,K., Kulberkytė,E., & Venclovas,Č. (2013). CAD ‐ score: a new contact area difference ‐ based function for evaluation of protein structural models. Proteins: Structure, Function, and Bioinformatics, 81(1), 149-162.
5. Olechnovič,K., & Venclovas,Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins: Structure, Function, and Bioinformatics, 85(6), 1131-1145.
6. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. Journal of chemical theory and computation, 13(5), 2123-2134.

HMS-Casper

Recurrent geometric networks using Frenet-Serret geometry and latent residue representations

Ratul Chowdhury[1], Nazim Bouatta[1], Surojit Biswas[1,2], and Mohammed AlQuraishi[1]
[1]*Laboratory of Systems Pharmacology, Harvard Medical School*
[2]*Nabla Bio, Inc.*
ratul_chowdhury@hms.harvard.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y.*

A novel version of the Recurrent Geometrical Network (RGN[1]) algorithm, which geometrically reasons over protein conformations, is used to predict protein structures. Two options are considered for inputs: (i) the raw amino acid sequence and position-specific scoring matrix (PSSM) of each protein and (ii) a context-based encoding of amino acid residues – AminoBert – derived strictly from raw amino acid sequences without making explicit use of any evolutionary information. Raw RGN structure predictions are subsequently refined using an energy-minimization protocol subject to dihedral constraints computed from family sequence alignments.

Methods

One-dimensional curves, in differential geometry, are described by the Frenet-Serret geometries (FSG). We use an improved version of the previously reported RGN[1], which parameterized protein backbones (C_α atoms) using dihedral angles, that leverages the fact that protein backbones are intrinsically discrete one-dimensional curves. The improved version implements a transfer matrix formalism which enables reasoning over protein backbones using a discrete version of Frenet-Serret geometries (dFSG[2]).

Inputs: dFSG-based RGNs are used with two different possible inputs to predict protein backbones:

(i) Raw amino acid sequences and PSSMs, as previously described [1]

(ii) *AminoBert*: a reformulated version of the BERT language model[3] is used to train a transformer[4] over protein sequences to predict missing amino acids conditioned on the flanking sequence. Amino acid residues are thus mapped onto a higher-dimensional representation.

Refinement: Raw structure predictions from dFSG-based RGNs, trained with sequence+PSSM (HMSCasper-PSSM) or AminoBERT representations (HMSCasper-Seq) are refined using a Rosetta-based protocol that first builds the remaining atoms and then alleviates steric clashes and fine-tunes folded domains. As an additional possibility, constraints coming from an orientogram populated with pairwise angular dependencies between residues, derived from a family sequence alignment and trRosetta, are imposed during energy minimization of the structure (HMSCasper-MSA).

Training: For training we used (a) the dFSG-based RGN model trained on ProteinNet12 dataset (comprising UniParc + JGI metagenomes and PDB) with sequence+PSSM inputs for making predictions under the HMSCasper-PSSM group and (b) AminoBERT models trained on SCOPe datasets for making predictions under the HMSCasper-Seq group.

Availability

Source code for training dFSG-based RGN models as well as trained models, including PSSM and AminoBert based versions as used for the CASP14 experiment, will be available on GitHub.

1. M.AlQuraishi, End-to-End Differentiable Learning of Protein Structure, Cell Syst. (2019). https://doi.org/10.1016/j.cels.2019.03.006.
2. S.Hu, M.Lundgren, A.J.Niemi, Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. (2011). https://doi.org/10.1103/PhysRevE.83.061908.
3. J.Devlin, M.W.Chang, K.Lee, K.Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., 2019.
4. A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, Ł.Kaiser, I.Polosukhin, Attention is all you need, in: Adv. Neural Inf. Process. Syst., 2017.

Template-based and Contact-assisted Docking for Homo-oligomeric Targets

Yumeng Yan[1], Hao Li[1] and Sheng-You Huang[1]

[1] – *School of Physics, Huazhong University of Science and Technology, Wuhan, China, 430074*
huangsy@hust.edu.cn

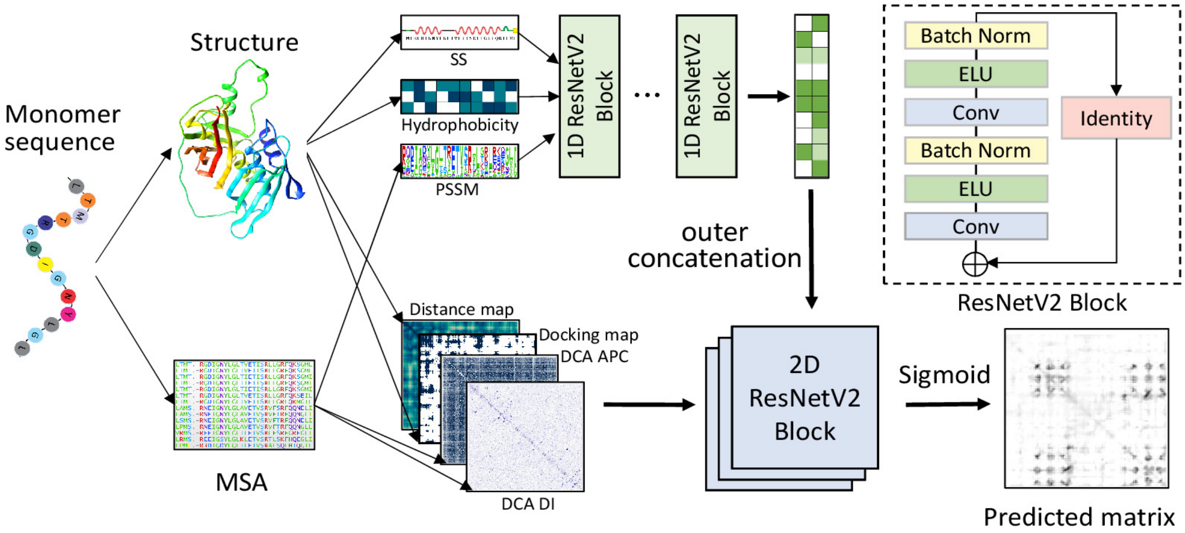
Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Protein-protein interactions play a fundamental role in all cellular processes. Therefore, determining the structure of protein-protein complexes is crucial to understand their molecular mechanisms and develop drugs targeting the protein-protein interactions. Although template-based docking has been well developed and demonstrated high accuracy in recent CASP-CAPRI challenges for those targets that have a template[1], it is still difficult to predict the structure of those protein-protein complexes that do not have a good template. A major portion of protein-protein interactions are formed by homo-oligomers. Recently, we have proposed a deep learning model to predict inter-protein residue-residue contacts across homo-oligomeric protein interfaces, named as DeepHomo[2], by integrating evolutionary coupling, sequence conservation, distance map, docking pattern, and physic-chemical information of monomers. In CASP 14, we have tested both our template-based docking method and contact-assisted docking approach on those homo-oligomeric targets.

Methods

For a given monomer sequence, we first used the HHblits program[3] to search the PDB database. It there was a homo-oligomer protein complex that has a sequence identity of 20% and a sequence coverage of 80% with the target sequence, we would use the complex as a template to construct the homo-oligomeric complex structure for the target through homology modeling using MODELLER[4]. The constructed complex structure was further refined to remove sever atom clashes through a short MD minimization using AMBER[5]. More homo-oligomeric complex structures would be created by our HSYMDOCK symmetric docking method[6] based on the monomer structure.

If there was no appropriate template available for a target, we would use our contact-assisted docking protocol to predicted the homo-oligomeric complex structure of monomers using the contacts predicted by our DeepHomo model[2]. The following figure shows the workflow of our deep learning-based inter-protein contact prediction for homo-oligomeric complexes. DeepHomo is designed to take full advantage of both the structure and sequence information of monomers, which can also be grouped into 1D sequential and 2D pairwise features. On one hand, the 1D sequential features, including the secondary structure (SS), hydrophobicity, and position-specific scoring matrix (PSSM) information, are first extracted from the monomer structure/sequence. Next, the 1D ResNet CNN is used to learn the high-dimensional features from the 1D sequential features. Then, a 2D pairwise matrix is constructed by outer concatenation from the high-dimensional sequential features. On the other hand, the 2D features, including the intra-residue distance map, docking map, and direct coupling analysis (DCA) scores are obtained from the structure and MSA of the monomer. Then, these 2D matrices plus the previously converted 2D map from sequential features are fed into a 2D ResNet CNN for training, resulting in the final matrix of predicted contacts. Finally, the predicted contacts by DeepHomo were integrated into our ab initio HSYMDOCK symmetric docking program[6] to predicted the homo-oligomeric complex structure from monomers.



Results

Our DeepHomo model was extensively tested on both experimentally determined structures and realistic CASP-CAPRI targets. It was shown that DeepHomo achieved a high accuracy of >60% for the top predicted contact and outperformed state-of-the-art direct-coupling analysis (DCA) and machine learning (ML)-based approaches. Integrating predicted contacts into protein docking with blindly predicted monomer structures also significantly improved the success rate from 42.9% to 64.3% on 28 realistic CASP-CAPRI targets when the top five binding modes were considered.

Availability

HSYMDOCK is freely available for academic use at http://huanglab.phys.hust.edu.cn/hsymdock/. DeepHomo is freely available for academic use through http://huanglab.phys.hust.edu.cn/DeepHomo/.

1. Yan Y, Tao H, He J, Huang S-Y. The HDOCK server for integrated protein-protein docking. Nat Protoc. 2020;15(5):1829-1852.
2. Yan Y, Huang S-Y. Accurate prediction of residue-residue contacts across homo-oligomeric protein interfaces through deep leaning. bioRxiv 2020.09.13.295196; doi: https://doi.org/10.1101/2020.09.13.295196
3. Remmert M, Biegert A, Hauser A, Sooding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173-175.

4. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29:291-325.
5. Case DA, Cheatham TE 3rd, Darden T, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26(16):1668-1688. doi:10.1002/jcc.20290
6. Yan Y, Tao H, Huang S-Y. HSYMDOCK: a docking web server for predicting the structure of protein homo-oligomers with Cn or Dn symmetry. Nucleic Acids Res. 2018;46(W1):W423-W431.
7. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol. 2017;13(1):e1005324.

Contact map prediction using a residual and fully convolutional neural network

C.A. Taylor[1] and J. Bacardit[1]

[1] – The Interdisciplinary Computing and Complex BioSystems (ICOS) research group, School of Computing, Newcastle University, UK

c.a.taylor2,jaume.bacardit@newcastle.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:N; Cont:Y; Dist:N; Tors:Y; DeepL:Y; EMA:N; MD:N*

Improving the accuracy of protein contact predictions has remained a key objective within bioinformatics within the past years. The biggest improvement to this was observed within CASP13 where deep learning models greatly outperformed all other models, though these models are still not sufficiently accurate to replace direct measuring procedures such as X-ray crystallography. The usefulness of protein contact maps in protein structure determination has motivated this research to further optimise prediction accuracy.

Methods

Our method uses a fully convolutional deep residual neural network architecture. Our neural network architecture uses a 1x1 2D convolutional layer to carry out a dimensionality reduction from an initial 256 features down to 64. These features are then passed through 12 residual blocks each consisting of two 2D convolutional layers with filter widths of 5 and 3, the number of features remains fixed at 64 throughout these blocks. The first seven residual blocks use increasingly large dilation factors that doubles with each block and both layers in each block use the same dilation factor – the first block has a dilation factor of 1, the second uses a factor of 2, up to the seventh which uses 64, the final 5 residual blocks have a dilation factor of 1. There is group normalization[1] used following each layer in these blocks as well as the initial dimensionality reduction layer with 4 groups total (16 channels per group). The final layer in the architecture is a 1x1 2D convolutional layer that performs a dimensionality reduction from 64 down to 1 and is followed by an Instance normalization[2] layer. ReLU is used as an activation function for all neural layers except the final layer and the network is trained with binary cross-entropy with a built in sigmoid calculation as a loss function and Adam[3] as the optimizer.

The models were trained on an extended dataset by including chains with missing residues. An initial list of protein chains was obtained using the PISCES[4] sequence culling server with a percentage identity cut-off of 30%, a resolution of less than 2.5Å, and an R-factor less than 1. Chains with less than 30 residues or greater than 700 residues were filtered along with chains that have more than 35 missing residues or have more than 15% of their residues missing. The positions of missing residues within these chains were predicted using an energy-based optimization algorithm resulting in a dataset size of 11627 chains. This was done to allow our models to attempt to capture the evolutionary information in the additional chains that other models may have missed.

The input features for the models were generated through the use of multiple sequence alignments to produce a pairwise information matrix to allow the models to capture the coevolution between each pair of residues. The multiple sequence alignments were generated using PSI-BLAST[5] against the Uniref90[6] database, HHBlits[7] against the Uniclust30[8] database, and

Jackhmmer[9] against the Uniref90 database. The input features include coevolutionary couplings generated using CCMpred[10], contact scores and mutual information scores from Freecontact[11], various statistical information from alnstats (MetaPSICOV[12]), and solvent accessibility predictions, secondary structure predictions, torsion angle predictions, and contact density predictions generated using Brewery[13]. The PSSM from PSI-BLAST and the HHM from HHBlits multiple sequence alignment searches were also included as input features.

Results

On CASP13 targets our best performing model achieved a long-range L/5 precision of 0.88 as opposed to the leading CASP13 models TripletRES[14] and RaptorX-Contact[15] having precisions of 0.74 and 0.77 respectively.

It was also observed that once a neural network had a sufficient receptive field, it was difficult to further improve the performance of that model even with extreme changes to the architecture which suggests that the bottleneck to our models contact map prediction performance existed within the data rather than the arrangement of convolutional layers within the networks. However, both the depth of our neural network models and the batch size was limited by the amount of RAM available on the GPU (12 GB) that the models were trained on which also potentially limited the performance of our models.

Availability

A paper on this method will be submitted for publication soon. A web interface and source code for our method will be made public at that point.

1. Wu,Y. & He,K. (2018) Group Normalization.
2. Ulyanov,D., Vedaldi,A. & Lempitsky,V. (2016) Instance Normalization: The Missing Ingredient for Fast Stylization.
3. Kingma,D.P. & Ba,J. (2014) Adam: A Method for Stochastic Optimization.
4. Wang,G. & Dunbrack,R.L. (2005) PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res. 33, W94−W98.
5. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389−3402.
6. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. & Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926−932.
7. Steinegger,M., Meier,M., Mirdita,M., Voehringer,H. & Haunsberger,S.J. (2019) HH-suite3 for fast remote homology detection and deep protein annotation. bioRxiv doi:10.1101/560029.
8. Mirdita,M. et al. (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 45, D170−D176.
9. Finn,R.D., Clements,J. & Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29−W37.

10. Seemayer,S., Gruber,M. & Söding,J. (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30, 3128−3130.
11. Kaján,L., Hopf,T.A., Kalaš,M., Marks,D.S. & Rost,B. (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15, 85.
12. Jones,D.T., Singh,T., Kosciolek,T. & Tetchner,S. (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31, 999−1006.
13. Torrisi,M. & Pollastri,G. (2020) Brewery: deep learning and deeper profiles for the prediction of 1D protein structure annotations. Bioinformatics 36, 3897−3898.
14. Li,Y., Zhang,C., Bell,E.W., Yu,D. & Zhang,Y. (2019) Ensembling multiple raw coevolutionary features with deep residual neural networks for contact ‐ map prediction in CASP13. Proteins Struct. Funct. Bioinforma. 87, 1082−1091.
15. Wang,S., Sun,S., Li,Z., Zhang,R. & Xu,J. (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Comput. Biol. 13, e1005324.

Fully Automated Prediction of Protein Tertiary Structures with Local Model Quality Scores Using the IntFOLD6 Server

L.J. McGuffin[1]
1 - School of Biological Sciences, University of Reading, Reading, UK
l.j.mcguffin@reading.ac.uk

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:Yv; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The IntFOLD server[1] integrates our latest methods for: tertiary structure (TS) prediction, domain boundary prediction, prediction of intrinsically disordered regions, prediction of protein-ligand interactions and the global and local quality assessment (QA) of predicted 3D models of proteins. Following the successes of our previous IntFOLD servers[2,3], which used ModFOLD variants[4] to rank models, our primary focus for the IntFOLD6 server at CASP14 was the further improvement of global model ranking and local model quality scoring, using our newly improved ModFOLD8_rank method.

Methods

For CASP14, a bespoke version of the IntFOLD6 server was developed in order to return appropriately formatted results for just the tertiary structure (TS) prediction category. Additionally, the local quality assessment predictions (QMODE3) were returned as scores in the B-factor column of each TS model file. (Predictions in the QMODE1 & QMODE2 QA categories were also returned by our separate servers (see our ModFOLD8 and ModFOLDclust2 abstracts for details.)

Our TS method was developed with the aim of fixing local errors, identified in an initial pool of single template models, through iterative multi-template modeling. The method attempts to exploit our previous CASP successes in accurately predicting local errors in our models[5] by taking the global and local per-residue errors into consideration during the multiple template selection stage[6].

For the initial fold recognition stage, 14 different methods were installed and run in-house to generate up to 10 sequence-to-structure alignments each - resulting in up to 140 alternative single-template based models being generated for each CASP target. The following fold recognition methods were used: SP3[7], SPARKS2[7], HHsearch[8], COMA[9], SPARKSX[10], CNFsearch[11] and the 8 alternative threading methods that are integrated into the current LOMETS package[12] (PPA, dPPA, dPPA2, sPPA, MUSTER, wPPA, wdPPA and wMUSTER).

In the first stage of the IntFOLD6 TS method, all single-template models were assessed using ModFOLDclust2[13] in order to assign global and local model quality scores. Using the single template model quality scores, and other criteria involving template coverage, the corresponding alignments were then selected from each fold recognition method and used to build multiple-template models, with the aim of minimizing local errors in the final models. The alternative multi-template modelling alignment selection methods resulted in the generation of a new population of up to 124 multi-template models for each target. Additionally, I-TASSER *light*[14] (for sequence <500 residues; run in "light mode" with wall-time restricted to 5h), HHpred[15] and DMPfold[16] were used to generate up to 5 models each, which were then added to the final pool of alternative multi-template models for ranking. In the final stage of the method, the models in the final reference set were then evaluated using our new ModFOLD8_rank QA method and the top 5 ranked models were submitted as the final IntFOLD6 TS predictions (see our ModFOLD8 abstract from more details about our ModFOLD8_rank method).

Results

The IntFOLD6 server is continuously benchmarked using the CAMEO resource[17] (identified as server 90). According to the CAMEO results, IntFOLD6 has shown improved performance over our last three methods (IntFOLD3, IntFOLD4 & IntFOLD5) and it is outperformed by just one public server in the benchmark.

Availability

The IntFOLD6 server is available at:
http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD6_form.html.

1. McGuffin,L.J., Adiyaman,R., Maghrabi,A.H.A., Shuid,A.N., Brackenridge,D.A., Nealon,J.O. & Philomina,L.S. (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Res. 47, W408-W413, doi: 10.1093/nar/gkz322
2. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Moult,J., Schwede,T., Tramontano,A. (2018) Evaluation of the template-based modeling in CASP12. Proteins. 86 S1, 321-334. doi: 10.1002/prot.25425.
3. McGuffin, L.J., Shuid, A.N., Kempster, R., Maghrabi, A.H.A., Nealon J.O., Salehe, B.R., Atkins, J.D. & Roche, D.B. (2018) Accurate Template Based Modelling in CASP12 using the IntFOLD4-TS, ModFOLD6 and ReFOLD methods. Proteins. 86 S1, 335-344. doi: 10.1002/prot.25360.
4. Cheng,J., Choe,M.H., Elofsson,A., Han,K.S., Hou,J., Maghrabi,A.H.A., McGuffin,L.J., Menéndez-Hurtado,D., Olechnovič,K., Schwede,T., Studer,G., Uziela,K., Venclovas,Č., Wallner, B. (2019) Estimation of model accuracy in CASP13. Proteins. 87, 1361-137. doi: 10.1002/prot.25767
5. McGuffin,L.J., Roche,D.B. (2011) Automated tertiary structure prediction with accurate local model quality assessment using the IntFOLD-TS method. Proteins. 79 S10, 137-46.
6. Buenavista,M.T., Roche,D.B., McGuffin,L. J. (2012) Improvement of 3D protein models using multiple templates guided by single-template model quality assessment. Bioinformatics. 28, 1851-1857.
7. Zhou,H., Zhou,Y. (2005) SPARKS2 and SP3 servers in CASP6. Proteins. 61 S7, 152-156.
8. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics. 21, 951-96.

9. Margelevičius,M., Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparisons. BMC Bioinformatics. 11, 89.
10. Yang,Y., Faraggi,E., Zhao,H., Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics. 27, 2076-2082.
11. Ma,J., Wang,S., Zhao,F., Xu,J. (2013) Protein threading using context-specific alignment potential. Bioinformatics. 29, i257-65.
12. Wu,S. and Zhang,Y. (2007) LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Research. 35, 3375-3382.
13. McGuffin,L.J. & Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics. 26, 182-188.
14. Roy,A., Kucukural,A., Zhang,Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 5, 725-738.
15. Meier,A., Söding,J. (2015) Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. PLoS Comput Biol. 11, e1004343.
16. Greener,J.G., Kandathil,S.M. & Jones,D.T. (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun. 10, 3977. doi: 10.1038/s41467-019-11994-0
17. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R., Schwede, T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins. 86 S1, 387-398. doi: 10.1002/prot.25431.

JLU_Comp_Struct_Bio

Molecular free energy optimization on a computational graph

Xiaoyong Cao[1] and Pu Tian[1,2]
[1] - School of Life Sciences, Jilin University, Changchun, China 130012, [2] - School of Artificial Intelligence, Jilin University, Changchun, China 130012
tianpu@jlu.edu.cn

Key: *Free energy, computational graph, auto differentiation, coordinates transformation*

Free energy is arguably the most important property of molecular systems. Despite great progress in both its efficient estimation by scoring functions/potentials and more rigorous computation based on molecular simulations, we remain far from accurately predicting and manipulating biomolecular structures and their interactions. There are fundamental limitations, including accuracy of interaction description and difficulty of sampling in high dimensional space, to be tackled. Computational graph underlies major artificial intelligence platforms and is proved to facilitate training, optimization and learning. This new framework greatly improves efficiency by replacing local sampling with differentiation and is demonstrated in protein structure refinement.

We introduce an new refinement protocol combining auto differentiation, coordinates transformation and generalized solvation free energy theory(GSFE)[1]. we construct a computational graph infrastructure to realize seamless integration of fully trainable molecular interaction description with end to end differentiable free energy optimization.

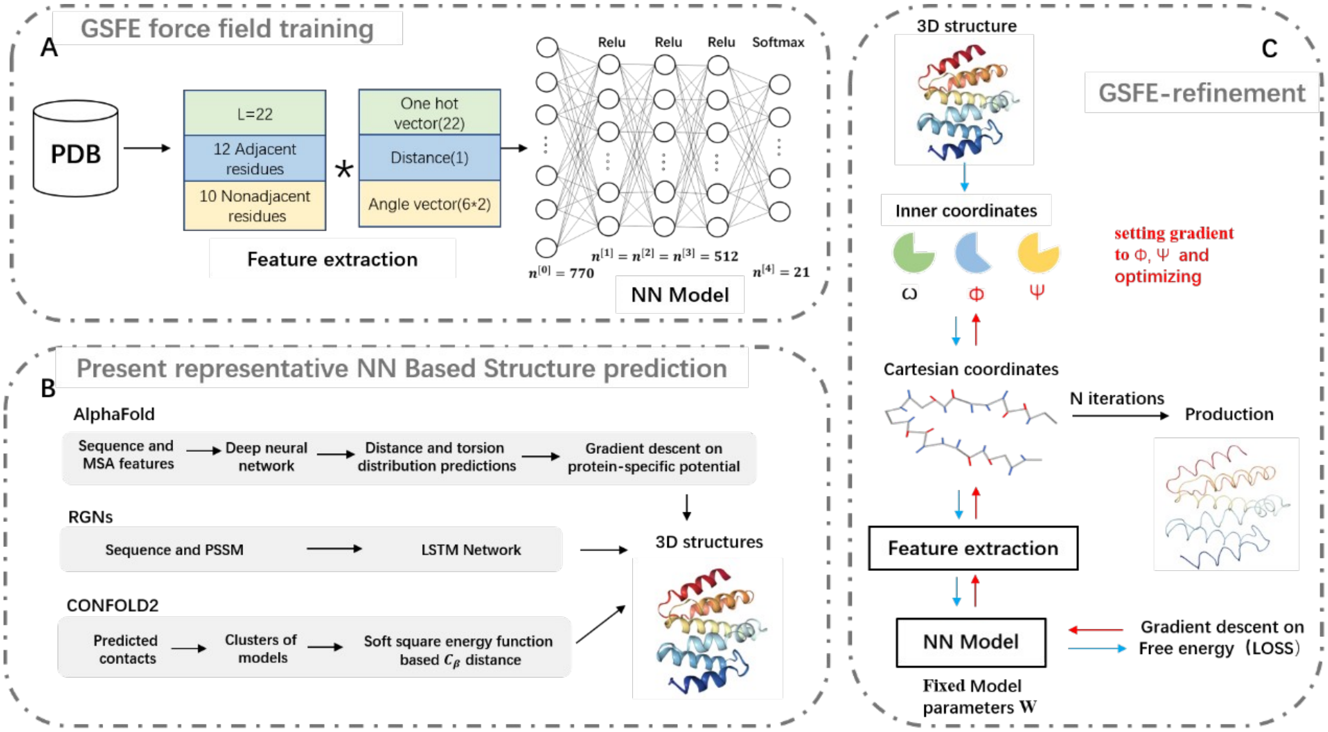
Methods

The GSFE-refinement[2] is a fast refinement protocol. As Figure1 show, Schematic representation of GSFE-refinement contain: (A) The neural network implementation of GSFE, Amino acid identity is used as labels for training; LMLA(local maximum likelihood approximation) is utilized for assessment of structural models. (B) major present protein structure prediction schemes based on NN(Neural Networks). All networks provide a map from sequence information (or contacts predicted from which) to structure. (C) Flowchart of the GSFE-refinement protocol, with Feature extraction and NN Model being the same as that of A). This scheme provides a map between structure and free energy. Through iterative minimization of free energy by differentiation with respect to structure with the NN model fixed, we realize differentiable structure optimization. In casp14, five iterations are made for each starting structure to generate five structures as the final result.

Results

We test our protocol in 31 proteins of CASP12 dataset(We remove the 8 start models don’t have native structure and 3 start models lack amino acids). In the best of top 5 models, the average GDT-HA increase 0.27% and 64.5%(20/31) start models improved. The average improvement of RMSD value is -0.02, and 100% (31/31) start models is successes. All structures are optimized within 170 seconds on a desktop computer , and each structure takes 5.5 seconds on average.

Abstracts - 1



1. Long, S.; Tian, P. A simple neural network implementation of generalized solvation free energy for assessment of protein structural models. RSC Advances 2019, 9 .
2. Xiaoyong Cao.; Pu Tian. (2020) End to end differentiable protein structure refinement. bioRxiv.

Jones-UCL

Manually curated tertiary structure and distance predictions using DMPfold2

S.M. Kandathil, J.G. Greener, A.M.C. Lau and D.T. Jones
Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
d.t.jones@ucl.ac.uk

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; MetaG:Y; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The Jones-UCL group used pipelines similar to those used in the DMP2 group submissions (see abstract for group DMP2) with manual intervention. Additional procedures included manual domain parsing, assembly of multi-domain models and the use of alternative multiple sequence alignments (MSAs).

Methods

The iterative MSA generation procedure was run for 5 iterations. Alternative MSAs were scored using a novel neural net which predicts expected model quality using only the MSA as input. This net was also used in a procedure that attempts to optimise an initial MSA by identifying a subset of sequences which give rise to an improved predicted model quality score. Where MSA optimisation resulted in significant improvements in predicted model quality, additional structure models were built using the optimised MSA. All models produced by the DMP2 group were also considered. Structure models were also built using alternative versions of the DMPfold2 pipeline, most notably including different sets of neural network weights, and procedures for automatically determining optimal threshold parameters for converting distance distributions to NOE distance ranges for CNS and Xplor NIH.

Domain boundaries were determined using one or a combination of: HHsearch scans against the PDB70 database; HHsearch against Pfam; or visual inspection of initial full-chain models. Where domain segmentation was deemed necessary, per-domain models were built using the methods described above. Models for each domain were scored using a combination of MODCHECK and MODELLER DOPE scores[1], and assembled into a full-chain model using MODELLER.

Submissions in the MQA category used a neural net that predicts per-residue and full-chain scores using the Cα coordinates of the query model. Refinement submissions used the iterated restraint generation and structure prediction section of the DMPfold2 pipeline, with Bayesian optimization used to sample bounds from the distance distributions, and the MODCHECK+DOPE score function used as the overall objective for minimization.

Results

Alignment optimization produced significant improvements in predicted MSA quality in a few cases. Automatic threshold parameter determination for distance predictions produced much more robust tertiary structure predictions at the expense of additional compute time (as evaluated on the CASP13 FM domains).

Availability

DMPfold2 will be made available on the PSIPRED GitHub page (https://www.github.com/psipred) under a permissive licence, and also via the PSIPRED Workbench[2] (http://bioinf.cs.ucl.ac.uk/psipred).

1. Greener,J.G., Kandathil,S.M. & Jones,D.T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat. Commun. 10, 3977.
2. Buchan,D.W.A., Jones,D.T. (2019). The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Research. 47, W402-W407.

Distance Prediction, Structure Prediction, Refinement, Quality Assessment, and Protein Docking in KiharaLab

Genki Terashi[1], Charles Christoffer[2], Jacob Verburgt[1],
Sai Raghavendra Maddhuri Venkata Subramanya[2], Aashish Jain[2], Yuki Kagaya[3],
Daipayan Sarkar[1], Tunde Aderinwale[2], Xiao Wang[2], and Daisuke Kihara[1,2]
[1] – Department of Biological Sciences, [2] – Computer Science, Purdue University, West Lafayette, IN, USA
[3] – Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi, Japan,
dkihara@purdue.edu

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Our KiharaLab group participated in the six prediction categories, TS (Tertiary Structure), TR (Refinement), QA (Quality Assessment), RR (Inter-residue distance prediction), H (assembly, connected with our CAPRI human group and server group (Kiharalab_Assembly)), and Data assisted targets.

Methods

Inter-residue Distance Prediction**:** Distance maps are predicted using a deep residual network that uses features from four different Multiple Sequence Alignments (MSA) with an attention mechanism to predict the importance of each MSA based feature. Providing multiple MSA's based feature increases the co-evolutionary information provided to the network leading to better performance than since MSA based features. We use four MSA's with different e-values cutoffs of 0.001, 0.1, 1, and 10. The input features are first fed into the feature encoding layers consisting of few ResNet blocks. Next, soft attention is applied to the encoded features from all the MSA's. Finally, the attended features are passed through a deep ResNet. The MSA's were generated using DeepMSA[1] pipeline. Input features include one hot encoding of amino acid type, PSI-BLAST position specific scoring matrix, HMM profile, secondary structure and solvent accessible surface area predicted by SPOT-1D[2], CCMPRED[3], mutual information, and statistical pairwise content potential.

Along with distance, the model also predicts the backbone phi-psi angles and the orientation angles. For training, the orientation ω, θ and φ angles were computed as mentioned in trRosetta[4]. Additionally, we trained two separate ResNet models to predict the sidechain center (SCE) distance and H-bond of backbone atoms. The SCE distance represents the distance between the center of the sidechain for a pair of residues in a protein. The H-bond prediction is the distance between the N atom of residue a and O atom of residue b where the N and O atoms form a hydrogen bond.

Tertiary Structure Prediction**:** For protein structure prediction, we used Rosetta's protein folding and energy minimization protocol with customized constraints. The constraints were computed from our prediction of distance distributions (Cβ-Cβ, SCE-SCE and backbone N-O) and angle distributions (backbone-phi, psi, inter-residue orientations) by normalizing with predicted reference distributions. We generated 2,000-4,000 decoy models with different folding paths and parameters. All decoy models were ranked by the sum of the ranks of multiple scoring functions.

To explore better conformation from our models, top ranked models were further refined by Rosetta large-scale energy optimization protocol[5]. For oligomeric targets, we searched for oligomer templates by HHsearch[6]. If appropriate template structures were not found, we used our protein-protein docking protocols, LzerD[7] and Multi-LzerD[8], with our model or top-ranked server models. In some oligomer targets, we performed literature searches. We manually selected the top 5 models based on our scoring and literature search.

***Refinement*:** For refinement targets, we used our MD-based refinement protocol developed during the past CASP rounds[9]. We performed sixty 1.2 nanosecond MD simulations with Cα atom position restraints of increasing strength. The protocol uses an implicit solvent model, FACTS, with the CHARMM force field, and a dialectic constant of 2.0. For the first eight targets, our submissions were the same as Kiharalab_Refine partially because the MD-based protocol was not ready.

***Quality Assessment*:** We used our QA method that combined a new single-model QA method PRESCO2 and a machine learning method. PRESCO2 searches similar residue environments observed in a query model in a reference database of representative native protein structures. The search results are subject to final quality prediction using machine learning method that was trained to distinguish near-native structures from other decoy structures. For the training datasets, we used the same dataset as our distance prediction.

***Protein Docking*:** We submitted protein docking models through CAPRI. Top 5 models of our CAPRI human and server submissions were automatically passed to CASP submissions of Kiharalab and Kiharalab_Assembly groups, respectively. In principle, we followed our protocol reported for earlier rounds of CAPRI[10, 11]. As described for TS above, we used template-based modeling and de novo docking with our LzerD suite. Decoys were ranked by the sum of the ranks of multiple scoring functions.

***Data Assisted Targets*:** In a SAXS data assisted target (S1063), we used traditional molecular dynamics flexible fitting (MDFF)[12] method to refine the protein structure. During MDFF, the force applied on each atom is the gradient of the potential energy function derived from the density map. We performed independent MDFF simulations in implicit (using GBIS model) and explicit solvents.

For NMR data assisted targets (N1077 and N1088), our MD based method applies restraints to selected group of atoms (protons) based on the NOE-measured distances. This is implemented using *distanceInv* colvar[13] in NAMD[14]. First, we minimize the protein structure and then apply regular harmonic restraints to protein backbone by gradually lowering the force constant while equilibrating the structure. Next, harmonic wall restraints based on NOE-measured distances are applied during the MD simulations, with an upper-bound set at 5Å. We also applied Rosetta relaxation protocol to our TS models with ambiguous NMR contact data and dihedral angle data.

Acknowledgement

We are grateful for ITaP Research Computing at Purdue University for providing us the additional computational resource for this project. This work is partially supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, DBI2003635).

1. Zhang, C., W. Zheng, S. M. Mortuza, Y. Li, and Y. Zhang. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics, 36, 2105-12.
2. Hanson, J., K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics, 35, 2403-10.
3. Seemayer, S., M. Gruber, and J. Soding. (2014). CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics, 30, 3128-30.
4. Yang, J., I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci, 117,1496-503.
5. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F., & Baker, D. (2018). Protein homology model refinement by large-scale energy optimization. Proc Natl Acad Sci, 115, 3054-3059.
6. Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9, 173-175.
7. Venkatraman, V., Yang, Y. D., Sael, L., & Kihara, D. (2009). Protein-protein docking using region-based 3D Zernike descriptors. BMC bioinformatics, 10, 407.
8. Esquivel ‐ Rodríguez, J., Yang, Y. D., & Kihara, D. (2012). Multi ‐ LZerD: Multiple protein docking for asymmetric complexes. Proteins, 80, 1818-1833.
9. Terashi, G. & Kihara, D. (2018). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. Proteins: Structure, Function, and Bioinformatics, 86 Suppl 1, 189- 201.
10. Lensink, M.F., Brysbaert, G., Nadzirin, N., Velankar, S., Chaleil, R.A.G., Gerguri, T., Bates, P.A., Laine, E., Carbone, A., Grudinin, S., Kong, R., Liu, R.R., Xu, X.M., Shi, H., Chang, S., Eisenstein, M., Karczynska, A., Czaplewski, C., Lubecka, E., Lipska, A. et al. (2019). Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. Proteins 87, 1200-1221.
11. Christoffer, C., Terashi, G., Shin, W. H., Aderinwale, T., Maddhuri Venkata Subramaniya, S. R., Peterson, L., Verburgt, J., & Kihara, D. (2020). Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38 ‐ 46. Proteins, 88, 948-961.
12. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. (2008) Structure 16, 673–683.
13. Fiorin, G., Klein, M. L. & Hénin, J. Using collective variables to drive molecular dynamics simulations. (2013) Mol. Phys. 111, 3345–3362.
14. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. (2020) J. Chem. Phys. 153, 044130.

Kiharalab_Contact

Protein Distance and Contact Prediction with Deep Learning

Aashish Jain[1], Sai Raghavendra Maddhuri Venkata Subramanya[1], Genki Terashi[2], Yuki Kagaya[3] and Daisuke Kihara[1,2]

[1] - Department of Computer Science, Purdue University, West Lafayette, IN, USA, [2] - Department of Biological Sciences, Purdue University, West Lafayette, IN, USA, [3] - Graduate School of Information Sciences, Tohoku University, Japan

dkihara@purdue.edu

***Key:** Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N*

We submitted models to distance/contact prediction category (RR) in CASP14. Our method is composed of two deep learning networks. The first network is a ResNet model that predicts distance distribution based on sequences features from four different Multiple Sequence Alignments (MSA). Distance prediction is converted to contact prediction and passed to the second network. The second network is a GAN model that refines the predicted noisy contacts and generates an improved contact map.

Methods

The first model is a deep residual network that predicts the protein distance along with backbone and orientation angles from multiple different MSA. Deep learning based distance prediction relies heavily on the input MSA as it contains information about evolutionary conserved positions and motifs. To provide more information to the model we use features from 4 MSA with e-value cutoffs of 0.001, 0.1, 1, and 10. We add an attention layer over the different MSA encoded features to let the model choose for every pair of residue which MSA to focus on. The input features are first fed into the feature encoding layers consisting of few ResNet blocks. Next, soft attention is applied to the encoded features from all the MSA's. Finally, the attended features are passed through a deep ResNet.

We used 8 sequence-based input features. The MSA was generated using DeepMSA[1] pipeline. The 1D features include one hot encoding of amino acid type, PSI-BLAST position specific scoring matrix, HMM profile, and secondary structure and solvent accessible surface area predicted by SPOT-1D. The 2D pairwise features includes CCMPRED, mutual information, and statistical pairwise content potential.

The predictions from the above network are then passed to ContactGAN[2] model. ContactGAN is a novel contact map denoising and refinement method using Generative Adversarial Networks (GAN)[3]. ContactGAN takes a contact map predicted by existing methods, which is considered as an imperfect, noisy input, and outputs an improved map that better captures correct residue-residue contacts compared to the original map. ContactGAN was trained with predicted noisy contact maps coupled with corresponding native contact maps, which the networks were guided to generate. Figure 1 outlines the architecture of ContactGAN.

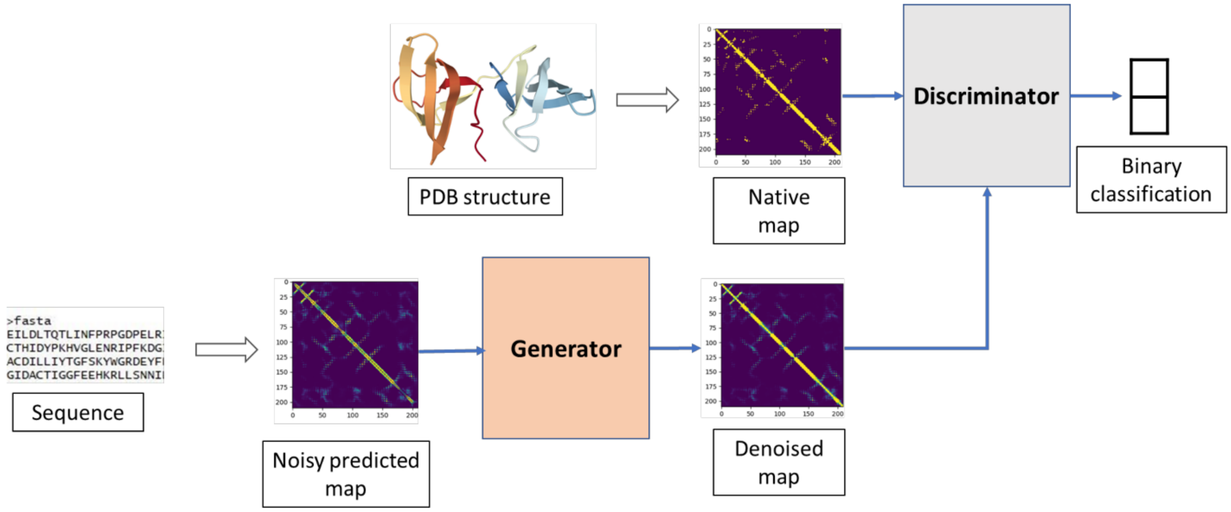


Figure 1. The architecture of ContactGAN showing overall structure that connects the generator and the discriminator networks. The generator network takes a noisy predicted contact map and outputs a refined map. The discriminator network is to discriminate a generated map by the generator network and the native map, so that the generator is trained to produce indistinguishable maps from native maps by the discriminator.

Acknowledgements
We are grateful for ITaP Research Computing at Purdue University for providing us the additional computational resource for this project. This work is partially supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, DBI2003635).

1. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 36, 2105-2112, doi:10.1093/bioinformatics/btz863 (2020).
2. Subramaniya, S.R.M.V., Terashi, G., Jain, A., Kagaya, Y. & Kihara, D., (2020). Protein Contact Map Denoising Using Generative Adversarial Networks. bioRxiv. doi: 10.1101/2020.06.26.174300
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y., (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672-2680).

Kiharalab_Refine

Integrative modeling for protein structure refinement using Molecular Dynamics with flat-bottom harmonic restraints, enhanced sampling and ROSETTA iterative hybridize

Daipayan Sarkar[1,#], Jacob Verburgt[1,#], Charles Christoffer[2], Yuki Kagaya[3], Genki Terashi[1], Karl Lundquist[4], Xiao Zhu[5], Lev Gorenstein[5], Daisuke Kihara[1, 2, c]

[1] - Department of Biological Sciences, Purdue University, West Lafayette, IN, USA, [2] - Department of Computer Science, Purdue University, West Lafayette, IN, USA, [3] - Graduate School of Information Sciences, Tohoku University, Japan,[4] - Markey Center for Structural Biology, Department of Biological Sciences, Purdue University, West Lafayette, Indiana, 47907,[5] - Research Computing, Purdue University, West Lafayette, IN, USA

[#]D.S. and J.V. contributed equally (co-first author)

[c]dkihara@purdue.edu

Key: *Protein structure refinement, Molecular Dynamics, Flat-bottom harmonic restraints, Enhanced Sampling, Metadynamics, NAMD, VMD, ROSETTA.*

Proteins are important biomolecules that are responsible for different biological function. Over the years, several experimental methods like X-ray crystallography (X-ray), Nuclear Magnetic Resonance (NMR), Small Angle X-ray Scattering (SAXS) and cryo-electron microscopy (cryo-EM) have been developed to determine protein structure. With advancement in technology, it is becoming very common now to solve high resolution protein structure and complex. However, the sparse nature of the experimental data, often requires deep learning[1] and physics-based using molecular dynamics (MD)[2–4] computational approach, to model and refine protein structures. Here, our group used a physics-based approach using MD to refine protein structures released during the ongoing CASP-14 season.

Methods

In this section, we outline the MD and enhanced sampling procedure adopted by our protein structure refinement group for the regular and extended refinement targets in CASP-14.

MD using flat-bottom harmonic restraints: Based on the success reported by Heo et al[5] to refine protein structures near experimental accuracy, we adopt a similar approach for this CASP experiment. Our method uses both harmonic restraints and flat-bottom harmonic restraints to refine protein structures. This combined approach showed refinement of several protein structures during our benchmark of the CASP-13 refinement targets. Specifically, in our MD based refinement approach, we initially minimize the protein structure, subsequently applying regular harmonic restraints to protein backbone by gradually lowering the force constant over a simulation time of 10 nanoseconds (ns). So far, the motivation here is to initially have a conservative approach of refinement from the starting structure. Next, we switch to flat-bottom harmonic restraints with a 4 Å width, applied to the C-α atoms in protein structure and perform MD simulation for an additional 100 ns. The total simulation time of a single MD trajectory is 110 ns. We perform 3 such cycles for a total time of 330 ns. The flat-bottom harmonic restraints are implemented using the Colvars[6] module in NAMD[7] and VMD[8]. All-atom MD simulations are performed using NAMD 2.13 with CHARMM36m[9] force-field in explicit solvent. Additionally, we use the hydrogen mass repartitioning[10] method to be able to use a longer timestep of 4 femtoseconds (fs) with flat-bottom harmonic restraints, instead of usual 1-2 fs.

Extended refinement: During CASP-14, for some refinement targets (8 out of 51) the organizers requested running longer MD simulations of earlier refined models to see if additional refinement is possible by further sampling the protein conformational space. The choice of the extended refinement targets provided by organizers included initial models with very high (R1034) and low (R1029) GDT_HA values, 70 and 28 respectively. For such extended refinement, we performed longer MD simulations, continuing for another 100 ns per trajectory using flat-bottom harmonic restraints. The total time for a refinement target under this category is 0.62 μs. As a parallel approach to our flat-bottom harmonic restraint MD, we implemented MD with enhanced sampling, specifically using metadynamics[11]. Here, we start from a refined model using flat-bottom harmonic restraints and perform enhanced sampling using metadynamics with a harmonic wall biased potential as implement in NAMD colvars. A two-dimensional potential of mean force (PMF) with appropriate thermodynamic weighting is constructed using root mean square deviation (RMSD) and radius of gyration (R_g) as our choice of the collective variables (or reaction coordinate). Finally, clustering is performed to identify ensemble of protein structures that are occupying low energy states.

Rosetta-Iterative Hybridize: Several (7 out of 51) of the extended refinement targets were also refined in our implementation of the Rosetta Iterative Hybridize[12] protein refinement protocol. The protocol first creates a diverse set of 50 models which are derived from the input structure by identifying and removing the flexible regions via MD and rebuilding them with Rosetta Comparative Modeling. These 50 models are then subjected to 50 iterations of the genetic algorithm-based refinement protocol, of which the top 5 scoring models (via the Rosetta Energy Function[13]) are forwarded to the pool for manual selection. Each iteration of the algorithm consists of 10 "parent structures" or seeds in which fragments of these structures undergo mutations from a fragment library or crossovers with other structures. Using the Rosetta Energy Function as selection criteria, 50 models are selected for the next iteration.

Results

In this section, we discuss how final models were selected based on the above refinement methodology using different MD and enhanced sampling methods.

MD using flat-bottom harmonic restraints: First, we calculated the DFIRE[14] energy potential and GDT-HA[15] from the starting model across 3 independent MD simulations per target. Based on our success in CASP-12, we use the same clustering[16] algorithm to structurally average protein structure. The resultant structure from their representative cluster was minimized for 1000 steps to refine the protein sidechains. Additionally, we ranked the models based on DFIRE energy function. Finally, from the five selected models we rank the models based on their MolProbity[17] score after visual inspection using PyMOL.

Extended refinement: Similar approach was adopted here as the MD using flat-bottom harmonic restraint method. Additionally, for our enhanced sampling MD simulations using metadynamics, each trajectory was clustered into five clusters representing the low energy states. The RMSD cutoff for clustering was set in the range 5 – 12 Å as necessary to obtain five clusters for any given target. Based on this, 15 models were selected, 5 each from regular flat-bottom harmonic restraint MD, metadynamics and ROSETTA iterative hybridize methods. Later, we ranked all 15 models by their MolProbity scores and selected the top five models after visual inspection using PyMOL.

Acknowledgement

This work is partially supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, DBI2003635). We also acknowledge the computational nodes provided to us by Research Computing at Purdue University, West Lafayette, IN, USA.

Availability

All data for input files for MD simulation (flat-bottom harmonic restraint and metadynamics) and final submitted models are available. Additional information is also available upon request.

1. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature (2020) doi:10.1038/s41586-019-1923-7.
2. Singharoy, A. et al. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. Elife 5, (2016).
3. Feig, M. Computational protein structure refinement: almost there, yet still so far to go. Wiley Interdiscip. Rev. Comput. Mol. Sci. 7, e1307 (2017).
4. Heo, L. & Feig, M. High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins Struct. Funct. Bioinforma. 88, 637–642 (2020).
5. Heo, L., Arbour, C. F. & Feig, M. Driven to near-experimental accuracy by refinement via molecular dynamics simulations. Proteins Struct. Funct. Bioinforma. 87, 1263–1275 (2019).
6. Fiorin, G., Klein, M. L. & Hénin, J. Using collective variables to drive molecular dynamics simulations. Mol. Phys. 111, 3345–3362 (2013).
7. Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153, 044130 (2020).
8. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. J. Mol. Graph. 14, 33–38 (1996).
9. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat. Methods 14, 71–73 (2017).
10. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. J. Chem. Theory Comput. 11, 1864–1874 (2015).
11. Laio, A. & Parrinello, M. Escaping free-energy minima. Proc. Natl. Acad. Sci. 99, 12562–12566 (2002).
12. Park, H. et al. High-accuracy refinement using Rosetta in CASP13. Proteins Struct. Funct. Bioinforma. 87, 1276–1282 (2019).
13. Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 13, 3031–3048 (2017).
14. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 11, 2714–2726 (2009).
15. Zhang, Y. . S. J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005).
16. Terashi, G. & Kihara, D. Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. Proteins Struct. Funct. Bioinforma. 86, 189–201 (2018).
17. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. 27, 293–315 (2018).

Kiharalab_Z_Server

Automatic Prediction of Protein Structure by Deep Learning and Rank Aggregation

Charles Christoffer[1], Aashish Jain[1], Sai Raghavendra Maddhuri Venkata Subramaniya[1], Genki Terashi[2], Yuki Kagaya[3], and Daisuke Kihara[1,2]

[1] – Department of Computer Science, Purdue University, West Lafayette, IN, USA, [2] – Department of Biological Sciences, Purdue University, West Lafayette, IN, USA,[3] – Graduate School of Information Sciences, Tohoku University, Japan

dkihara@purdue.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

We used a fully automatic protein structure prediction pipeline to submit to the TS and RR categories as a server predictor in CASP14. Our server predicts protein structures entirely without any templates or fragment libraries, relying on multiple sequence alignment (MSA) features and deep learning to predict a residue-residue distance distribution. After structures are extracted from the distance distributions using an energy minimization procedure, they are ranked by multiple scoring functions. The rankings are then aggregated to select the five models for submission.

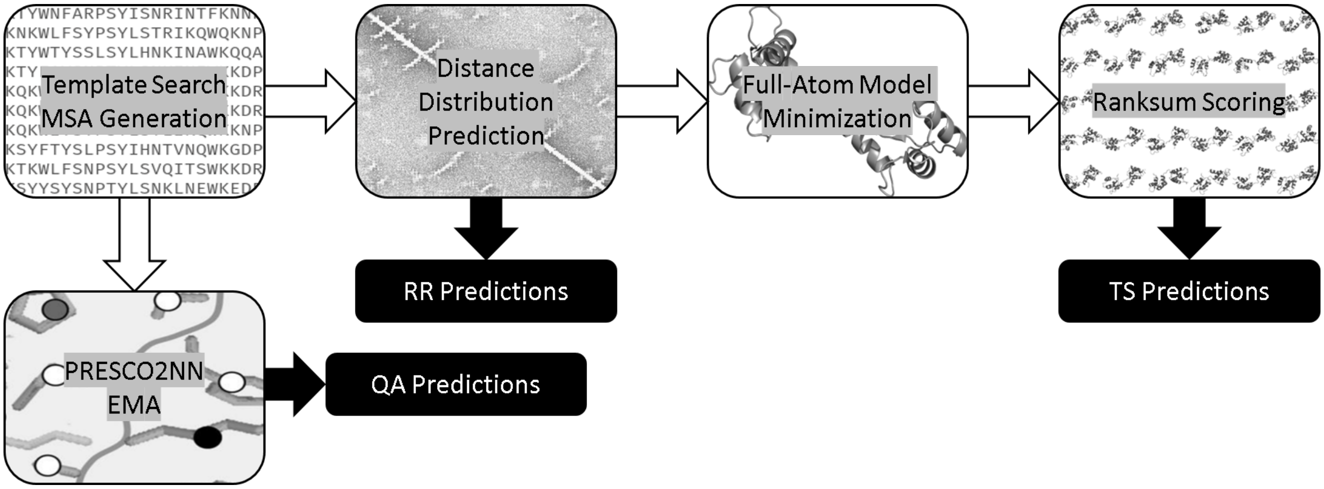
Methods

The first stage of our pipeline generates four MSAs with e-value cutoffs of 0.001, 0.1, 1, and 10 using the DeepMSA[1] pipeline, which uses HH-suite and HMMER programs to generate alignments from the UniClust30[2], UniRef90[3], and Metaclust[4] databases. Eight sequence-based input features were fed into the neural network stage to generate a predicted residue-residue distance distribution. The 1D features include one-hot encoding of amino acid type, PSI-BLAST[5] position specific scoring matrix, HMM profile, and secondary structure and solvent accessible surface area predicted by SPOT-1D[6]. The 2D pairwise features include CCMPRED[7], mutual information, and statistical pairwise content potential. The distance distributions from this stage were submitted for the RR category.

Once generated, the predicted distance distribution is converted into full-atom structure models by L-BFGS minimization of predicted short-, medium-, and long- range distance restraints in sequence using PyRosetta[8]. Rankings of the model pool are then calculated using the knowledge-based scoring functions GOAP[9], DFIRE[10], and ITScorePro[11], as well as Rosetta's REF2015[12] score. These rankings are then aggregated using the ranksum method[13-15], where the ranks of a given model by each of the component scores are added to produce a new ordering. The top five models by ranksum were submitted for the TS category.

For the QA category, we combined template search with a single-model QA method PRESCO2 and machine learning. PRESCO2 searches similar residue environments observed in a query model in a reference database of representative native protein structures. The PRESCO2 search results are subjected to final quality prediction using machine learning method that was trained to distinguish near-native structures from other decoy structures. For targets where no template was detected, the output of this neural network was used to rank the models. For targets with a detected template, the models were instead ranked by TM-score from TM-align[16] to the template.

Figure 1



Flowchart of our automated prediction pipeline. All structure predictions are derived from information from sequence databases, processed by machine learning models and restraint minimization. QA predictions refer to a template when available.

Acknowledgements
We are grateful for ITaP Research Computing at Purdue University for providing us additional computational resources for this project. This work is partially supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941, MCB1925643, DBI2003635). C.C. was supported by a National Institute of General Medical Sciences-funded predoctoral fellowship to C.C. (T32 GM132024).

1. Zhang,C., Zheng,W., Mortuza,S.M., Li,Y., Zhang,Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics. 2020;36:2105-12.
2. Mirdita,M., von den Driesch,L., Galiez,C., Martin,M.J., Soding,J., Steinegger,M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45:D170-D6.
3. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H., UniProt, C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926-32.
4. Steinegger,M., Soding,J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9:2542.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-402.
6. Hanson,J., Paliwal,K., Litfin,T., Yang,Y., Zhou,Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted

contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics. 2019;35:2403-10.

7. Seemayer,S., Gruber,M., Soding,J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics. 2014;30:3128-30.
8. Chaudhury,S., Lyskov,S., Gray,J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics. 2010;26:689-91.
9. Zhou,H., Skolnick,J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J. 2011;101:2043-52.
10. Zhou,H., Zhou,Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002;11:2714-26.
11. Huang,S.Y., Zou,X. ITScorePro: an efficient scoring program for evaluating the energy scores of protein structures for structure prediction. Methods Mol Biol. 2014;1137:71-81.
12. Park,H., Bradley,P., Greisen,P.,Jr., Liu,Y., Mulligan,V.K., Kim,D.E., et al. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016;12:6201-12.
13. Christoffer,C., Terashi,G., Shin,W.H., Aderinwale,T., Maddhuri Venkata Subramaniya,S.R., Peterson,L., et al. Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38-46. Proteins. 2019.
14. Peterson,L.X., Shin,W.H., Kim,H., Kihara,D. Improved performance in CAPRI round 37 using LZerD docking and template-based modeling with combined scoring functions. Proteins. 2018;86 Suppl 1:311-20.
15. Peterson,L.X., Kim,H., Esquivel-Rodriguez,J., Roy,A., Han,X., Shin,W.H., et al. Human and server docking prediction for CAPRI round 30-35 using LZerD with combined scoring functions. Proteins. 2017;85:513-27.
16. Zhang,Y., Skolnick,J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302-9.

Kozakov-Vajda

Template-assisted docking and docking of protein models.

Dima Kozakov[1], Sandor Vajda[2,3], Kathryn Porter[2], Dzmitry Padhorny[1], Israel Desta[2], Dmitri Beglov[2], Mikhail Ignatov[1], Sergey Kotelnikov[1]

[1]Laufer Center for Physical and Quantitative Biology, Stony Brook University; Departments of [2]Biomedical Engineering and [3]Chemistry, Boston University

In spite of the significant progress in scoring functions, template-based modeling of protein assemblies usually outperforms free docking when good templates are available. Thus, as a human predictor group we focused on the enhancement of the template-based modeling methodology and explored our ability to improve the quality of assembly modeling by combining the predictions of our ClusPro automated server with the models of the individual subunits generated by CASP participants.

On the other hand, for many complexes one does not have assembly templates, and sometimes even subunit templates are not available. In such cases, a protocol that unifies modeling of protein subunits with free docking is required. Thus, our second focus was application of protein docking methodology to the docking of template-based and *ab initio* models of protein subunits.

Methods

Model Preparation: For model preparation we either use the top template provided by HHPRED or, in difficult cases, build an “ab initio” model of the subunit using TrRosetta. For each “easy” target most models had the same fold, with variations in loops and tails. Removal of the uncertain regions resulted in reliable “consensus” models that were used for docking.

Template based docking. If a template of the biological complex satisfying the requird stoichiometry is found then we chose the best template for each unique monomer of the complex, align multiple copies of this monomer template to the complex template and then model the whole complex using Modeller. In order to diversify the pool of models, we also used the CASP server models of the subunits by aligning those onto our initial template based models and minimizing the resulting structures. Per rules of CAPRI we generate up to 10 models.

Free Docking: Our free docking approach consists of two steps. The first step is running PIPER, a docking program that performs systematic search of complex conformations on a grid using the fast Fourier transform (FFT) correlation approach. The scoring function includes van der Waals interaction energy, an electrostatic energy term, and desolvation contributions calculated by a structure based pairwise potential. We can effectively account for cross-linking and SAXS data in the global search as described above.

The second step of the algorithm is clustering the top 1000 structures generated by PIPER using pairwise RMSD as the distance measure. The radius used in clustering is defined in terms of C_α interface RMSD. For each docked conformation we select the residues of the ligand that have any atom within 10 Å of any receptor atom, and calculate the C_α RMSD for these residues from the same residues in all other 999 ligands. Thus, clustering 1000 docked conformations involves computing a 1000 × 1000 matrix of pairwise C_α RMSD values. Based on the number of structures that a ligand has within a (default) cluster radius of 9 Å RMSD, we select the largest cluster and rank its cluster center as number one. The members of this cluster are removed from the matrix, and we select the next largest cluster and rank its center as number two, and so on. After clustering with this hierarchical approach, the ranked complexes are subjected to a straightforward (300 step and fixed backbone) minimization of the van der Waals energy using the CHARMM potential to remove potential side chain clashes.

LamoreuxLab

Modeling and assessment of CASP14 Targets using 3DCNN and template-based methods

Devlina Chakravarty[1], Talant Ruzmetov[1], Georgy Derevyanko[2], and Guillaume Lamoureux[1]

1- Department of Chemistry and Center for Computational and Integrative Biology, Rutgers University, Camden, New Jersey, USA, 2Department of Chemistry and Biochemistry and Centre for Research in Molecular Modeling (CERMM), Concordia University

Goal of the CASP experiment is to advance the state of the art in modeling protein structure from amino acid sequence to protein structure and assembly. Our team has participated in two sections of the competition; quality assessment of protein models (QA) and prediction of protein complexes or tertiary assemblies (TS). Overall, we participated in 78 QA and 28 TS prediction rounds. For QA we have relied on scores generated by our pre-trained 3DCNN model,1 adopting the protocols used in CASP13. Individual monomers were either directly modelled or obtained by ranking the predictions from CASP server stage 2 rounds. Protein assemblies were built in two main ways: (1) using a template-free approach based on the CNN docking algorithm2 recently developed in our lab and (2) using template-based docking3 whenever templates were available.

The CNN docking algorithm was trained to predict complex three-dimensional representations from the atomic densities of a protein. Each protein is represented by 48 features, which are combined with those of the partner protein to produce a score that can be optimized by adjusting the relative position and orientation of the two proteins. Complexes of more than two proteins are assembled in a stepwise manner, one protein at a time. The CNN docking algorithm is trained on decoy conformations generated from the dataset of Huang and Zou4 and is tested on structures from the Protein-Protein Docking Benchmark Version 4.0.5

When good templates were found, docking poses were generated by aligning the proteins with TM-align6 and were scored by global score from 3DCNN-LQA algorithm. During the TS rounds, most of the protein assemblies were built using a combination of the CNN docking algorithm and template-based methods, and the complexes built were ranked using the 3DCNN-LQA algorithm. Our neural network-based approach enabled us to deal with assemblies of any size and stoichiometry. By interpreting the CNN docking score as an energy, we could also estimate the binding affinity of some of the complexes.

1. Derevyanko, G., Grudinin, S., Bengio, Y., & Lamoureux, G. (2018). Deep convolutional networks for quality assessment of protein folds. Bioinformatics, 34(23), 4046-4053.
2. Derevyanko, G., & Lamoureux, G. "Protein-protein docking using learned three-dimensional representations." bioRxiv (2019): 738690.
3. Chakravarty, D., McElfresh, G. W., Kundrotas, P. J., & Vakser, I. A. (2020). How to choose templates for modeling of protein complexes: Insights from benchmarking template-based docking. Proteins: Structure, Function, and Bioinformatics.
4. Huang, S.-Y. and Zou, X. An iterative knowledge-based scoring function for protein–protein recognition. Proteins: Structure, Function, and Bioinformatics, 72(2):557–579, 2008.
5. Hwang, H., Vreven, T., Janin, J., & Weng, Z. (2010). Protein–protein docking benchmark version 4.0. Proteins: Structure, Function, and Bioinformatics, 78(15), 3111-3114.
6. Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research, 33(7), 2302-2309.

Laufer_ros

Protein Folding with MELD Molecular Dynamics

R. Nassar[1], C. Liu[1], E. Brini[1], S. Parui[1], G. Dignon[1] and K.A. Dill[1,2,3]

1– Laufer Center for Physical and Quantitative Biology, Stony Brook University, 2– Department of Physics, Stony Brook University, 3– Department of Chemistry, Stony Brook University

dill@laufercenter.org

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

Molecular dynamics (MD) simulations possess the ability to sample structures with atomic-scale resolution. A major bottleneck, however, is the vast configurational space that a protein can sample when folding. To alleviate this burden, we use our MD accelerator MELD, which leverages external information to limit the search space of physics based folding simulations[1,2]. MELD accelerated MD (MELD x MD) complements knowledge-based, machine learning, and experimental approaches by integrating data from all of them into a technique that delivers high resolution structures with free energy based scoring[3]. In CASP14, we combine MELD with structures and contact information from trRosetta[4], secondary structure predictions from PSIPRED[5], and general protein properties[2] to guide the generation of 3D conformations in the free modeling as well as the refinement categories. All generated structures from MELD x MD are clustered into conformational macrostates. In the limit of a converged simulation, the most populated macrostate is the most stable one, i.e. it is the native state. The centroids of the five most populated clusters are submitted to CASP14 as representatives of each macrostate.

Methods

MELD (Modeling Employing Limited Data) is a Bayesian sub-haystacking method that uses external knowledge and even noisy data to speed up protein folding in a Replica Exchange MD (REMD[6]) construct. Various types of data can be used once converted to distance and angle restraints to guide the MD towards low energy structures that are compatible with the information[7]. To allow for noise and ambiguity in the data, the restraints are not all activated at the same time during the simulation (see below for details).

MELD is a freely available plugin to the OpenMM[8] simulation software. In our CASP14 protocol, we set up the simulations to use the AMBER ff14SBonlysc[9] force field and gbNeck2 implicit solvent[10]. By the start of CASP14, the machine learning algorithm trRosetta was the most accurate for inferring contacts and generating structures from them[4]. We integrate predictions from trRosetta into our MELD pipeline as follows: the trRosetta server predictions submitted to CASP are downloaded, built using tleap and minimized before inputting into MELD x MD as starting points. We obtain trRosetta's predicted contacts from CASP and filter out low probability contacts ($p<0.5$) and local contacts ($CO<6$). We then input the remaining ones into MELD as separate short-range ($6<CO<12$), medium-range ($11<CO<24$) and long-range ($CO>23$) contacts. We enforce these restraints on the distances between Cβ atoms (Cα for glycine residues) of each pair of residues using a flat bottom energy term added to the force field. This energy term has a favorable zero value region when $2 < d_{C\beta\text{-}C\beta} < 9$ Å, a quadratic region when $9 < d_{C\beta\text{-}C\beta} < 10$ Å and linear when $d_{C\beta\text{-}C\beta} > 10$ Å. We set MELD to enforce 80% of each short, medium and long range restraint lists to allow for inaccurately predicted contacts.

Secondary structure information is constructed using predicted helix and strand residues from PSIPRED[5]. Helical distance and angle restraints, as well as pairwise distance restraints between strand residues, are incorporated into MELD at 70% enforcement to allow for inaccurate assignments. Here, as well as for all other restraint types input into MELD, the force field determines which 70% fraction of all possible secondary structure restraints are favorable to keep. Hydrophobic restraints are also enforced in MELD to create protein hydrophobic cores[2]. This is done using pairwise distance restraints ($d_{\mathrm{ij}} < 5$Å) between any heavy atoms of two hydrophobic residues. We enforce a total of 0.8 x number of hydrophobic residues in the sequence.

All the restraint sets were incorporated into a 1D Hamiltonian and Temperature Replica Exchange Molecular Dynamics scheme (HT-REMD)[1,2]. Each target simulation consists of 32 temperature replicas (300-500 K) running on 32 GPUs. The force constant for the restraints ranges between 0 KJ/mol/nm^2 at high T to release the restraints and allow for large-scale conformational sampling, and 250 KJ/mol/nm^2 at low T to refine the structures inside the discovered energy minima. At intervals of MD timesteps, MELD ranks the restraint energies in each category (e.g predicted contacts) and enforces the ones with the lowest restraint energies to discover a data-compatible conformation at that time. This way MELD x MD examines all possible combinations of contacts and lets the force field decide which ones are favorable and which are not. Consequently, more populated conformations correspond to thermodynamically favorable structures on the MELD x MD landscape. The simulations are run for up to 2.5μs of simulation time to allow for as much sampling as possible. Structures from the lowest five temperature replicas are then collectively clustered using CPPTRAJ[11]. The centroids of the five most populated clusters are then minimized[12] and submitted to CASP14.

Refinement targets are simulated with the same protocol, with the one difference being that we use the provided template as a starting configuration rather than server predictions, and we back-calculate secondary structures from this template instead of relying on PSIPRED predictions.

Results

Few of the target structures in CASP14 have been resolved experimentally and publicly released. Comparison to the best five MELD x MD structures shows 3.4, 3.2 and 3.8 Å in backbone RMSD for the three targets T1035, T1046s1 and T1046s2 respectively.

Availability

https://github.com/maccallumlab/meld.git

1. MacCallum,J.L., Perez,A. & Dill,K. (2015). Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proc. Natl. Acad. Sci. U. S. A. 112, 6985-6990.
2. Perez,A., MacCallum,J.L. & Dill,K. (2015). Accelerating molecular simulations of proteins using Bayesian inference on weak information. Proc. Natl. Acad. Sci. U. S. A. 112, 11846-11851.
3. Perez,A., Morrone,J.A., Brini,E., MacCallum,J.L. & Dill,K. (2016). Blind protein structure prediction using accelerated free-energy simulations. Science Advances. 2, e1601274.
4. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. & Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. 117, 1496-1503.

5. Jones,D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
6. Sugita,Y. & Okamoto,Y. (1999). Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. 314, 141−151.
7. Robertson,J., Nassar,R., Liu,C., Brini,E., Dill,K. & Perez,A. (2019). NMR-assisted protein structure prediction with MELDxMD. Proteins. 87, 1333-1340.
8. Eastman,P., Swails,J., Chodera,J.D., McGibbon,R.T., Zhao,Y., Beauchamp,K.A., Wang,L.P., Simmonett,A.C., Harrigan,M.P., Stern,C.D., Wiewiora,R.P., Brooks,B.R. & Pande,V.S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comp. Biol. 13: e1005659.
9. Maier,J.A., Martinez,C., Kasavajhala,K., Wickstrom,L., Hauser,K.E. & Simmerling,C. (2015). Improving the accuracy of protein side chain and backbone parameters from ff99SB J. Chem. Theory Comput. 11, 8, 3696-3713.
10. Mongan,J., Simmerling,C., McCammon,J.A., Case,D.A. & Onufriev,A. (2007). Generalized Born model with a simple, robust molecular volume correction. J. Chem. Theory. Comput. 3, 156−169.
11. Roe,D.R. & Cheatham,T.E. (2013). PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. J. Chem. Theory. Comput. 9, 3084-3095.
12. Perez,A., Roy,A., Kasavajhala,K., Wagaman,A., Dill,K. & MacCallum,J.L. (2014). Extracting representative structures from protein conformational ensembles. Proteins. 82, 2671−2680.

LAW, MASS

Predicting protein residue-residue contacts and tertiary structures using deep networks with varying dilation rates

Tong Liu[1] and Zheng Wang[1, *]
Department of Computer Science, University of Miami
Zheng.Wang@miami.edu

We present two protein residue-residue contact and tertiary structure prediction servers named LAW and MASS. For residue-residue contact prediction, we collected 73 residue-specific 1D features and five 2D features and designed two deep convolutional networks with varying dilation rates. We used two significant factors (predicted top contacts and predicted secondary structures) as the input to CNS for modelling protein structures.

Methods

The domain sequences and corresponding native structures for residue-residue (RR) contact training are from SCOPe 2.07. We only kept sequences with length in [30, 600] and with resolution of corresponding native structures less than or equal to 2.0 Å, resulting in 6113 sequences. For each of the retaining sequences, we executed DeepMSA[1] to obtain its multiple sequence alignment (MSA). Position-Specific Scoring Matrix (PSSM) was obtained from Psiblast. We used two different normalization methods (i.e., row normalization and sigmoid function) for normalizing entries in PSSM in the range 0 to 1. PSIPRED[2] was used for predicting secondary structure (SS) and solvent accessibility (ACC). We collected 73 residue-specific 1D features in total: 40 for PSSM, 3 for SS, 1 for ACC, 5 for factor solution scores, 4 for sinusoidal positional encoding, and 20 for hot coding. We also used five 2D features: 1 for CCMpred (GPU version)[3], 1 for mfDCA[4], and 3 for mutual information and pairwise potential implemented in MetaPSICOV[5]. After generating all of these features and discarding proteins with some missing features, we finally used 6085 sequences for training and validation of RR contact prediction.

For RR contact prediction, we designed two types of dilated deep convolutional networks (i.e., LAW and MASS), both containing two main successive parts: the first 1D convolutional part is used to learning useful representations of 1D residue-specific features, and the second 2D convolutional part is for combining 1D and 2D features and for better learning the underlying relationships between 2D features and ground truth (native RR contact maps), similar to the network described in [6]. For LAW network, the first part contains six 1D residual blocks; each block consists of two 1D convolutional layers with kernel size set to 15, and the second part contains 18 2D residual blocks; each block consists of two 2D convolutional layers with kernel size set to 5 and different dilation rates. For MASS network, the first part contains 18 1D residual blocks; each block consists of two 1D convolutional layers with kernel size set to 5 and different dilation rates, and the second part is similar to the corresponding one in LAW network. Both networks were trained using dynamic batch sizes and Adam optimizer with initial learning rate set to 0.001. The loss function we used is binary cross entropy with positive weight set to 3.

For structure prediction, we used top 2.3×L (L is the protein length) predicted contacts and predicted secondary structures from PSIPRED as the input to CONFOLD[7], which is a two-stage protein structure prediction method based on the Crystallography & NMR System (CNS).

Results

We evaluated our two contact prediction servers based on the targets of CASP13 experiment. The evaluation results are shown in the following table, indicating that LAW is better in terms of TBM-easy targets, whereas MASS performs slightly better on FM targets.

	Short			Medium			Long		
	L	L/2	L/5	L	L/2	L/5	L	L/2	L/5
	FM targets								
LAW	0.266	0.421	0.586	0.298	0.448	0.569	0.335	0.412	0.495
MASS	0.26	0.401	0.599	0.321	0.447	0.569	0.36	0.465	0.567
	TBM-easy targets								
LAW	0.249	0.427	0.701	0.345	0.561	0.763	0.705	0.861	0.941
MASS	0.242	0.421	0.711	0.341	0.554	0.743	0.695	0.832	0.899

1. Zhang, C., Zheng, W., Mortuza, S., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics (2019).
2. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292, 195-202 (1999).
3. Seemayer, S., Gruber, M. & Sӧding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30, 3128-3130 (2014).
4. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108, E1293-E1301 (2011).

5.. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31, 999-1006 (2014).

6. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology 13, e1005324 (2017).
7. Adhikari, B., Bhattacharya, D., Cao, R. & Cheng, J. CONFOLD: residue ‐ residue contact ‐ guided ab initio protein folding. Proteins: Structure, Function, and Bioinformatics 83, 1436-1449 (2015).

Protein Single-Model Accuracy Estimation Using Graph and Residual Neural Networks

Chenguang Zhao[1], Tong Liu[1], Zheng Wang[1, *]
Department of Computer Science, University of Miami
Zheng.Wang@miami.edu

We developed two computational methods, named LAW and MASS, for in the category of Estimation of Model Accuracy (EMA). LAW was implemented as a combination of graph and convolutional neural networks, whereas MASS was built as a residual neural network (Resnet). Both servers were new designs compared to our previous servers[1,2].

Methods

We used the targets of CASP7 to CASP12, at the level of evaluation units (EUs)[3], to train the networks. For each residue of a protein model, we generated 83 features, which can be classified into six categories: (1) hot coding of amino acid sequence; (2) position-specific scoring matrix (PSSM) created using PSI-BLAST from the multiple sequence alignment (MSA); (3) normalized Rosetta energies; (4) SOV_refine scores[4] indicating the similarity between the sequence-based and model-based secondary structures (SS) and solvent accessibilities; (5) sinusoidal positional encoding; and (6) MASS protein statistical potentials[1] including pseudo-bond angle potential (PAP), accessible surface potential at the atomic level (ASPA), sequence separation-dependent potential (SSDP), contact-dependent potential (CDP), relative solvent accessibility potential (RSAP), and volume-dependent potential (VDP).

LAW used a 5-layer graph network[5] as the basis component for both local and global quality assessments. We defined each individual protein model as a graph and generated node, edge and global features. The node features are residue-specific as described above. The global features are the global SOV_refine consistency scores for SS and solvent accessibility. The edge features are related to Euclidean distances, sequence separation, and angle between two corresponding residues. The LAW-local network for residue-specific deviation prediction contains a 5-layer graph network followed by a 3-layer 1D convolutional network. The LAW-global network for GDT-TS prediction is a 5-layer graph network followed by a fully connected layer. Both LAW-local and LAW-global used dropout for reducing overfitting.

The MASS network for predicting local deviations contains 24 blocks. Each block has two 1DConv-BatchNorm-Dropout-LeakyReLU layers. The MASS network begins with two branches; the input of the first branch, consisting of six blocks, contains the features of the current residue, whereas the input of the other branch contains the concatenated features of the top five residues that are most spatially proximate to the current residue. The two branches are then combined and followed by another 18 blocks along with a fully connected layer converting channels into one as the predicted residue-specific deviations. The MASS network for predicting global accuracies also starts with the same two branches as in MASS-local but is followed by a combination of a fully connected layer and a spatial pyramid pool layer[6], the latter one converting varying lengths of different models into a fixed-length scalar as the global GDT-TS prediction for the input model.

Results

We used 79 targets in stage 2 in CASP13 experiment for evaluating the two methods[3]. Performances of LAW and MASS in terms of local and global accuracy estimation are shown in Table 1.

For evaluating local accuracy estimation, we followed the evaluation criteria as in CASP13[3] including the correlation between the real and predicted deviations, residue-wise S-score error (ASE), area under the curve (AUC), and unreliable local region (ULR). For evaluating global accuracy estimation, top 1 GDT-TS loss and top 1 LDDT loss were calculated.

Score Server	Correlation	ASE/100	AUC	ULR	Top 1 loss GDT- TS	Top 1 loss LDDT
LAW	0.552	0.833	0.839	0.246	13.322	8.184
MASS	0.572	0.832	0.852	0.268	13.965	9.735

Table 1. Evaluation results of local and global accuracy estimations.

1. Liu, T. & Wang, Z. J. B. b. MASS: predict the global qualities of individual protein models using random forests and novel statistical potentials. 21, 1-10 (2020).
2. Liu, T., Wang, Y., Eickholt, J. & Wang, Z. J. S. r. Benchmarking deep networks for predicting residue-specific quality of individual protein models in CASP11. 6, 19301 (2016).
3. Won, J. et al. Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning. 87, 1351-1360 (2019).
4. Liu, T., Wang, Z. J. S. C. f. B. & Medicine. SOV_refine: A further refined definition of segment overlap score and its significance for protein structure similarity. 13, 1-10 (2018).
5. Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. (2018).
6. He, K., Zhang, X., Ren, S., Sun, J. J. I. t. o. p. a. & intelligence, m. Spatial pyramid pooling in deep convolutional networks for visual recognition. 37, 1904-1916 (2015).

Manual Prediction of Protein Tertiary and Quaternary Structures and 3D Model Refinement

L.J. McGuffin[1], R. Adiyaman[1], D.A. Brakenridge[1], N.S. Edmunds[1], L.S. Philomina[1]
1 - School of Biological Sciences, University of Reading, Reading, UK
l.j.mcguffin@reading.ac.uk

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:Y; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

For our manual predictions we used several components from our latest servers[1,2,3] (see our IntFOLD6 and ModFOLD8 server abstracts). For our tertiary structure (TS) predictions we made use of the CASP hosted 3D server models, which we ranked using ModFOLD8_rank and then refined with our new refinement method (ReFOLD3). For our quaternary structure predictions, we used a docking and template-based approach (MultiFOLD) along with our assembly quality assessment pipeline (ModFOLDdock). Finally, clues from likely ligand binding sites (predicted with FunFOLD3), aided our manual evaluation of submitted models.

Methods

Tertiary structure predictions: The server models were ranked according to the ModFOLD8_rank global quality scores (see our ModFOLD8 abstract). The top-ranked initial model was then selected and submitted to the ReFOLD3 and MultiFOLD pipelines described below. For each model, the ModFOLD8 predicted per-residue error scores were added into the B-factor column for each set of atom records.

Refinement (ReFOLD3): For the refinement of 3D models of proteins we used a modified version of our automated ReFOLD method[3]. Our new refinement pipeline, ReFOLD3, consisted of four protocols that were similar to the original version. The major improvement for ReFOLD version 3 was the accommodation of the two new MD-based strategies. The first protocol used a rapid iterative strategy (i3Drefine[4]), and the second and third protocols both employed a more CPU/GPU intensive molecular dynamic simulation strategy (using NAMD[5]) to refine each starting model.

The second protocol included the introduction of molecular dynamics simulations that were guided by the per-residue accuracy scores obtained from ModFOLD8. The per-residue accuracy scores were used to identify the poorly modelled regions, which were then targeted for refinement to improve the overall model quality. A gradual restraint strategy based on the per-residue accuracy score was applied, which considered the degree of refinement for each residue during the MD simulations. The gradual restraints ranged from weak (0.05 kCal/mol/Å2) and strong (1 kCal/mol/Å2) harmonic positional restraints on all atoms including C-alphas according to the distribution of the per residue accuracy score produced by ModFOLD8.

For the third protocol, residue-residue contact predictions were used to guide the MD simulation. We used our Contact Distance Agreement (CDA) score, which is based on the agreement between the residue contacts predicted by DeepMetaPSICOV (see our ModFOLD8 abstract). If the CDA score was high, a stronger restraint was applied to keep the residues in contact as in the predicted 3D model. A lower CDA score indicated that the residue may be further away from the native structure, therefore it was targeted for refinement to improve the quality of the predicted 3D model. Therefore in the third protocol, another gradual restraint strategy was preferred, which considered the distribution of the CDA scores during the MD simulation.

Refined models generated from the first three protocols were then assessed and ranked using ModFOLD8_rank. The fourth protocol was a combination of the first 2 approaches, where the top-ranked model from the 2nd and 3rd protocol was then further refined using i3Drefine. Finally, all of the refined models generated by each of these protocols and the starting model were pooled and re-ranked again using ModFOLD8_rank and the final top 5 models were selected and submitted.

Quaternary structure predictions (MultiFOLD): The highest scoring models from the ReFOLD3 procedure, described above, were used to generate predicted quaternary structures using, MEGADOCK[6], FRODOCK[7], PatchDock[8] and ZDOCK[9] for dimeric complexes, and M-ZDOCK[10] for multimeric complexes. In addition to the docking strategy, a multimeric fold recognition approach was also deployed. The fold template lists (with PDB and chain IDs) generated by the IntFOLD server[1] were filtered using multimeric data extracted from PISA[11] for each template. Model assemblies were then constructed using TM-align[12] for structural superposition of tertiary models onto assemblies and PyMOL was used for visualisation and manual quality checking of the template generated models. The final predicted quaternary structures were then ranked for submission using the newly developed ModFOLDdock method described below. Furthermore, the information from our FunFOLD3 method (regarding the function and locations of putative bound ligands) along with visual inspection was used for some targets in order to manually filter the modelled complexes.

Quaternary structure model quality assessment (ModFOLDdock): The ModFOLDdock protocol uses a hybrid consensus approach for producing both global and local (interface residue) scores for predicted quaternary structures. The ModFOLDdock global score was taken as the mean score from four individual methods: ProQDock[13], QSscoreJury, DockQJury, VoroMQA[14] and ModFOLDIA. For each interacting pair of chains in a modelled complex, the ProQDock scores were simply taken and averaged to produce a global score for the complete assembly. For the QSscoreJury and DockQJury methods, pairwise comparisons were made for each quaternary structure model to every other model made for the target and then the mean QS[15] and DockQ[16] scores were calculated. The ModFOLDIA method also carries out structure-based comparisons of alternative oligomer models and can produce both global and local/per-residue interface scores. The first stage of the ModFOLDIA method was to identify the interface residues in the model to be scored (defined as <= 5Å between the heavy atoms in different chains) and then obtain the minimum contact distance (D_{min}) for each contacting residue. The second stage was to locate the equivalent residues in all other models and then obtain the mean minimum distances of those residues in all other models ($MeanD_{min}$). The final IA score for each of the interface residues in the model was the absolute difference in the S_i from the mean S_i : $IA = 1-|S_i-MeanS_i|$, where $S_i = 1/(1+(D_{min}/20)^2)$ and $MeanS_i = 1/(1+(MeanD_{min}/20)^2)$. The global ModFOLDIA score for a model was then taken as the total interface score (sum of residue scores) normalised by the maximum of either the number of residues in the interface or the mean number of interface residues across all models for the same target.

Availability

Server methods are available via http://www.reading.ac.uk/bioinf/. Software is free to download via http://www.reading.ac.uk/bioinf/downloads/.

1. McGuffin,L.J., Adiyaman,R., Maghrabi,A.H.A., Shuid,A.N., Brackenridge,D.A., Nealon,J.O. & Philomina,L.S. (2019) IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Res. 47, W408-W413. doi: 10.1093/nar/gkz322
2. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. Nucleic Acids Res., 45, W416-W421. doi: 10.1093/nar/gkx332.
3. Shuid,A.N., Kempster,R., McGuffin,L.J. (2017) ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. Nucleic Acids Res. 45, W422-W428. doi: 10.1093/nar/gkx249.
4. Bhattacharya,D., Cheng,J. (2013) i3Drefine software for protein 3D structure refinement and its assessment in CASP10. PLoS One. 8, e69648.
5. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L., Schulten,K.J. (2005) Scalable molecular dynamics with NAMD. Comput Chem. 26, 1781-802.
6. Ohue,M., Shimoda,T., Suzuki,S., Matsuzaki,Y., Ishida,T., Akiyama,Y. (2014) MEGADOCK 4.0: an ultra–high-performance protein–protein docking software for heterogeneous supercomputers. Bioinformatics. 30, 3281–3283.
7. Garzon,J.I., Lopéz-Blanco,J.R., Pons,C., Kovacs,J., Abagyan,R., Fernandez-Recio,J., Chacon,P. (2009). FRODOCK: a new approach for fast rotational protein–protein docking. Bioinformatics. 25, 2544–2551.
8. Duhovny,D., Nussinov,R., Wolfson,H.J. (2002) Efficient unbound docking of rigid molecules, in: International Workshop on Algorithms in Bioinformatics. Springer, pp. 185–200.
9. Chen,R., Li,L., Weng,Z. (2003) ZDOCK: An initial-stage protein-docking algorithm. Proteins. 52, 80–87.
10. Pierce,B., Tong,W., Weng,Z. (2005) M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics. 21, 1472–1478.
11. Krissinel,E., Henrick,K. (2007) Inference of Macromolecular Assemblies from Crystalline State. J. Mol. Biol. 372, 774–797.
12. Zhang,Y., Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-9.
13. Basu,S., Wallner,B. (2016) Finding correct protein-protein docking models using ProQDock. Bioinformatics. 32, i262-i270. doi: 10.1093/bioinformatics/btw257.
14. Olechnovič,K., Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins. 85, 1131-1145. doi: 10.1002/prot.25278.

15. Bertoni,M., Kiefer,F., Biasini,M., Bordoli,L., Schwede,T. (2017) Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. 7, 10480. doi: 10.1038/s41598-017-09654-8.

16. Basu,S, Wallner,B. (2016) DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS One. 11, e0161879. doi: 10.1371/journal.pone.0161879.

MESHI

The human MESHI group in CASP14

T. Sidi, M. Bitton and C. Keasar
Ben Gurion University of the Negev
keasar@bgu.ac.il

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y.*

The human MESHI group elaborated on the EMA predictions of the automatic MESHI servers (See the abstracts of MESHI_EMA, MESHI_consensus, and MESHI_server). Our major aim was to explore the power of EMA beyond the limits of the EMA track. Most importantly, to apply EMA at the domains level rather than whole chains. When relevant, and time permitted, we also tried to improve the decoys by introducing structural constraints derived from predicted secondary structure, disulfide bonds and ligand binding. In addition, the human MESHI group also submitted EMA predictions. These were practically automatic predictions by an experimental variant of the MESHI_consensus server.

Methods

Apparent single-domain targets: We re-submitted server decoys after side-chain repacking by SCWRL4[1] and energy minimization by the OPTIMIZE program of MESHI[2]. Model no. 1 was typically the highest ranking decoy according to MESHI_consensus and the other four were selected by visual inspection from decoys ranked high by MESHI_EMA and MESHI_SERVER.

We used structural consistency among the best models as an estimate of local quality (temperature factor). To this end, we structurally aligned each of the five submitted decoys with the other 19 top scoring decoys. The average distance between a C-alpha atom and its counterparts served as the quality estimate of all the residue's atoms.

Multi-domain targets**:** Multi domain targets and the domain boundaries within them were identified by superposition, and visual inspection of the top ranking decoys. Each domain was sent to the three MESHI EMA servers and its decoys were ranked independently. The top scoring domain models, typically from different server decoys, were manually combined and submitted.

EMA prediction

EMA predictions were performed by a submitting-group-aware variant of MESHI_consensus (see abstract) that was trained with the addition of one-hot features that depict the server submitting each decoy.

Availability

All relevant software: Latest version of MESHI and ML models are freely available upon request.

1. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795 (2009).
2. Kalisman, N. et al. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 21, 3931–3932 (2005).

MESHI_consensus

The MESHI_concensus server for estimation of model accuracy

M. Bitton and C. Keasar
Ben Gurion University of the Negev
keasar@bgu.ac.il

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y.*

The MESHI_consensus server uses LightGBM[1] based random forest regressor to estimate decoy accuracies (measured in GDT_TS) based on a unique set of structural, sequence based and ensemble-based features. The features are extracted from energy minimized decoys by the MESHI molecular modeling package[2]. In addition to the EMA submissions, the top ranking minimized structures were also considered for submission by the human group MESHI (see abstract).

Methods

Structural Features: MESHI_consensus uses the same structural features as MESHI_server (See the MESHI_server abstract). All structural features are generated by our MESHI[2] package. Most of them were developed in-house[3–5] and a few other are implementations of energy terms from the literature [6–8]. An important set of features consider the compatibility of the decoy's secondary structure and solvent accessibility, calculated by DSSP[9], with predicted ones. Specifically, we use predictions of 3-class (HCE) secondary structure by PSI-PRED[10], 8-class (DSSP) predictions by DeepCNF[11], and an in-house prediction of 13-class secondary structure[12].

***Sequence based Features*:** In addition to the structural features MESHI_EMA uses a set of 26 sequence-based features. These features depict the relative abundance of amino-acid types (20) and physio-chemical properties (e.g., hydrophobicity and positive charge) in the sequence. Obviously, all the decoys of a given target share the same values for these features, however the regressor seems to be able to use them as regularizers of the structural features.

***Ensemble based Features*:** MESHI_consensus uses two types of features that rely on the availability of an ensemble of decoys for each target. The first set of features compares each decoy with the other ensemble members, by calculating pairwise RMSD and GDT_TS and associate the decoy with the average measure. The other set of features includes ensemble averages and median valued of the structural features. Like the sequence based features, these features are shared by the ensemble members, yet the regressor is able to use them as normalizers of the other features.

***Model training*:** Towards CASP14 MESHI_consensus was trained with a non-redundant set of CASP0-13 targets.

***Prediction pipeline*:** MESHI_consensus shares most of its prediction pipeline with MESHI-server. The side-chain configurations of the decoys are refined by SCWRL4[13], and the models are subjected to energy minimization by OPTIMIZE (A MESHI program). At the end of the minimization OPTIMIZE extracts the features (both structural, sequence-based and ensemble based), hands them to the regressor and the predicted GDT_TS values are submitted.

Availability

All relevant software: Latest version of MESHI and ML models are freely available upon request.

1. Ke, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in Advances in Neural Information Processing Systems 30 (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).
2. Kalisman, N. et al. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 21, 3931–3932 (2005).
3. Levy-Moonshine, A., Amir, E. D. & Keasar, C. Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics 25, 2639–2645 (2009).
4. Amir, E.-A. D., Kalisman, N. & Keasar, C. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins: Structure, Function, and Bioinformatics 72, 62–73 (2008).
5. Maximova, T., Kalisman, N. & Keasar, C. Unpublished.
6. Summa, C. M. & Levitt, M. Near-native structure refinement using in vacuo energy minimization. PNAS 104, 3177–3182 (2007).
7. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275, 895–916 (1998).
8. Zhou, H. & Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal 101, 2043–2052 (2011).
9. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).
10. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405 (2000).
11. Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports 6, 18962 (2016).
12. Sidi, T. & Keasar, C. Redundancy-weighting the PDB for detailed secondary structure prediction using deep-learning models. Bioinformatics 36, 3733–3738 (2020).
13. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795 (2009).

MESHI_EMA

The MESHI_EMA server for estimation of model accuracy

M. Bitton and C. Keasar
Ben Gurion University of the Negev
keasar@bgu.ac.il

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y.*

The MESHI_EMA server uses LightGBM[1] based random forest regressor to estimate decoy accuracies (measured in GDT_TS) based on a unique set of structural and sequence based features. The features are extracted from energy minimized decoys by the MESHI molecular modeling package[2]. In addition to the EMA submissions, the top ranking minimized structures were also considered for submission by the human group MESHI (see abstract).

Methods

Structural Features: MESHI_EMA uses the same structural features as MESHI_server (See the MESHI_server abstract). All structural features are generated by our MESHI[2] package. Most of them were developed in-house[3–5] and a few other are implementations of energy terms from the literature [6–8]. An important set of features consider the compatibility of the decoy's secondary structure and solvent accessibility, calculated by DSSP[9], with predicted ones. Specifically, we use predictions of 3-class (HCE) secondary structure by PSI-PRED[10], 8-class (DSSP) predictions by DeepCNF[11], and an in-house prediction of 13-class secondary structure[12].

Sequence based Features: In addition to the structural features MESHI_EMA uses a set of 26 sequence-based features. These features depict the relative abundance of amino-acid types (20) and physio-chemical properties (e.g., hydrophobicity and positive charge) in the sequence. Obviously, all the decoys of a given target share the same values for these features, however the regressor seems to be able to use them as regularizers of the structural features.

Model training: Towards CASP14 MESHI_EMA was trained with a non-redundant set of CASP0-13 targets.

Prediction pipeline**:** MESHI_EMA shares most of its prediction pipeline with MESHI-server. The side-chain configurations of the decoys are refined by SCWRL4[13], and the models are subjected to energy minimization by OPTIMIZE (A MESHI program). At the end of the minimization OPTIMIZE extracts the features (both structural and sequence-based), hands them to the regressor and the predicted GDT_TS values are submitted.

Availability

All relevant software: Latest version of MESHI and ML models are freely available upon request.

1. Ke, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. in Advances in Neural Information Processing Systems 30 (eds. Guyon, I. et al.) 3146–3154 (Curran Associates, Inc., 2017).
2. Kalisman, N. et al. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 21, 3931–3932 (2005).
3. Levy-Moonshine, A., Amir, E. D. & Keasar, C. Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics 25, 2639–2645 (2009).

4. Amir, E.-A. D., Kalisman, N. & Keasar, C. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins: Structure, Function, and Bioinformatics 72, 62–73 (2008).
5. Maximova, T., Kalisman, N. & Keasar, C. Unpublished.
6. Summa, C. M. & Levitt, M. Near-native structure refinement using in vacuo energy minimization. PNAS 104, 3177–3182 (2007).
7. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275, 895–916 (1998).
8. Zhou, H. & Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal 101, 2043–2052 (2011).
9. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).
10. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405 (2000).
11. Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports 6, 18962 (2016).
12. Sidi, T. & Keasar, C. Redundancy-weighting the PDB for detailed secondary structure prediction using deep-learning models. Bioinformatics 36, 3733–3738 (2020).
13. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795 (2009).

MESHI_server

The MESHI-server pipeline for estimation of model accuracy and template based structure prediction

T. Sidi and C. Keasar
Ben Gurion University of the Negev
keasar@bgu.ac.il

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y.*

MESHI server submitted predictions in two CASP14 tracks: EMA and tertiary structure predictions. At the core of the server lies our EMA method MESHI-score[1,2]. MESHI-score uses a unique set of structural features to predict the quality of protein decoys in terms of GDT_TS. The structural features are extracted from energy minimized decoys by our MESHI molecular modeling package[3].

For tertiary structure prediction we used MESHI-score to choose the top ranking decoys among 200alternatives, which are generated by HHPRED[4] and MODELLER[5] and energy minimized by the OPTIMIZE program (part of the MESHI package). For EMA prediction we first minimize the energies of server decoys and then extract their features for MESHI-score. The top ranking minimized structures are also considered for submission by the human group MESHI (see astract).

Methods

MESHI-score and features: MESHI-score is an ensemble learning method that apply non-linear regression to features-vector. All the features are generated by the MESHI package. Most of them were developed in-house[6–8] and a few other are implementations of energy terms from the literature [9–11]. An important set of features consider the compatibility of the decoy's secondary structure and solvent accessibility, calculated by DSSP[12], with predicted ones. In previous CASPs we used predictions of 3-class (HCE) secondary structure by PSI-PRED[13] and 8-class predictions by DeepCNF[14]. For CASP14 we augmented them with an in-house prediction of 13-class secondary structure[15]. Towards CASP14 MESHI-score was trained with a non-redundant set of CASP0-13 targets.

Tertiary structure prediction pipeline: The pipeline for tertiary structure prediction relies on HHPRED[4] for template identification and alignment. Thus, it can submit tertiary structure predictions only for TBM targets. For each of these targets we use MODELLER[5] to build 200 alternative decoys. The side-chain configurations of the decoys are further refined by SCWRL4[16], and the models are subjected to energy minimization by OPTIMIZE. At the end of the minimization OPTIMIZE extracts the structural features and hands them to MESHI-score. The 5 top ranking decoys are selected for submission. We use structural consistency among the best models as an estimate of local quality (temperature factor). To this end, the server structurally align each of the five submitted decoys with the other 19 top scoring decoys. The average distance between a C-alpha atom and its counterparts served as the quality estimate of all the residue's atoms.

Estimating the accuracy of server models: For the EMA task, MESHI-server uses almost the same pipe-line as used for structure prediction. The decoy generation steps are bypassed by the server decoys, and the MESHI-score results are submitted.

Availability

All relevant software: Latest version of MESHI and ML models are freely available upon request.

1. Elofsson, A. et al. Methods for estimation of model accuracy in CASP12. Proteins: Structure, Function, and Bioinformatics 86, 361–373 (2018).
2. Mirzaei, S., Sidi, T., Keasar, C. & Crivelli, S. Purely structural protein scoring functions using support vector machine and ensemble learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics 16, 1515–1523 (2019).
3. Kalisman, N. et al. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 21, 3931–3932 (2005).
4. Zimmermann, L. et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Journal of Molecular Biology 430, 2237–2243 (2018).
5. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 234, 779–815 (1993).
6. Levy-Moonshine, A., Amir, E. D. & Keasar, C. Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. Bioinformatics 25, 2639–2645 (2009).
7. Amir, E.-A. D., Kalisman, N. & Keasar, C. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins: Structure, Function, and Bioinformatics 72, 62–73 (2008).
8. Maximova, T., Kalisman, N. & Keasar, C. Unpublished.
9. Summa, C. M. & Levitt, M. Near-native structure refinement using in vacuo energy minimization. PNAS 104, 3177–3182 (2007).
10. Samudrala, R. & Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275, 895–916 (1998).
11. Zhou, H. & Skolnick, J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal 101, 2043–2052 (2011).
12. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637 (1983).
13. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405 (2000).
14. Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports 6, 18962 (2016).
15. Sidi, T. & Keasar, C. Redundancy-weighting the PDB for detailed secondary structure prediction using deep-learning models. Bioinformatics 36, 3733–3738 (2020).
16. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795 (2009).

ModFOLD8, ModFOLD8_cor, ModFOLD8_rank

Automated 3D Model Quality Assessment using the ModFOLD8 Server

L.J. McGuffin[1]
1 - School of Biological Sciences, University of Reading, Reading, UK
l.j.mcguffin@reading.ac.uk

***Key:** Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:Y; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

The ModFOLD8 server is the latest version of our web resource for the Quality Assessment (QA) of 3D models of proteins[1,2,3].

Methods

ModFOLD8 is our new approach to QA that combines the strengths of multiple pure-single and quasi-single model methods for improving prediction accuracy. For CASP14, again our emphasis was on increasing the accuracy of: per-residue assessments for single models, single model ranking and score consistency. Each model was considered individually using 9 pure-single model methods: CDA[3], SSA[3], ProQ2[4], ProQ2D[5], ProQ3D[5], VoroMQA[6], ProQ4[7], CDA_DMP and CDA_SC. CDA_DMP and CDA_SC are two new pure single model scoring methods, based on the original Contact Distance Agreement (CDA) score, which relates to the agreement between the predicted residue contacts according to MetaPSICOV[8] and the measured Euclidean distance (in Å) between residues in the model[3]. However, the contact predictions from DeepMetaPSICOV[9] and SPOT-Contact[10] were used as inputs for CDA_DMP and CDA_SC respectively. Additionally, sets of reference 3D models (generated using IntFOLD6 - see our other abstract) were used to score models using 4 alternative quasi-single model methods: DBA[3], MF5s[3], MFcQs[3] and ResQ[11]. Neural networks (NNs) were then used to combine the component per-residue/local quality scores from each of the 13 alternative scoring methods, resulting in a final consensus of per-residue quality scores for each model.

Per-residue/local quality scoring methods: Two ModFOLD8 NNs were trained using two separate target functions for each residue in a model: the superposition based S-score used previously[3] (for ModFOLD8_res), and the residue contact based lDDT score[12] (for ModFOLD8_res_lddt). The local scoring methods were trained using a simple multilayer perceptron (MLP). The MLP inputs consisted of a sliding window (size=5) of per-residue scores from all 13 of the scoring methods described above, and the output was a single quality score (i.e. either the S-score or lDDT as the target function) for each residue in the model (65 inputs, 33 hidden, 1 output). The RSNNS package for R was used to construct the NNs, which were trained using data derived from the evaluation of CASP11 & 12 server models versus native structures. For both of the per-residue methods, the similarity scores, s, for each residue were converted back to distances, d, with $d = 3.5\sqrt{((1/s)-1)}$.

Global scoring methods: Global scores were calculated by taking the mean per-residue scores (the sum of the per-residue similarity scores divided by sequence lengths) for each of the 13 individual component methods, described above, plus the NN output from ModFOLD8_res and ModFOLD8_res_lddt. Furthermore, 3 additional quasi-single global model quality scores were generated for each model based on the original ModFOLDclust, ModFOLDclustQ and ModFOLDclust2 global scoring methods[13]. Thus, we ended up with 18 alternative global QA scores, which could be combined in various ways in order to optimize for the different facets of the quality estimation problem. We registered three ModFOLD8 global scoring variants:

The ModFOLD8 global score ((ModFOLDclust2 + DBA + ProQ3D + VoroMQA + CDA_SC)/5) was found to have a good balance of performance both for correlations of predicted and observed scores and rankings of the top models.

The ModFOLD8_cor global score variant *((ModFOLDclustQ + MFcQs + DBA + ProQ3D + ResQ + VoroMQA)/6)* was found to be an optimal combination for producing good correlations with the observed scores, i.e. the predicted global quality scores produced should produce closer to linear correlations with the observed global quality scores.

The ModFOLD8_rank global score variant (*(SSA + ProQ3D + VoroMQA + CDA_DMP + CDA_SC + ProQ4 + ModFOLD8res + ModFOLD8res_lDDT)/8)* was found to be an optimal combination for ranking, i.e. the top ranked models (top 1) should be closer to the highest observed accuracy, but the relationship between predicted and observed scores may not be linear.

The local scores provided in the submission files for the ModFOLD8 and ModFOLD_rank variants used the output from the ModFOLD8_res NN, whereas the ModFOLD_cor variant used the local scores from the ModFOLD8_res_lddt NN.

Results

The ModFOLD8_rank method is used to evaluate models as part of the IntFOLD6 server (see our IntFOLD6 abstract), which is continuously benchmarked using the CAMEO resource[14]. According to the CAMEO results, IntFOLD6 has shown improved performance over our last three methods (IntFOLD3, IntFOLD4 & IntFOLD5) and it is outperformed by just one public server in the benchmark.

Availability

The ModFOLD8 server is available at:
https://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD8_form.html

1. McGuffin,L.J. (2008) The ModFOLD Server for the Quality Assessment of Protein Structural Models. Bioinformatics. 24, 586-587.
2. McGuffin,L.J., Buenavista, M.T., Roche, D.B. (2013) The ModFOLD4 Server for the Quality Assessment of 3D Protein Models. Nucleic Acids Res., 41, W368-72.
3. Maghrabi,A.H.A. & McGuffin,L.J. (2017) ModFOLD6: an accurate web server for the global and local quality estimation of 3D models of proteins. Nucleic Acids Res., 45, W416-W421, doi: 10.1093/nar/gkx332.
4. Uziela,K., Wallner,B. (2016). ProQ2: estimation of model accuracy implemented in Rosetta. Bioinformatics. 32, 1411-3.

5. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B., Elofsson,A. (2017) ProQ3D: improved model quality assessments using deep learning. Bioinformatics. 33, 1578-1580. doi: 10.1093/bioinformatics/btw819.
6. Olechnovič,K., Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins. 85, 1131-1145. doi: 10.1002/prot.25278.
7. Cheng,J., Choe,M.H., Elofsson,A., Han,K.S., Hou,J., Maghrabi,A.H.A., McGuffin,L.J., Menéndez-Hurtado,D., Olechnovič,K., Schwede,T., Studer,G., Uziela,K., Venclovas,Č. & Wallner,B. (2019) Estimation of model accuracy in CASP13. Proteins. 87, 1361-137. doi: 10.1002/prot.25767
8. Jones,D.T., Singh,T., Kosciolek,T., Tetchner,S. (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 31, 999–1006.
9. Kandathil,S.M., Greener,J.G., Jones,D.T. (2019) Prediction of inter-residue contacts with DeepMetaPSICOV in CASP13. Proteins. 87, 1092-1099. doi: 10.1002/prot.25779
10. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y. (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics. 34, 4039–4045.
11. Yang,J., Wang,Y., Zhang,Y. (2016) ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. J Mol Biol. 428, 693-701. doi: 10.1016/j.jmb.2015.09.024.
12. Mariani,V., Biasini,M., Barbato,A. & Schwede,T. (2013) lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 29, 2722-8.
13. McGuffin,L.J., Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics. 26, 182-188.
14. Haas,J., Barbato,A., Behringer,D., Studer,G., Roth,S., Bertoni,M., Mostaguir,K., Gumienny,R., Schwede,T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins. 86 S1, 387-398. doi: 10.1002/prot.25431.

ModFOLDclust2

Automated 3D Model Quality Assessment using ModFOLDclust2

L.J. McGuffin[1]

1 - School of Biological Sciences, University of Reading, Reading, UK

l.j.mcguffin@reading.ac.uk

***Key:** Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

The ModFOLDclust2 method[1] is a leading automatic clustering based approach for both local and global 3D model quality assessment[2].

Methods

The ModFOLDclust2 server that was evaluated during CASP14 was identical to that tested during the CASP9, CASP10, CASP11, CASP12 & CASP13 experiments. The ModFOLDclust2 method was originally developed to provide increased prediction accuracy, over the original ModFOLDclust method[3,4], with minimal additional computational overhead. The global QA score from ModFOLDclust2 is simply the mean of the global QA scores obtained from the ModFOLDclustQ method and the original ModFOLDclust method. ModFOLDclustQ is similar to our previous ModFOLDclust method, however a modified version of the structural alignment free Q-measure[5] is used instead of the TM-score[6] in order to carry out all-against-all pairwise model comparisons. The per-residue QA scores for ModFOLDclust2 were just taken directly from ModFOLDclust, as no advantage was gained from simply combining the per-residue scores with those from ModFOLDclustQ.

Results

ModFOLDclust2 has been independently evaluated by the CASP assessors since CASP9 and has consistently ranked among the top performing QA methods[2,7,8,9,10].

Availability

ModFOLDclust2 can be run as an option via the ModFOLD server (version 3.0):
http://www.reading.ac.uk/bioinf/ModFOLD/ModFOLD_form_3_0.html
The ModFOLDclust2 software is also available to download as a standalone program via:
http://www.reading.ac.uk/bioinf/downloads/.

1. McGuffin,L.J., Roche,D.B. (2010) Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics. 26, 182-188.
2. Kryshtafovych,A., Fidelis,K., Tramontano,A. (2011) Evaluation of model quality predictions in CASP9. Proteins. 79 S10, 91-106.
3. McGuffin,L.J (2007) Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics. 8, 345.

4. McGuffin,L.J. (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins. 77 S9, 185-190.
5. Ben-David,M., Noivirt-Brik,O., Paz,A., Prilusky,J., Sussman,J.L. and Levy,Y. (2009) Assessment of CASP8 structure predictions for template free targets, Proteins. 77 S9, 50-65.
6. Zhang,Y., Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. Proteins. 57, 702-710.
7. Kryshtafovych,A., Barbato,A., Fidelis,K., Monastyrskyy,B., Schwede,T., Tramontano,A. (2013) Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins. 82 S2, 112-26.
8. Kryshtafovych,A., Barbato,A., Monastyrskyy,B., Fidelis,K., Schwede,T., Tramontano,A. (2015) Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins. 84 S1, 349-69. doi: 10.1002/prot.24919.
9. Kryshtafovych,A., Monastyrskyy,B., Fidelis,K., Schwede,T., Tramontano,A. (2018) Assessment of model accuracy estimations in CASP12. Proteins. 86 S1, 345-360. doi: 10.1002/prot.25371.
10. Won,J., Baek,M., Monastyrskyy,B., Kryshtafovych,A., Seok,C. (2019) Assessment of Protein Model Structure Accuracy Estimation in CASP13: Challenges in the Era of Deep Learning. Proteins. 87, 1351-1360. doi: 10.1002/prot.25804.

MUFOLD

MUFOLD_REFINE: Iterative Protein Structure Refinement Using Potential Functions based on Distance Distributions

Junlin Wang[1], Dong Xu[1,2] and Yi Shang[1]

[1] -Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA, [2] - Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA
jwyn@umsystem.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:Y*

MUFOLD_REF is a new population-based iterative protein structure refinement method including optimization, model hybridization, and model selection. It uses distance distributions for each amino-acid pair of the target protein, which are generated based on a set of reference models. A new potential function has been designed base on the distance distributions for model optimization.

Methods

MUFOLD_REF takes a protein model to be refined (called target model) and a set of reference models. In CASP 14, we used filtered CASP server predictions as the reference models to generate distance distributions for each amino-acid pair of the target protein. A potential function is designed base on distance distributions and a gradient descent algorithm (e.g., L-BFGS) is used to maximize the potential function. The optimized models are then used for model selection and structure hybridization. After a few iterations (e.g., 3), the top five models selected by MUFOLD single models QA method will be the final refinement output. The main steps are as follows:

1. A set of ten models, including the target model and nine other models selected from the set of reference models by the MUFOLD single model QA method, are optimized individually to generate 10 new models.

2. MODELLER[1] is used to generate two models from each of the ten models. Then, from the total 30 models, the MUFOLD single model QA method is used to select top 5 models. These 5 models together with the target model are given to a model hybridization process to generate 150 new models. During hybridization, the target model is combined with the top 1 selected model to generate 10 new models, combined with top two selected models to generate 20 new models, and so on. A set of 180 models are formed to include these 150 models and the 30 models after MODELLER.

3. MUFOLD single model QA method is used to select top 9 models from the 180 models. Go to Step 1), where the new set of ten models includes these 9 models and the target model.

After three iterations of these steps, MUFOLD single model QA method is used to select the top five models as final output.

MUFOLD single model QA method is a two-stage single model QA. A QA score generated base on energy and structural features using the machine learning algorithm, random forests, are combined with Rosetta[2] energy score and Proq3[3] score by a SVM model to generate the final QA score.

1. Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 1993, 234, 779−815.

2. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation 13, 3031-3048 (2017).

3. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B., & Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. Bioinformatics, 33, 1578-1580.

MUFOLD_H

MUFOLD_HUMAN: Protein Structure Prediction Using Potential Functions based on Distance Distributions

Junlin Wang[1], Dong Xu[1,2] and Yi Shang[1]

[1] -Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA, [2] - Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA
jwyn@umsystem.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:Y*

MUFOLD_HUMAN is a new population-based iterative method for protein structure prediction, including optimization, model hybridization, and model selection. It uses distance distributions for each amino-acid pair of the target protein, which are generated based on a set of reference models. A new potential function has been designed base on the distance distributions for model optimization.

Methods

MUFOLD_HUMAN performs protein structure generation based on a set of reference models. In CASP 14, we used filtered CASP server predictions as the reference models to generate distance distributions for each amino-acid pair of the target sequence. A potential function is designed base on distance distributions and a gradient descent algorithm (e.g., L-BFGS) is used to maximize the potential function. The optimized models are then used for model selection and structure hybridization. After a few iterations (e.g., 3), the top five models selected by MUFOLD single models QA method will be the final prediction output. The main steps are as follows:

1. A set of ten models selected from the set of reference models by the MUFOLD single model QA method, are optimized individually to generate 10 new models.

2. MODELLER[1] is used to generate two models from each of the ten models. Then, from the total 30 models, the MUFOLD single model QA method is used to select top 6 models. These 6 models are given to a model hybridization process to generate 150 new models. During hybridization, the top 2 selected models are combined to generate 10 new models; the top 3 selected models are combined to generate 20 new models; and so on. A set of 180 models is formed to include these 150 models and the 30 models after MODELLER.

3. MUFOLD single model QA method is used to select top 10 models from the 180 models. Then, go to Step 1).

After three iterations of these steps, MUFOLD single model QA method is used to select the top five models as final output.

MUFOLD single model QA method is a two-stage single model QA. A QA score generated base on energy and structural features using the machine learning algorithm, random forests, are combined with Rosetta[2] energy score and Proq3[3] score by a SVM model to generate the final QA score.

1. Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 1993, 234, 779−815.

2. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation 13, 3031-3048 (2017).

3. Uziela,K., Menéndez Hurtado,D., Shu,N., Wallner,B., & Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. Bioinformatics, 33, 1578-1580.

MUfoldQA_G

MUfoldQA_G: A New Multi-Model QA Method Based on Machine Learning

Wenbo Wang[1], Dong Xu[1,2] and Yi Shang[1]

[1] *Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA,* [2] *Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA*
wwr34@mail.missouri.edu

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

MUfoldQA_G is a new multi-model QA method featuring a newly designed 2-stage machine learning scheme to improve over the naïve consensus method. In this method, first, a pre-trained model is used to make initial predictions of QA scores of a set of candidate models. Then, a second model is trained on demand to generate more accurate predictions.

Methods

The input of MUfoldQA_G is a target protein sequence S and a set of n candidate models (M_i, i=[1, n]) to be evaluated. Its output is a score for each candidate model, in the range of 0 to 1, with 1 being the highest quality–identical to the native structure of the protein.

Step 1. Use the target protein sequence S to query a PDB database with Blast[1] and HHsearch[2], to find similar proteins as templates.

Step 2. Select a subset of the candidate models as reference models. If the number of candidate model is smaller than 50, use the entire set as the reference model set. Otherwise, sort all candidate models using their MQAPRank[3] scores and choose top 45% as the reference model set.

Step 3. Run our previously published MUfoldQA_S[4,5] method to calculate the local scores, W, for each C-alpha position of each reference model using templates generated in Step 1.

Step 4. Between each candidate model M_x and each reference model, calculate pairwise GDT-TS value G.

Step 5. For each candidate model, calculate a score Z equal to the weighted average of W and G.

Step 6. If the size of the candidate model set is larger than 50, run *Subroutine V=ReCon(M)*, and the final score of a candidate model is a linear combination of Z and V. Otherwise, directly output Z as the final score.

Subroutine V=ReCon (M):

Step 1. Calculate pairwise GDT-TS score R_{xy} between each candidate model M_x and all n candidate models (M_y, y=[1, n]).

Step 2. For each candidate model M_x, calculate naïve consensus score C_x = sum(R_{xy},y=[1,n])/n.

Step 3. Sort candidate models (M_x, x=[1, n]) based on C_x from high to low as (P_x, x=[1, n]).

Step 4. Calculate pairwise GDT-TS score Q_{xy} between each candidate P_x model and all n candidate models (P_y, y=[1, n]).

Step 5. For each candidate model P_x, generate a feature set F_x=[$Q_{x,1}$, $Q_{x,2}$, … , $Q_{x,n}$], then feed F_x into pretrained model U to generate initial QA score T_x. The model U is based on bagged trees algorithm and the training set comprises of server prediction models from CASP5 to CASP12. Each training model is run through Subroutine ReCon's Step 1-5 to generate its input feature vector, and its true GDT-TS score is used as the label.

Step 6. Generate a new training set with model quality distribution mimicking the distribution of T_x (x=1, n) and train a new on-demand model Z using bagged trees algorithm.

Step 7. For each candidate model P_x, generate a feature set F_x=[$Q_{x,1}$, $Q_{x,2}$, … , $Q_{x,n}$], then feed F_x into on-demand model Z to generate new QA score V_x.

Return the QA score V.

Availability

At this moment, MUfoldQA_S is available at http://qas.wangwb.com/~wwr34/mufoldqa/. Other tools like Blast, HHsearch and MQAPRank can be downloaded from their corresponding websites. We plan to release an integrated version in the near feature.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
2. Soding,J. (2004). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21, 951–960.
3. Jing,X. Dong,Q. Liu,X & Liu,B. (2015). Protein model quality assessment by learning-to-rank. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 91-96.
4. Wang,W., Li,Z., Wang,J., Xu,D. & Shang,Y. (2019). PSICA: a fast and accurate web service for protein model quality analysis, Nucleic Acids Res. 47, W443-W450, doi: 10.1093/nar/gkz402
5. Wang,W., Wang,J., Xu,D. & Shang,Y. (2020). Two New Heuristic Methods for Protein Model Quality Assessment. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17, 1430-1439, doi: 10.1109/TCBB.2018.2880202.

MufoldQA_X: A New Quasi-Single-Model QA Method Combining Template Based and Deep Learning Based Features

Wenbo Wang[1], Dong Xu[1,2] and Yi Shang[1]

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, 65211, USA, [2] Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri, 65211, USA

wwr34@mail.missouri.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

MUfoldQA_X is a new quasi-single-model QA method in the MUfoldQA family. This method takes advantage of information from both templates and predicted pairwise distance probability distributions by deep neural networks. Machine learning models, such as bagged trees, have been trained to combine the information from multiple sources and generate a QA score prediction for a protein model.

Methods

The input is a target protein sequence S and a model (or decoy) M, and the output is a quality score in the range of 0 to 1, with 1 being the highest quality–the same as the native structure of the target protein. MUfoldQA_X calculates a set of features and feeds the feature array into a machine learning model to make the final prediction. The machine learning model we used in CASP14 is bagged trees and the training set we used to train the model contains server prediction models from CASP5 to CASP12. The features of the machine learning models include

1. MUfoldQA_S[1-2] score (1 feature).

Generated by running our previously published MUfoldQA_S method on target protein sequence S and predicted model M.

2. ProSPr-based[3] features (300 features).

Run ProSPr with target protein sequence S to generate a 64-bin probability distribution for each cell in the distance matrix between C-alpha atoms of the protein S. Extract the 3D coordinates of C-alpha atoms of model M and calculate the pairwise distance matrix T of M. For each number in T, find its corresponding probability in the ProSPr prediction result. Repeat this process for the entire T to generate a probability matrix P. Then, calculate the product of all numbers along 1, 2, …, 300 off-diagonal of P, respectively, to generate 300 features.

3. Template-based features (300 features).

Run the same process as in 2, but replace ProSPr predictions with probability distributions derived from templates generated by Blast[4] or HHsearch[5], respectively, which leads to another 300 features.

Availability

At this moment, MUfoldQA_S is available at http://qas.wangwb.com/~wwr34/mufoldqa/. Other tools like Blast, HHsearch and ProSPr can be downloaded from their corresponding websites. We plan to release an integrated version in the near feature.

1. Wang,W., Li,Z., Wang,J., Xu,D. & Shang,Y. (2019). PSICA: a fast and accurate web service for protein model quality analysis, Nucleic Acids Res. 47, W443-W450, doi: 10.1093/nar/gkz402
2. Wang,W., Wang,J., Xu,D. & Shang,Y. (2020). Two New Heuristic Methods for Protein Model Quality Assessment. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17, 1430-1439, doi: 10.1109/TCBB.2018.2880202.
3. Billings,W.M., Hedelius,B., Millecam,T., Wingate,D. & Della Corte,D. (2019). ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. BioRxiv, 830273.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
5. Soding,J. (2004). Protein homology detection by HMM-HMM comparison. Bioinformatics, 21, 951–960.

MULTICOM (TS)

CASP14 Tertiary Structure Prediction by MULTICOM Human Group

Jian Liu[1], Jie Hou[2], Tianqi Wu[1], Zhiye Guo[1], J. Cheng[1*]

1-Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211; 2-Department of Mathematics and Computer Science, St. Louis University, St. Louis, MO 63103

*Corresponding author: chengji@missouri.edu

The main improvement of our CASP14 MULTICOM human tertiary structure predictor over our CASP13 human predictor[1] is to extensively use deep learning-based inter-residue distance prediction in template-free (ab initio) tertiary structure prediction and model quality assessment.

Methods

The input for MULTICOM predictor included all the CASP14 server models plus some extra *ab initio* models built from inter-residue distance maps predicted by DeepDist[2] with deeper alignments generated from larger updated protein sequence databases if necessary.

The redundant models with high similarity from the same server predictor were filtered out. The five automated quality assessment (QA) methods (MULTICOM-CLUSTER, MULTICOM-CONSTRUCT, MULTICOM-HYBRID, MULTICOM-DEEP, MULTICOM-DIST) that integrated a number of single-model and multi-model QA scores and inter-residue distance/contact features were applied to evaluate the quality of the models (for more details, see our MULTICOM QA abstracts). The consensus of these QA predictions and human inspections were used to select top five models. Each top ranked model was combined with other similar models to generate a combined model. If the combined model was sufficiently similar to the original model, it was used as one of final top models. Otherwise, the top ranked model was refined by 3Drefine[3] or Modrefiner[4] to generate the final model.

If a target was predicted to have multiple domains, the same protocol above was applied to each domain separately to generate five top models for each domain. The top five models of all the domains were joined together to form final five full-length models for the target.

1. Hou, J., Wu, T., Cao, R., & Cheng, J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1165-1178, 2019.
2. Wu, T., Guo, Z., Hou, J., and Cheng, J. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. bioRxiv, 2020.
3. Bhattacharya, D., Nowotny, J., Cao, R., & Cheng, J. 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic acids research, 44(W1), W406-W409, 2016.
4. Xu, D. and Zhang, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophysical Journal, vol 101, 2525-2534, 2011.

MULTICOM-AI (TS)

Prediction of Protein Interchain Contacts and Complex Structures in CASP14-CAPRI Experiment

Raj S. Roy, Farhan Quadir, Jian Liu and Jianlin Cheng*
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
*chengji@missouri.edu

Our MULTICOM-AI protein complex structure predictor uses an *ab initio* deep learning-based intrachain contact prediction tool (DNCON2[1]) as well as a template-based prediction method to predict interchain residue-residue contacts for protein complex targets in CASP14 and CAPRI 50. Interchain residue pairs are considered interchain contacts if the Euclidean distance between the heavy atoms of the residues is less than or equal to 6.0 Å[2]. These contacts are then used to predict the final structure of the multimeric complex using the distance-geometry protocol of Crystallography & NMR System (CNS)[3].

Methods

Our system takes as input the individual sequences and predicted ("known") tertiary structures of the subunits of the target protein complex. Tertiary structures are predicted by our MULTICOM-CLUSTER predictor. The workflow of our system is illustrated in **Figure 1** and the details described in the following steps:

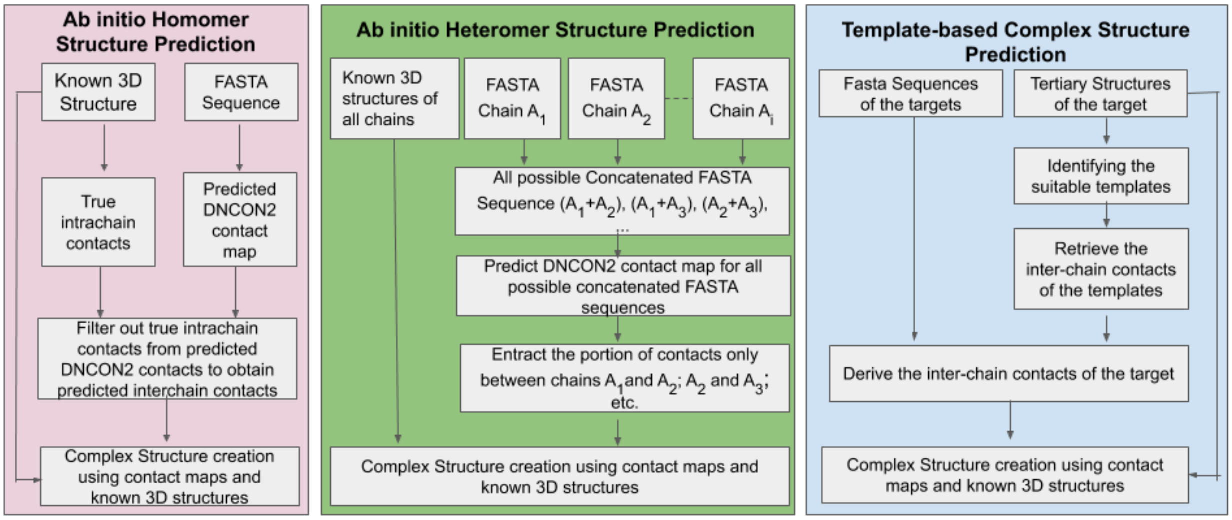

Ab initio Homomer Structure Prediction
Known 3D Structure
FASTA Sequence
True intrachain contacts
Predicted DNCON2 contact map
Filter out true intrachain contacts from predicted DNCON2 contacts to obtain predicted interchain contacts
Complex Structure creation using contact maps and known 3D structures
Ab initio Heteromer Structure Prediction
Known 3D structures of all chains
FASTA Chain A_1
FASTA Chain A_2
FASTA Chain A_i
All possible Concatenated FASTA Sequence (A_1+A_2), (A_1+A_3), (A_2+A_3), ...
Predict DNCON2 contact map for all possible concatenated FASTA sequences
Entract the portion of contacts only between chains A_1 and A_2; A_2 and A_3; etc.
Complex Structure creation using contact maps and known 3D structures
Template-based Complex Structure Prediction
Fasta Sequences of the targets
Tertiary Structures of the target
Identifying the suitable templates
Retrieve the inter-chain contacts of the templates
Derive the inter-chain contacts of the target
Complex Structure creation using contact maps and known 3D structures

Figure 1: The Pipeline of the MULTICOM-AI Complex Structure Prediction System

1. Interchain Contact Prediction: The first step is to determine the pairwise contacts between the residues of the individual chains, i.e. the interchain contacts between two chains. This prediction can thus be treated as contact prediction between dimers - homodimer if the chains are identical; heterodimer otherwise. For *ab initio* homodimer contact prediction, the system predicts interchain contacts using DNCON2. Since the multiple sequence alignment (MSA) generated (using UniRef30 database) for intrachain contact prediction is the same as that used for homomeric complexes, it is assumed that DNCON2 predicts both intrachain and interchain contacts from MSAs. Interchain contacts in homomeric complexes are obtained by removing the intrachain contacts inferred from the tertiary structure of a protein chain from all the predicted contacts. The process for *ab initio* heterodimer contact prediction is slightly different. At first, we concatenate the individual sequences of the two chains. In order to generate MSA for heterodimers, we run JackHMMER on a custom database of concatenated sequences of all possible interacting heterodimers derived from the protein data bank (PDB). This is then fed to DNCON2 to generate the interchain co-evolutionary features and predict the output contact map. The desired portion from this DNCON2 predicted contact map that captures interchain residue-residue relationships is extracted as the final interchain contact map. For template-based contact prediction, we search the predicted tertiary structures of the protein chains against our custom dimer database to find complex templates and then extract the interchain contacts. The tertiary structures of the dimers and their respective interchain contact maps are then fed into our complex structure creation system to generate the final structures of the entire protein complex which is described next.

2. Complex Structure Prediction: This is implemented using the distance geometry protocol of the Crystallography & NMR System (CNS) which uses a stochastic simulated annealing method to build the complex structure of protein dimers leveraging interchain protein contacts and tertiary structures of individual protein chains. The tool can build complex structures consisting of two or more protein chains, keeping the individual protein chains unchanged and also satisfying the provided interchain contacts between them as much as possible. It is used to generates 100 models, which are then sorted in ascending order of the distance-restrain energy and the top 5 models are selected.

3. Model Selection: The best five models from each of the methods described above are largely selected based on two criterion - the minimum distance-restrain energy and the maximum number of inter-chain contact satisfaction.

1. Adhikari, B., Hou, J., & Cheng, J. (2018). DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics, 34(9), 1466-1472.
2. Hopf, T. A., Schärfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., ... & Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife, 3, e03430.
3. Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. Nature protocols, 2(11), 2728.

MULTICOM-CLUSTER, MULTICOM-CONSTRUCT, MULTICOM-AI (QA)

Protein model quality assessment with deep learning and residue-residue contact and distance predictions

Jian Liu[1$], Xiao Chen[1$], Jie Hou[2], Tianqi Wu[1], Zhiye Guo[1], J. Cheng[1*]
[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211;
[2]Department of Mathematics and Computer Science, St. Louis University, St. Louis, MO 63103
[$]equal contribution; *corresponding author: chengji@missouri.edu

In CASP14, our MULTICOM-CLUSTER and MULTICOM-CONSTRUCT quality assessment (QA) methods is based on DeepRank[1] that were successfully tested in CASP13. DeepRank uses a deep learning network to integrate a number of QA features to predict the global quality of protein structural models. MULTICOM-AI is based on DeepRank2[2] that adds new inter-residue distance features on top of DeepRank.

Methods

The features used by MULTICOM-CLUSTER and CONSTRUCT include single-model QA features (i.e. DNCON2[3], SBROD[4], OPUS_PSP[5], RF_CB_SRS_OD[6], Rwplus[7], DeepQA[8], ProQ2[9], ProQ3[10], Dope[11], Voronota[12], QMEAN[13] and Model evaluator[14]) and multi-model QA features (i.e. APOLLO[15], Pcons[16] and ModFOLDclust2[17]). Residue-residue contact predictions generated by DNCON2 are used to calculate the percentage of top contacts (i.e. short-range, medium-range, long-range contacts) matched with the contact map of a protein structural model. These features are used by 10 deep networks to predict 10 global quality scores.

MULTICOM-CLUSTER combine the 10 predicted scores with the original input features as inputs for another deep neural network to generate the final quality scores, while the final scores of MULTICOM-CONSTRUCT are the simple average of the 10 predicted scores.

The new inter-residue distance feature added into MULTICOM-AI is the correlation between a selected set of inter-residue distances in a structural model and that of the distance map of a target predicted by DeepDist[18]. A distance between two residues is selected if the sequence separation between them is greater than or equal to a threshold (i.e. 6), and the distance is less than or equal to a threshold (i.e. 16 Angstrom). Moreover, the inter-residue contact-based features of MULTICOM-AI are updated by replacing contact predictions made by DNCON2 with those made by DNCON4[19]. We trained and validated the deep neural networks of MULTICOM-AI on the models of CASP8-12 using five-fold cross-validation before tested it in CASP14.

1. Cheng, J., Choe, M. H., Elofsson, A., Han, K. S., Hou, J., Maghrabi, A. H., ... & Studer, G. (2019). Estimation of model accuracy in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1361-1377.
2. Chen, X., Akhter, N., Guo, Z., Wu, T., Hou, J., Shehu, A. and Cheng, J. 2020. Deep Ranking in Template-free Protein Structure Prediction. The 11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB).
3. Adhikari, B., Hou, J. & Cheng, J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics (2017).

4. Karasikov, M., Pagès, G., & Grudinin, S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics, bty1037, https://doi.org/10.1093/bioinformatics/bty1037
5. Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of molecular biology 376, 288-301 (2008).
6. Rykunov, D. & Fiser, A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67, 559-568 (2007).
7. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one 5, e15386 (2010).
8. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC bioinformatics 17, 495 (2016).
9. Uziela, K. & Wallner, B. ProQ2: Estimation of Model Accuracy Implemented in Rosetta. Bioinformatics, btv767 (2016).
10. Uziela, K., Shu, N., Wallner, B. & Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. Scientific reports 6, 33509 (2016).
11. Shen, M. y. & Sali, A. Statistical potential for assessment and prediction of protein structures. Protein science 15, 2507-2524 (2006).
12. Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry 35, 672-681 (2014).
13. Benkert, P., Tosatto, S. C. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics 71, 261-277 (2008).
14. Wang, Z., Tegge, A. N. & Cheng, J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins: Structure, Function, and Bioinformatics 75 (2009).
15. Wang, Z., Eickholt, J. & Cheng, J. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 27, 1715-1716 (2011).
16. Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science 15, 900-913 (2006).
17. McGuffin, L. J. & Roche, D. B. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 26, 182-188 (2009).
18. Wu, T., Guo, Z., Hou, J., and Cheng, J.. 2020. DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. bioRxiv.
19. Wu, T., Guo, Z., and Cheng, J.. 2019. DNCON4 V1.0 https://github.com/jianlin-cheng/DNCON4_system.

MULTICOM-CONSTRUCT, MULTICOM-CLUSTER, MULTICOM-DIST, MULTICOM-HYBRID, MULTICOM-DEEP (TS)

CASP14 Protein Tertiary Structure Prediction by MULTICOM Server Predictors

Tianqi Wu[1$], Jian Liu[1$], Zhiye Guo[1], Jie Hou[2], J. Cheng[1*]
1-Department of Electrical Engineering and Computer Science, University of Missouri, Columbia; 2-Department of Computer Science, Saint Louis University, St. Louis
[$]equal contribution; *corresponding author (chengji@missouri.edu)

In CASP14, we tested multiple versions of our MULTICOM integrated protein structure prediction system based on distance-based template-free structure modeling and template-based modeling as five tertiary structure prediction servers (MULTICOM-DIST, MULTICOM-HYBRID, MULTICOM-DEEP, MULTICOM-CLUSTER, and MULTICOM-CONSTRUCT). Main improvements made on both template-based and ab initio predictors of the MULTICOM system since CASP13 include (1) new template-free (ab initio) modeling methods empowered by the deep-learning based protein inter-residue distance prediction; (2) a fast, light version of template-based modeling system using deeper sequence alignments to build sequence profiles; and (3) consensus model ranking methods that leverage predicted residue-residue distance information.

Methods

Four servers (MULTICOM-HYBRID, MULTICOM-DEEP, MULTICOM-CLUSTER, and MULTICOM-CONSTRUCT) are rebuilt from the CASP13 MULTICOM system[1], while MULTICOM-DIST is a brand-new, pure ab initio server predictor. In general, they shared a similar protocol composed of the following four parts: (1) full-length model generation using both template-based and template-free modeling or template-free modeling only (i.e. MULTICOM-DIST); (2) domain identification and domain-based model generation and assembly if needed; (3) model evaluation for both full-length and domain-based models; and (4) model combination and refinement.

For the template-based modeling, MULTICOM-HYBRID and MULTICOM-DEEP apply a fast, light version of the MULTICOM template-based prediction pipeline, which mainly uses deep sequence alignments with HHsuite search against several updated template libraries to identify the best hits for the target. As a comparison, MULTICOM-CLUSTER and MULTICOM-CONSTRUCT employs the same template-based modeling pipeline used in CASP13[1] that leverages a number of sequence alignment and fold recognition methods.

For the template-free modeling, several inhouse distance-guided ab initio modeling tools are used (e.g., DFOLD[2] for MULTICOM-DIST, MULTICOM-HYBRID, and MULTICOM-DEEP, and CONFOLD2[3] for MULTICOM-CLUSTER and MULTICOM-CONSTRUCT). Moreover, MULTICOM-DIST, MULTICOM-HYBRID, MULTICOM-DEEP use the inter-residue distance map predicted by DeepDist[4] as input for trRosetta[5] to generate ab initio models, while MULTICOM-CLUSTER and MULTICOM-CONSTRUCT use trRosetta[5] with multiple sequence alignments generated by DNCON4 (*https://github.com/jianlin-cheng/DNCON4_system*) to generate ab initio models.

Different model ranking methods are applied in the final step to select top five models from both full-length and domain-based models. MULTICOM-CLUSTER and MULTICOM-HYBRID models are primarily ranked by the pairwise similarity scores between models. MULTICOM-CONSTRUCT applies DeepRank[6] to select models. MULTICOM-DEEP selects the top models largely based on the average ranking scores of pairwise ranking, SBROD[7], and inter-residue distance scores. The pure ab initio server predictor MULTICOM-DIST mainly uses SBROD[7] for scoring models. The final selected models of MULTICOM-CLUSTER or MULTICO-HYBRID may be combined with other similar models in the pool to generate final models.

Availability

Some source code of several tools of MULTICOM servers is available here: https://github.com/multicom-toolbox.

1. Hou,J., Wu,T., Cao,R., Cheng,J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics. 87, 1165-1178.
2. https://github.com/jianlin-cheng/DFOLD.
3. Adhikari,B., Cheng,J. (2018) CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC bioinformatics. 19:22.
4. Wu,T., Guo,Z., Hou,J., Cheng,J. (2020) DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. bioRxiv. 2020.03.17.995910.
5. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S., Baker,D. (2020) Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences. 117,1496-1503.
6. Cheng,J., Choe,M.H., Elofsson,A., Han,K.S., Hou,J., Maghrabi,AHA., et al. (2019) Estimation of model accuracy in CASP13. Proteins: Structure, Function, and Bioinformatics. 87, 1361-1377.
7. Karasikov,M., Pagѐs,G., Grudinin,S. (2019) Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics. 35, 2801-2808.

MULTICOM-DEEP, MULTICOM-DIST, MULTICOM-HYBRID (QA)

Improving protein single-model and consensus quality assessment using inter-residue distance prediction and deep learning

Jie Hou[1], Zhiye Guo[2], Jian Liu[2] , Tianqi Wu[2] and Jianlin Cheng[2]

1. *Department of Computer Science, Saint Louis University, St. Louis, MO, 63103, USA*
2. *Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA. *corresponding author: chengji@missouri.edu*

Residue-residue contact prediction and deep learning have demonstrated their effectiveness in improving the protein model quality assessment (QA)[1]. The deep learning techniques have significantly enhanced inter-residue contact and distance prediction[2]. Moreover, deep learning network has shown the great potentials for effectively integrating multiple complementary QA metrics as well as the structural constraints derived from contact predictions. In CASP14, we introduced the improved deep learning *consensus* QA method (MULTICOM-HYBRID) and two new *single-model* QA methods (MULTICOM-DIST and MULTICOM-DEEP) that aim to fully utilize inter-residue distance constraints for predicting the global quality of stage1 and stage2 models of CASP14 targets.

Methods

Given a pool of structural models, our methods firstly analyze the structural discrepancies between the distance map parsed from each structural model and the distance map predicted from the protein sequence. The full-length real-value distance map from the target protein sequence and its high-quality predicted contact map are generated by our latest distance map predictor (DeepDist)[2]. We adopt several image similarity metrics used in the field of computer vision to evaluate the consistency of the structural patterns between the distance map calculated from the model and the distance map predicted from the protein sequence. The major distance-based metrics include: GIST Descriptor[3], Oriented FAST and Rotated BRIEF (ORB)[4], PHASH[5], PSNR & SSIM[6], Pearson correlation (PCC), and root mean square error (RMSE). In addition, the percentage of predicted contacts (i.e., short-range, medium-range, and long-range contacts) existing in a model of the target are also used as features by converting the real-value distance map to binary contact predictions at 8 Å threshold, along with the precisions of top 2L contact predictions and recall of top L contact predictions (L: sequence length). All these distance/contact-based features for an input model are integrated with other model quality metrics using a *deep neural network* to make a final quality prediction. The other input features include energy scores from 9 single-model methods (i.e., SBROD[7], OPUS_PSP[8], RF_CB_SRS_OD[9], Rwplus[10], DeepQA[11], ProQ2[12], ProQ3[13], Dope[14] and Voronota[15]) and three multi-model QA methods (i.e., APOLLO[16], Pcons[17], and ModFOLDclust2[18]).

The deep neural network was trained on the models of CASP8-12 experiments. 10 trained deep neural networks were obtained from 10-fold cross-validation. All input features of each model are fed into the 10 trained networks to generate 10 quality scores. Next, the 10 predicted quality scores and the initial input features are used together by another deep neural network to predict the final quality score. Among the three QA methods, MULTICOM-HYBRID include all the features above as input to predict the final quality score, which leads to a *consensus* method. MULTICOM-DEEP excludes three multi-model QA methods (i.e. APOLLO, Pcons and ModFOLDclust2) from the input features to perform the *single-model* quality assessment. MULTICOM-DIST simply integrates the *single-model* distance-based features and 6 *single-model* energy-based features (i.e., SBROD[7], OPUS_PSP[8], RF_CB_SRS_OD[9], Rwplus[10], Dope[14] and Voronota[15]) for single-model quality assessment. Prior to the CASP14 experiment, the three methods were benchmarked on the CASP13 dataset and showed a significant improvement over the individual QA methods used to generate input features.

1. Hou, J., Wu, T., Cao, R. & Cheng, J. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics 87, 1165-1178 (2019).
2. Wu, T., Guo, Z., Hou, J. & Cheng, J. DeepDist: real-value inter-residue distance prediction with deep residual network. bioRxiv (2020).
3. Oliva, A. & Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42, 145-175 (2001).
4. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. in 2011 International conference on computer vision. 2564-2571 (Ieee).
5. Kozat, S. S., Venkatesan, R. & Mihçak, M. K. in 2004 International Conference on Image Processing, 2004. ICIP'04. 3443-3446 (IEEE).
6. Hore, A. & Ziou, D. in 2010 20th international conference on pattern recognition. 2366-2369 (IEEE).
7. Karasikov, M., Pagès, G. & Grudinin, S. Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics 35, 2801-2808 (2019).
8. Lu, M., Dousis, A. D. & Ma, J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of molecular biology 376, 288-301 (2008).
9. Rykunov, D. & Fiser, A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance ‐ dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics 67, 559-568 (2007).
10. Zhang, J. & Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one 5, e15386 (2010).
11. Cao, R., Bhattacharya, D., Hou, J. & Cheng, J. DeepQA: improving the estimation of single protein model quality with deep belief networks. BMC bioinformatics 17, 495 (2016).
12. Uziela, K. & Wallner, B. ProQ2: Estimation of Model Accuracy Implemented in Rosetta. Bioinformatics, btv767 (2016).
13. Uziela, K., Shu, N., Wallner, B. & Elofsson, A. ProQ3: Improved model quality assessments using Rosetta energy terms. Scientific reports 6, 33509 (2016).
14. Shen, M. y. & Sali, A. Statistical potential for assessment and prediction of protein structures. Protein science 15, 2507-2524 (2006).
15. Olechnovič, K. & Venclovas, Č. Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry 35, 672-681 (2014).
16. Wang, Z., Eickholt, J. & Cheng, J. APOLLO: a quality assessment service for single and multiple protein models. Bioinformatics 27, 1715-1716 (2011).
17. Wallner, B. & Elofsson, A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Science 15, 900-913 (2006).

18. McGuffin, L. J. & Roche, D. B. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics 26, 182-188 (2009).

MULTICOM, MULTICOM-DEEP, MULTICOM-CONSTRUCT, MULTICOM-DIST, MULTICOM-AI, MULTICOM-HYBRID, MULTICOM-CLUSTER (RR)

Prediction of protein inter-residue distance and contacts with deep learning

Zhiye Guo[1], Tianqi Wu[1], Jian Liu[1], Jie Hou[2], Jianlin Cheng[1*]
1-Department of Electrical Engineering and Computer Science, University of Missouri, Columbia; 2-Department of Computer Science, Saint Louis University, St. Louis
* chengji@missouri.edu

In CASP14, we tested several residue-residue distance predictors and one contact predictor based on different deep learning models trained on residue-residue co-evolution features and several other sequence and structural features.

Methods

We use four sets of features with deep neural networks. Three of four feature sets are mostly coevolution-based features, i.e. covariance matrix (COV)[1], pseudolikelihood maximization matrix (PLM)[2], and precision matrix (PRE)[3] calculated from multiple sequence alignments. And one set of features contains non-coevolution sequence-based features (OTHER)[4] in case multiple sequence alignments are shallow. The OTHER feature set has the sequence profile generated by PSI-BLAST[5], solvent accessibility from PSIPRED[6] and so on. The architecture of the 2D deep network[7] is shown in **Fig.1**. Different input feature sets have different input sizes. The dimension of COV, PLM, PRE and OTHER is L*L*483, L*L*482, L*L*484 and L*L*47 respectively (L: sequence length). The inputs are fed to an instance normalization layer[8], followed by one convolutional layer and one Maxout layer[9]. The output of the Maxout layer is fed into 20 residual blocks. Each residual block contains three RCIN blocks which are composed of instance normalization layer, row normalization layer, column normalization layer[10], five convolutional layers with 64 filters and kernel size are 1*1, 3*3, 7*1, 1*7, 1*1 respectively, one squeeze-and-excitation block[11] and one dropout layer with a dropout rate at 0.2. After the last residual block employing a convolutional layer followed by instance normalization layer, the softmax activation function is used to predict the distance distribution.

Even though all seven distance predictors share the similar network architectures, they differ in distance interval labels used to train them, how the prediction output is produced, and how input multiple sequence alignments (MSAs) are generated. The distance intervals of MULTICOM-CONSTRUCT are 0 to 4 Å, 4 to 6 Å, 6 to 8 Å, …, 18 to 20 Å and > 20 Å. We discretize inter-residue distance into 42 bins for MULTICOM-DIST, i.e. dividing 2 to 22 Å into 40 bins with bin size 0.5 Å, plus a 0 - 2 Å interval and a > 22 Å interval. The distance range (0 to 20 Å) of MULTICOM-AI is binned into 37 equally spaced interval of 0.5 Å, plus one > 20 Å interval. MULTICOM-HYBRID shares a similar segmentation strategy with MULTICOM-DIST, but the difference is that it starts with an interval 0 - 3.5 Å, and its last interval is set to > 19 Å. All predictions are converted into the official intervals of CASP14. The predictions of MULTICOM-DEEP and MULTICOM are averaged from the outputs of all the other servers above. Unlike the multi-interval distance predictors, MULTICOM-CLUSTER is a binary contact map predictor. Furthermore, different alignment methods are used by the predictors to generate input MSAs, including DeepMSA[12], our in-house tool DeepAln and one approach that uses HHblits[13] search against BFD[14] database.

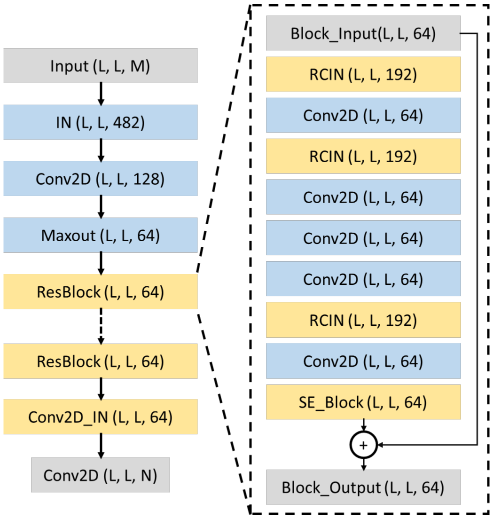


Fig 1. Deep learning network architecture for protein inter-residue distance prediction.

Availability

The source code is available at https://github.com/multicom-toolbox/deepdist .

1. Jones, D. T.; Kandathil, S. M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018;34(19):3308-3315.
2. Seemayer, S.; Gruber, M.; Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
3. Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 2019;35(22):4647-4655.
4. Adhikari, B.; Hou, J.; Cheng, J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 2018;34(9):1466-1472.
5. Bhagwat, M.; Aravind, L., Psi-blast tutorial. In Comparative genomics, Springer: 2007; pp 177-186.
6. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology 1999;292(2):195-202.
7. Wu, T.; Guo, Z.; Hou, J.; Cheng, J. DeepDist: real-value inter-residue distance prediction with deep residual network. bioRxiv 2020.
8. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 2016.
9. Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. Maxout networks. arXiv preprint arXiv:1302.4389 2013.
10. Mao, W.; Ding, W.; Xing, Y.; Gong, H. AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. Nature Machine Intelligence 2019:1-9.

11. Hu, J.; Shen, L.; Sun, G. In Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp 7132-7141.
12. Zhang, C.; Zheng, W.; Mortuza, S.; Li, Y.; Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2019.
13. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 2012;9(2):173.
14. Steinegger, M.; Mirdita, M.; Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nature methods 2019;16(7):603-606.

netris

netris: Contact Prediction through Network Inference Methods

I. Johansson-Åkhe[1]

[1] - *Linköping University*

isak.johansson-akhe@liu.se

Key: *Auto:Y; CASP_serv:N; Templ: N; Fragm: N; Cont: Y; Dist: N; Tors: N; DeepL: N; EMA: N; MD: N*

The bioinformatics field of network inference has classically outpaced the structural bioinformatics field with regards to implementing mathematical concepts. For instance, efficient filtering of indirect correlation effects was introduced to structural bioinformatics in 2010[1] and popularised in CASP10 (2012), but ARACNE[2] used advanced indirect correlation filtering for gene regulatory network inference already in 2006.

However, in recent years the structure prediction field has seen a considerable rise in performance through the use of machine learning, while the network inference field still relies on methods using information theory and manual correlation definitions, with only minor increases in predictive ability[3].

Can network inference methods developed for another field which has classically adopted mathematical concepts earlier compete with the carefully engineered and machine-learning based methods used in structural bioinformatics today?

Netris is a consensus-method combining several top-performing methods for gene regulatory network inference and adapting them to the protein contact matrix prediction problem with minimal changes.

Methods

Netris uses averaged predictions between three common gene regulatory network inference methods: ARACNE[2], pearson correlation coefficient, and clr[4], all implemented through the comhub[5] package. Throughout the development of netris, other new network inference methods were also considered, such as GENIE3[6] or TIGRESS[7]. The combination and selection of methods was optimised on contact prediction performance on CASP11-12 data.

Briefly, netris uses the aforementioned method with minor changes to facilitate using multiple sequence alignments as input rather than the GWAS input these methods normally expect. For ARACNE and clr, netris uses an MSA-compatible method of calculating the mutual information matrix, which is then directly fed to the algorithms as substitute for the regular GWAS-based mutual information matrix. For pcc, a short auto-encoder reduces the MSA to numerical values.

Contacts to residues present in few sequences in the MSA are disregarded. The resulting interaction-networks between residues are averaged by edge-confidence.

Results

On a subset of published data from CASP13, netris achieved a top L/5 interactions long range contact precision of 21.5%. While comparable to top predictors in CASP11 and earlier, from the early days of efficient filtering of indirect contacts before the paradigm of machine learning, it is still far from the performance of modern contact prediction methods. Although not close to the state-of-the-art, netris will serve as a baseline for comparison to the un-optimized but powerful network inference methods of old.

Availability

The method is not publicly available.

1. Burger, L., & Van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS computational biology, 6(1), e1000633.
2. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006, March). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In BMC bioinformatics (Vol. 7, No. S1, p. S7). BioMed Central.
3. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., & Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 17(2), 147-154.
4. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biol, 5(1), e8.
5. Åkesson, J., Lubovac-Pilav, Z., Magnusson, R., & Gustafsson, M. (2019). ComHub: Community predictions of hubs in gene regulatory networks. bioRxiv, 840959.
6. Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PloS one, 5(9), e12776.
7. Haury, A. C., Mordelet, F., Vera-Licona, P., & Vert, J. P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. BMC systems biology, 6(1), 145.

NOVA: protein structure prediction with structural information and deep learning-based folding framework

Tong Wang[1], Bin Shao[1], Siyuan Liu[1*], Qijiang Xu[1*], Wenze Ding[1*], Jianwei Zhu[1] and Tie-Yan Liu[1]

1-Microsoft Research Asia, Beijing, China

**Work performed as an intern in Microsoft Research Asia*

watong@microsoft.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:Y.v; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N.*

NOVA is a de novo protein structure prediction pipeline which consists of three components: a protein property prediction module, a novel deep learning-based folding framework, and a proteinspecific quality ranking module. First, multiple protein properties are predicted from sequence information via multiple sequence alignment (MSA) and structural information by high-quality fragments. Second, a novel deep learning-based protein folding framework which takes predicted protein properties as input and automatically adjusts the input constraints in a self-adaptive way (Ding et al, in preparation). Finally, NOVA integrates all protein structures predicted by our folding framework as well as those from our server predictor, TOWER and scores all structures for final submission by our protein-specific quality ranking module.

Methods

Different from most protein property predictors relying on sequence information from MSA, our protein property prediction module fully utilizes structural information by high-quality fragment libraries derived from resolved structures in Protein Data Bank (PDB). Fragment libraries exploit near-native template fragments as much as possible for each position of the target protein, which can provide rich structural information that MSA cannot capture. In NOVA, we adopted DeepFragLib[1, 2] which achieved the best performance among all fragment library construction algorithms when benchmark tested on recent CASPs. Using fragment library for fragment assembly based protein structure prediction is generally computationally intensive; thus, in NOVA we proposed an algorithm that utilizes the structural information of fragment libraries for protein structure prediction much more efficiently (Liu et al, in preparation). First, protein properties such as torsion angles, backbone angles, inter-residue distances between Cα and Cβ atoms within fragments were extracted from each fragment of the fragment library constructed by DeepFragLib. Then all these properties were gathered for a position of the target protein to form property distributions, where each distribution represents a kind of property for this position. Third, to utilize the structural information for gradient-based protein folding in a differentiable way, these distributions were parameterized with a set of weighted Gaussian Mixture Models (wGMM). Finally, the wGMM models were adopted into our folding framework to work as protein-specific potentials by negative log likelihood functions. In addition, sequence information derived from

MSA was also used to predict inter-residue distances of Cβ atoms and orientations, and then fed into folding process via cubic spline functions.

A novel deep learning-based protein folding framework was developed to fold proteins from constraints in a self-adaptive way. Direct optimization folding process generally takes protein constraints as input and folds protein structures via gradient descent. Both properties derived from sequence information such as inter-residue distances and orientations and those derived from structural information such as distributions extracted from fragment libraries provide rich and sometimes conflicting constraints for gradient-based protein folding. However, conflicts and redundancy among these constraints predicted from different sources are inevitable, which can trap the optimization process into poor local minima and thus damages the accuracy of the final model. In our protein folding framework, an automatic constraint optimization system was proposed to make full use of all constraints in a self-adaptive way. The constraints were first analyzed and scored by a deep neural network and the scores were then gradually adjusted according to the output signals during the folding process. Furthermore, the protein folding framework adopts an ensemble of deep neural networks with sophisticated loss functions, which makes accurate ranking of the predicted structures.

Finally, the highly ranked protein structures as well as those generated by our server predictor TOWER were refined using Rosetta FastRelax[3] and then re-ranked by a protein-specific quality analysis module. This module consists of an ensemble of quality analysis software packages including ProQ3D[4], ProQ4[5], and a newly designed model which describes structural similarity between proteins. A linear regression model was trained to assign a weight for each score given by the quality analysis software. The protein structures with highest scores were then picked up for final submission.

1. Wang, T., Qiao, Y., Ding, W., Mao, W., Zhou, Y., Gong, H. 2019. Improved fragment sampling for ab initio protein structure prediction using deep neural networks. Nature Machine Intelligence. 1: 347-355.
2. Wei, G.-W. 2019. Protein structure prediction beyond AlphaFold. Nature Machine Intelligence. 1: 336-337.
3. Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D. 2004. Protein structure prediction using Rosetta, pp. 66-93. Methods in enzymology, Ed. Elsevier,
4. Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B., Elofsson, A. 2017. ProQ3D: improved model quality assessments using deep learning. Bioinformatics. 33: 1578-1580.
5. Hurtado, D.M., Uziela, K., Elofsson, A. 2018. Deep transfer learning in the assessment of the quality of protein models. arXiv preprint arXiv:1804.06281.

Ornate

Protein model quality assessment using 3D oriented convolutional neural network Ornate

G. Pagès[1], B. Charmettant[1], D. Zhemchuzhnikov[1], and S. Grudinin[1]

[1] - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
Sergei.Grudinin@inria.fr

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Protein model quality assessment (QA) is a crucial and yet open problem in structural bioinformatics. The current best methods for single-model QA typically combine results from different approaches, each based on different input features, both structure-based and sequence-based, constructed by experts in the field. Then, the prediction models are trained using machine-learning algorithms. Recently, with the development of convolutional neural networks (CNN), the training paradigm has been changed. In computer vision, the expert-developed features have been significantly overpassed by automatically trained convolutional filters. This motivated us to apply a three-dimensional (3D) CNN to the problem of protein model QA and to develop a novel QA method called Ornate[1].

Methods

Ornate (Oriented Routed Neural network with Automatic Typing) is a recent method for single-model QA. Ornate comprises a residue-wise scoring function that takes as input 3D density maps around each of the protein residues. It predicts the local (residue-wise) and the global model quality through a deep 3D CNN. Specifically, the Ornate method aligns the input density maps, constructed from each residue and its neighborhood, with the backbone topology of the corresponding residue. This circumvents the problem of ambiguous orientations of the initial models[2]. Also, Ornate includes automatic identification of protein atom types.

The input of the network is constituted of 167 density maps, each consisting of 24×24×24 voxels with a 0.8 Å side. Each map represents the density of one type of atoms among the 167 that can be found in proteins. Such a representation is very sparse. To make it dense and to reduce the number of model variables, we linearly project the 167 types into a 15-dimensional space. We wanted to be as rigorous as possible on making the assumptions about classifying the atoms. Therefore, we let the network to learn the projection automatically upon training by designing a "retyper" projection layer. This is followed by three 3D convolutional layers that learn structural features on different scales. Then, two last fully connected layers process the features from the previous layers, select their best combination, and output a scalar prediction corresponding to the local model quality. We trained the method on structures from the previous CASP experiments using local CAD-scores[3] of each residue as the ground truth.

Results

We used Ornate for the first time during CASP13, in a semi-automatic fashion. In CASP14, we applied Ornate to the QA category of targets using a fully automated server. The method is rather robust, but it requires an extensive amount of training data and computational resources. This motivated us to develop a new generation of CNN-based methods, that are faster to train using the current amount of structural data and computational resources. These are our new models VoroCNN and S-GCN, described elsewhere[4,5].

Availability

Ornate is made publicly available on our website at https://team.inria.fr/nano-d/software/Ornate/.

1. Pagès,G., Charmettant,B., & Grudinin,S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics, 35(18), 3313-3319.
2. Derevyanko,G., Grudinin,S., Bengio,Y., & Lamoureux,G. (2018). Deep convolutional networks for quality assessment of protein folds. Bioinformatics, 34(23), 4046-4053.
3. Olechnovič,K., Kulberkytė,E., & Venclovas,Č. (2013). CAD - score: a new contact area difference - based function for evaluation of protein structural models. Proteins: Struct., Funct., Bioinf., 81, 149-162.
4. Igashov,I., Olechnovič,K., Kadukova,M., Venclovas,Č., & Grudinin,S. (2020). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. bioRxiv.
5. Igashov,I., Pavlichenko,N., & Grudinin,S. (2020). Spherical convolutions on molecular graphs for protein model quality assessment. Submitted.

PBuild

A Rust-based Protein Tertiary Structure Builder

T. Oda
odat1248@gmail.com

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:N; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

For CASP13, I made a Java application named PepBuilderJ. However, I had some difficulty implementing the required functions such as atom-level scoring functions in Java, mostly due to its inadequate performance. Therefore, I used Rust, a much more robust and optimal language, to develop a new application for CASP14.

Methods

The multiple sequence alignments (MSAs) and Hidden Markov Models (HMMs) of the prediction targets were constructed as follows; first, the blastp[1] searches against UniRef90[2] were performed. Next, jackhmmer[3] was used to build intermediate-MSAs from blastp results. Finally, the intermediate-MSAs were processed by hhblits[4] with the UniRef30_2020_01 database, which then produced the final-MSAs and HMMs.

For template base modelling (TBM), the template candidates were obtained by hhblits search against the profile database which was used in CASP13 (hmm_casp13) and by hmmsearch[3] against a recent PDB database[5] (https://www.rcsb.org/). Two HMM profiles were constructed for every template candidate by hhblits using hmm_casp13 and UniRef30_2020_01. The HMM-HMM alignments between targets and template candidates were made by hhalign[4]. The templates and alignments used in the next step were manually selected based on the scores, length of aligned regions, and variations of templates.

Backbone atoms were assigned according to the corresponding template residues, and side-chains were constructed using the rotamer library used in CASP13. Simulated annealing was performed to build the gap regions (nearly whole regions for (possible) free modelling (FM) targets; e.g. targets where good templates and alignments couldn't be found) using the phi-psi angles predicted by the customized version of ProSPr[6], which is then scored using a scoring function built from several scores and parameters as follows; inter-residue contact scores produced by plmDCA[7;8] (https://github.com/pagnani/PlmDCA), a number of energy terms used in CHARMM19[9], a number of energy terms used in EvoEF2[10], and phi-psi angle scores produced by the customized ProSPr. The predicted inter-residue distances were not used to avoid the potential infringement of the patent (https://patents.google.com/patent/WO2020058176A1). To get the ProSPr and plmDCA results, intermediate-MSA and final-MSA were merged into one MSA (merged-MSA). The merged-MSAs were processed by plmDCA to obtain the scores and ProSPr input files, by hhmake[4] to obtain HMMs, and by PSI-BLASTexB[11] to obtain PSSMs.

The refinements of the structures were done using the scoring function mentioned above. Each model was constructed using 1 CPU core for several hours up to a few days. The submitted models were manually selected based on the scores produced by the scoring function, the visualization by PyMOL(https://sourceforge.net/projects/pymol/files/Legacy/), and variations of templates.

The protocol was fixed in the middle of the season and was not applied to the early targets.

Results

While it appears that I could detect several templates for TBM targets, the alignments were not sufficient as I could see some gaps/insertions in the alpha helix and beta sheet regions of the templates. In the case of FM targets, the models were rarely folded into meaningful structures. It's conclusive that revising alignments for TBM targets and more effective structure sampling algorithms for FM targets are needed.

Conflict of Interest

The author is an employee of Lifematics Inc. This work was done by the author using his private time.

Availability

The customized version of ProSPr is available from https://github.com/yamule/prospr.

1. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402 (1997).
2. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926-932, doi:10.1093/bioinformatics/btu739 (2015).
3. Potter, S. C. et al. HMMER web server: 2018 update. Nucleic Acids Res 46, W200-W204, doi:10.1093/nar/gky448 (2018).
4. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics 20, 473, doi:10.1186/s12859-019-3019-7 (2019).
5. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res 28, 235-242 (2000).
6. Billings, W., Hedelius, B., Millecam, T., Wingate, D. & Della Corte, D. ProSPr: Protein Structure Prediction via Interatomic Distances. Bulletin of the American Physical Society 65 (2020).
7. Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. Journal of Computational Physics 276, 341-356 (2014).
8. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87, 012707, doi:10.1103/PhysRevE.87.012707 (2013).
9. Brooks, B. R. et al. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. Journal of computational chemistry 4, 187-217 (1983).
10. Huang, X., Pearce, R. & Zhang, Y. EvoEF2: accurate and fast energy function for computational protein design. Bioinformatics 36, 1135-1142, doi:10.1093/bioinformatics/btz740 (2020).
11. Oda, T., Lim, K. & Tomii, K. Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. BMC Bioinformatics 18, 288, doi:10.1186/s12859-017-1686-9 (2017).

PerezLab_Gators

MELDxMD: incorporating ML derived distograms and heuristics into simulations

A. Mondal[1], A. Perez[1]
[1] - *University of Florida*
perez@chem.ufl.edu

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:Y*

We have previously established the value of MELDxMD as a Bayesian inference approach to incorporate information into Molecular Dynamics simulations, determining high accuracy structures for several targets. Previously, the largest success has come from using NMR data, and using heuristics on small targets. Here we try to break the size barrier for MELDxMD in the absence of NMR data by using distograms derived from combining sequence co-evolution with machine learning.
In the current CASP 14 event, we have submitted 99 predictions, in the TS, TR and data driven categories. All simulations were performed with our local resources at the HiperGator Supercomputer center at the University of Florida.

Methods

MELD (Modeling Employing Limited Data) uses Bayesian inference to accelerate molecular dynamics with data[1]. The data we use can be mapped to distances between pairs of atoms, which we impose as flat-bottom harmonic restraints. There is no energy penalty when the data is satisfied, and increases quadratically (and then linearly after a cutoff) otherwise. The data has the peculiarity that some of the data might not be correct, we generally know what per cent to trust, but not which data to trust. The Bayesian aspect of the method comes from inferring which subset of the data is most compatible with the protein given a prior (given by the molecular dynamics force field). All simulations were run for at least a microsecond using 30 replicas, the GBneck2 implicit solvent model[2] and the ff14SB force field[3] for side chains with ff99SB[4] for the backbone.

Data used in simulations: We use secondary structure predictions from psipred[5] and enforce them at 60% accuracy (90% in the case of membrane proteins). We only keep information about helix and strand residues. For small proteins (under 110 residues) we used our Coarse Physical Insights based on hydrophobic packing and strand pairing (see ref [6]).

We used the hhssuite and the strategy derived by the Baker group[7] to feed multiple sequence alignments to trRosetta. When the alignments were good, we used their trained neural network to produce distograms of residue-residue distances (we did not use the orientation information). We analyzed the distograms and selected those that had 80% of probability within an 8Å window (e.g. those distributions that were narrow, no matter what the average of the distribution was. We then imposed this as MELD restraints centered at the peak of the distribution, with a 7Å flat-bottom region. We trusted 70% of the distograms.

We select the 5 structures to submit by performing hierarchical clustering on the lowest 5 temperature replicas with an epsilon value of 2. We use the RMSF fluctuations in each cluster as a measure of the error on the precision of each amino acid.

Experimental data: For the SAX data we used a combination of MDFF and MELD to fit the structures within the SAX derived envelope[8]. For the two NMR targets we used our previous protocols[9,10], changing the H,T-REMD ladder with respect to previous attempts to favor identification of the native state. We seeded NMR runs with our previous TS predictions and performed runs combining NMR and distograms as well as NMR data by itself. We selected the 5 structures to submit based on agreement with the NMR data. We trust 60% of the NMR data.

Results

Our main focus was on NMR targets, which in this edition only accounted for 2 targets. However, both targets were membrane proteins — which result in a different set of challenges from our previous Casp11 and casp13 experiences. We were not able to use an implicit membrane solvent model, hence the force field struggled to stabilize structures in which the hydrophobic residues were facing implicit water and the inside of the protein was mostly polar. For target 1088 we noticed that this deficiencies in our solvent model tended to collapse the beta barrel structure we predicted. Refinement of the models with proper membranes might lead to recovery of the correct structures.

Availability

OpenMM, AmberTools, MELD and the MELD-OpenMM plugin are all available and free to use. Our MELD frontend can be accessed at: git@github.com:maccallumlab/meld.git, and the MELD-openMM plugin can be accessed at: git@github.com:maccallumlab/meld-openmm-plugin.git.

1. J.L. MacCallum, A. Perez, K. Dill, Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference, Proc. Natl. Acad. Sci. U. S. a. 112 (2015) 6985–6990.
2. Nguyen, H., Roe, D., Simmerling, C. (2013). Improved Generalized Born Solvent Model Parameters for Protein Simulations. Journal of chemical theory and computation 9(4), 2020 - 2034.
3. Maier, J., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K., Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. Journal of chemical theory and computation 11(8), 3696 - 3713.
4. Okur, A., Wickstrom, L., Layten, M., Geney, R., Song, K., Hornak, V., Simmerling, C. (2006). Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/ Implicit Solvation Model Journal of Chemical Theory and Computation 2(2), 420-433
5. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195-202.
6. Perez, A., MacCallum, J., Dill, K. (2015). Accelerating molecular simulations of proteins using Bayesian inference on weak information. PNAS 112(38), 11846-51.
7. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. PNAS 117(3), 1496-1503.
8. Mrinal Shekhar, et al. CryoFold: Ab-initio structure determination from electron density maps using molecular dynamics. bioRxiv 687087
9. Blind proteins structure prediction using accelerated free-energy simulations. Science Advances 2(11), e1601274 - e1601274.
10. Robertson, J., Nassar, R., Liu, C., Brini, E., Dill, K., Perez, A. (2019). NMR ‐ assisted protein structure prediction with MELDxMD Proteins 36(1), D402.

PerillaGroup

High throughput structural refinement for large-scale MD simulations

Chaoyi Xu[1], Nidhi Katyal[1] and Juan R. Perilla[1]
[1] *Department of Chemistry and Biochemistry, University of Delaware*
jperilla@udel.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:Y*

The refinement category in Critical Assessment of Structure Prediction (CASP) is a competition for improving the structural quality of selected protein models for which the experimental structures are not yet available. It has encouraged the development of many innovative strategies for protein refinements. Successful refinement strategies from previous CASP rounds[1,2] used physics-based force fields as scoring functions, utilized effective sampling techniques like molecular dynamics simulations and incorporated appropriate restraints to prevent structural deterioration during refinement. By incorporating all these measures, we present here a novel protein structure refinement protocol used in CASP14.

Methods

A schematic overview of our CASP14 refinement protocol is shown in Figure 1. The protocol consists of four steps. The first step includes optimization of the geometric outliers in the initial starting structure using High throughPut structure REFinement (HPREF)[3] developed in our group. HPREF identifies Molprobity outliers[4-7] in the structure and refine them in Cartesian space using Rosetta[8,9], while the rest of the structure is kept fixed.

After the local refinement, the resulting structure is subject to simulated annealing (SA) using NAMD2.14[10]. Before running SA, the protonation states of the titratable residues in the refined structure are determined using propka[11]. Subsequently, the model is solvated by TIP3P water molecules[12], neutralized and the bulk NaCl concentration is set to 0.15 mM. The prepared system is then minimized in two-stages using a conjugate gradient[13] and line search algorithm[14]. In the first stage, the backbone atoms of the protein are fixed and then restrained in the second stage. Thereafter, the minimized system is heated from 50 K to 310 K in 20 K increments for 1 ns and equilibrated for 5 ns at 310 K. The backbone atoms of the model are restrained by a force constant of 1.0 Kcal/mol Å^2 during the heating and equilibration steps. After equilibration, an iterative refinement step similar to SA is employed to sample the conformational space of the model approximate to the native state. A weaker force constant of 0.25 Kcal/mol Å^2 is applied to prevent structural deterioration. Each cycle of the iterative SA simulation comprises three steps, a heating step, a cooling step, and a minimization step. The temperature of the system increments at a step of 50 K and 10 K in the heating and cooling step, respectively. At each temperature, the system is sampled for 0.5 ns. The temperature range of the SA is 350 K to 150 K for the model with a GDT_HA score larger than 50 and 450 K to 150 K for those having a GDT_HA score less than 50. After cooling, the model is then further minimized for 1,000 steps. The resulting model is saved and used as the initial structure for the next iteration. The SA refinement step is conducted with NAMD2.14[10]. The temperature and pressure during the simulation is controlled by the stochastic rescaling thermostat[15] and the Nosé-Hoover Langevin-piston pressure control, respectively. Long range electrostatic force calculations used the particle mesh Ewald method, with a 1.2 nm cutoff. An integration time step of 1 fs was utilized, with the non-bonded interactions evaluated every 2 fs and electrostatics updates every 4 fs. CHARMM36m protein[16] force field is employed in all molecular dynamics (MD) simulations here.

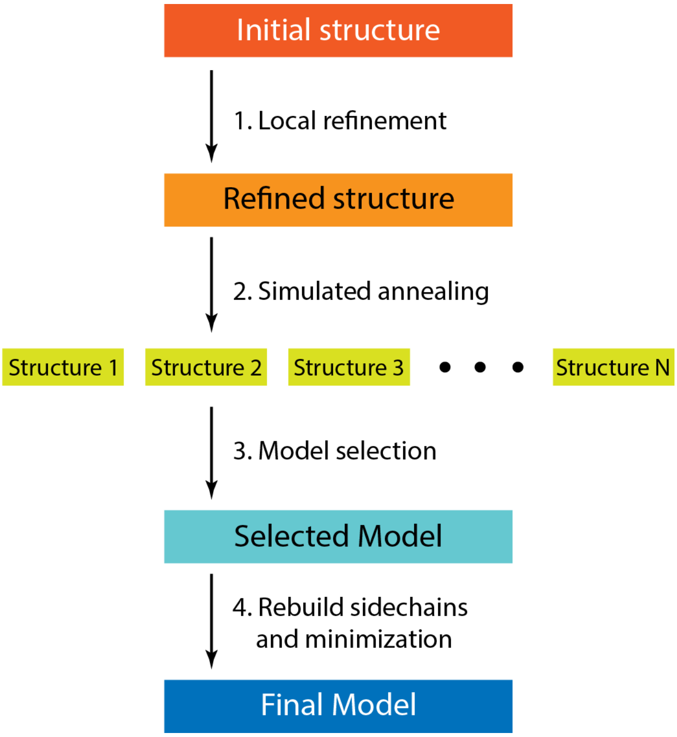


Figure 1. CASP14 refinement protocol used in Perilla Group. Starting model is firstly refined using a local refinement tool developed by our group to optimize the Molprobity outliers in the initial structure. Subsequently, the refined model is subject to an iterative simulated annealing molecular dynamics simulation using NAMD2.14[10]. During the simulation, the backbone atoms of the structure are restrained by a force constant of 0.25 Kcal/mol Å^2. From the MD simulated annealing step, over 120 of structures are generated. Among them, the model with lowest Rosetta energy is selected. The sidechain atoms of selected model are rebuild using SCWRL4[17] and then further minimized in NAMD2.14[10]. At last, the resulting model is validated by GDT_TS, GDT_HA and Molprobity before submitted to CASP14.

After the MD sampling step, at least 120 models are generated for each input structures. These models are then sorted by the Rosetta ref2015 scoring function. The model with lowest Rosetta energy is selected for the next step. The sidechain atoms of the selected model are removed and rebuilt using SCWRL4[17]. After that, the resulting model is then minimized in explicit solvent in NAMD2.14 for 1,000 steps. The model after minimization is referred as “Model 1”. Before submission, Model 1 for each input structure is confirmed by GDT_TS, GDT_HA, and Molprobity scores. All the targets were refined with the same protocol without manual intervention.

Availability

The codes used in the refinement protocol are available at: https://github.com/Perilla-lab/hpRefStruct.

1. Hovan, L., et al., Assessment of the model refinement category in CASP12. Proteins: Structure, Function, and Bioinformatics, 2018. 86(S1): p. 152-167.
2. Read, R. J., et al. Evaluation of model refinement in CASP13. Proteins: Structure, Function, and Bioinformatics, 2019.87(12): p. 1249-1262.
3. Xu, C.Y., et al., High throughput structural refinement for large-scale MD simulations. In preparation.
4. Davis, I.W., et al., MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Research, 2004. 32(suppl_2): p. W615-W619.
5. Davis, I.W., et al., MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Research, 2007. 35(suppl_2): p. W375-W383.
6. Chen, V.B., et al., MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D, 2010. 66(1): p. 12-21.
7. Williams, C.J., et al., MolProbity: More and better reference data for improved all-atom structure validation. Protein Science, 2018. 27(1): p. 293-315.
8. Das, R., et al., Macromolecular Modeling with Rosetta. Annual Review of Biochemistry, 2008. 77(1): p. 363-382.
9. Bender, B.J., et al., Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. Biochemistry, 2016. 55(34): p. 4748-4763.
10. Phillips, J.C., et al., Scalable molecular dynamics with NAMD. J Comput Chem, 2005. 26(16): p. 1781-802.
11. Dolinsky, T.J., et al., PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res, 2007. 35: p. W522-5.
12. Jorgensen, W.L., et al., Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. Journal of Computational Chemistry, 1998. 19(10): p. 1179-1186.
13. Fletcher, R., et al., Function Minimization by Conjugate Gradients. Computer Journal, 1964. 7(2): p. 149-154.
14. Sun, W., et al., Optimization theory and methods: nonlinear programming. Vol. 1. 2006: Springer Science & Business Media.
15. Bussi, G., et al., Canonical sampling through velocity rescaling. The Journal of Chemical Physics, 2007. 126(1): p. 014101.
16. Huang, J., et al., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods, 2017. 14(1): p. 71-73.
17. Krivov, G.G., et al., Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function, and Bioinformatics, 2009. 77(4): p. 778-795.

Pharmulator Structure Prediction Module: Angle-Based Structure Prediction using Bidirectional LSTM and Random Omega Generation

Manuel Collazo[1], Byeongcheol Jo[1], Minjun Jeon[1], Suhyun Park[2], Sang Won Rhee[2], Hoon Dong Kim[3], Sungmin Ahn[1], Sukho Jung[1], Sung Jong Lee[4], Sangwook Wu*[1,2]

[1] - PharmCADD, Busan, 48060, Republic of Korea,
[2] - Department of Physics, Pukyong National University, Busan, 48513, Republic of Korea,
[3] - SK Telecom, Seoul, 100-999, Republic of Korea,
[4] - Research Institute for Basic Sciences, Changwon National University, Changwon, 51140, Republic of Korea
*s.wu@pharmcadd.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

The prediction of a native fold for an unknown protein chain remains challenging, especially if the amino acid sequence is particularly unique. The methodology presented here tries to break ground in the prediction of distinct protein folds through the application of deep learning and statistics based on decades of accumulated knowledge and theory centered around the protein folding process.

Methods

Processes described below harness the capabilities of Artificial Intelligence (AI) alongside statistical physics to achieve viable predictions for native protein folds. A sequence based deep neural network architecture was trained on data implicitly extracted from residues and their respective ordering. This network of sequential procedures achieved the goal of simulating the folding process as it would take place under protein folding theory.

Data Preparation: Datasets were developed from parsing 150,000+ protein structure files downloaded from Protein Data Bank (PDB). An internally developed script was built to parse and evaluate structures based on completeness, experimental procedure, resolution, R-value and number of outliers in terms of inter-atomic distances and bond angles. From the final viable structures, training data was prepared through the generation of an amino acid matrix, PSSM from PSI-BLAST[1], Secondary Structure probabilities from SPOT-1D[2], Solvent Accessible Surface Area (SASA) values from VMD[3], and physical properties[4]. SASA values for protein chains with partial transmembrane regions were updated by membrane topology predictions generated from TOPCONS[5]. Finally, angle and distance targets were calculated from the coordinates of backbone atoms available in PDB files.

Baseline Structure Prediction: A Bidirectional LSTM based deep neural network was developed in order to accurately predict structural elements of an unknown target protein. To increase accuracy for prediction, the base model was trained on a dataset of protein chains with highest sequence similarity to the target. Targets that were evaluated to be membrane proteins were predicted from a model trained strictly on a dataset containing only membrane protein chains. All datasets were constructed from clean, high resolution, non-homologous protein chains so as to further improve predictive capabilities. Final target predictions provided an initial baseline structure from which further processing steps were applied.

***Protein Folding Simulation*:** For unknown structures, phi and psi dihedral angles were obtained from the prediction realized by the Bidirectional LSTM deep neural network. In addition, omega angle values were obtained through random generation based on a probability distribution from the whole population of protein structures, simulating a variation of the Random Energy Model (REM)[6] in protein folding theory.

The random omega generation process was utilized to simulate realistic folding intermediates with a conformational sampling space of 10^7 unique folds. Also, to account for inherent error in the LSTM prediction of dihedral angles phi and psi, random noise was added to evaluate the possibility for small augmentations that may have resulted in a structure with lower potential energy. Discrete conformational modifications also served to overcome potential energy barriers between local minima and move further toward the native fold. The angle-based NERF[7] (Natural Extension Reference Frame) algorithm was used to realize the cartesian coordinates of fold samples in phi, psi, omega dihedral space for structural evaluation.

Potential energy, radius of gyration and van der Waals radii were used as filtering criteria. Viable structures left over from the filtering step were then subjected to clustering via MUFold-CL[8] for more isolated comparisons. Folds with perceived minimal potential energy values with the highest structural integrity were selected as prospects for fine tuning and evaluation.

***Fine Tuning: Minimizing Potential*:** Side chains for all residues in each remaining candidate fold were generated by the psfgen tool which utilizes the CHARMM36[9] force field for accurate structure creation. Structures were ranked based on an orientation-dependent atomic potential calculated using calRW+[10]. The highest ranking folds were selected for structural optimization through Molecular Dynamics (MD) simulation.

MD simulation was used in order to further minimize potential energy through the calculation of inter-atomic forces with the added presence and influence of water molecules using the NAMD[11] package. The application of either explicit or implicit solvent calculations was decided taking into account chain size and computational capacity. Final results from molecular simulation provided further insights into the quality of each fold candidate.

***Distance Based Corroboration*:** Final folded structures were evaluated through a confirmation process based on corroborating candidate inter-atomic distance values with predicted inter-atomic distances from DeepMetaPSICOV[12]. In this step, discernment based on intuition of inter-residue interactions and experience in protein folding theory from human intervention was utilized.

Template Based Modeling: In the case that a prediction target's amino acid sequence achieved at least 80% similarity with a sequence of a protein with known structure, the template based modeling algorithm HHpred[13] was used. When significant, independent regions of a protein chain were unmatched or evaluated to have structures that did not corroborate with expert intuition, the LSTM based structural prediction process was utilized to predict these localized structures.

1. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.
2. Hanson, J., Paliwal, K., Litfin, T., Yang, Y., & Zhou, Y. (2019). Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics, 35(14), 2403-2410.

3. Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. Journal of molecular graphics, 14(1), 33-38.

4. Zhou, Y., & Faraggi, E. (2010). Prediction of one-dimensional structural properties of proteins by integrated neural networks. Introduction to protein structure prediction: methods and algorithms, 45-74.

5. Tsirigos, K. D., Peters, C., Shu, N., Käll, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic acids research, 43(W1), W401-W407.

6. Bryngelson, J. D. and Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding), J. Phys. Chem. 93 (19), 6902–6915.

7. Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., & Strauss, C. E. (2005). Practical conversion from torsion space to Cartesian space for in silico protein synthesis. Journal of computational chemistry, 26(10), 1063-1068.

8. Zhang, J., & Xu, D. (2013). Fast algorithm for population-based protein structural model analysis. Proteomics, 13(2), 221-229.

9. Huang, J., & MacKerell Jr, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Journal of computational chemistry, 34(25), 2135-2145.

10. Zhang, J., & Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one, 5(10), e15386.

11. Phillips, J. C., Hardy, D. J., Maia, J. D., Stone, J. E., Ribeiro, J. V., Bernardi, R. C., ... & McGreevy, R. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. The Journal of Chemical Physics, 153(4), 044130.

12. Kandathil, S. M., Greener, J. G., & Jones, D. T. (2019). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins: Structure, Function, and Bioinformatics, 87(12), 1092-1099.

13. Söding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. W244–W248.

PrayogRealDistance

Real-valued protein distance prediction

Badri Adhikari[1], Bikash Shrestha[1] and Matthew Bernardini[1]
[1] - *University of Missouri-St. Louis*
adhikarib@umsl.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

Contacts or distograms (distance bins) are human defined zero-one or multi-class labels. After Kukic et al.[1] and Walsh et al.[2] and introduced the idea of real-valued distance prediction, we re-introduced this paradigm in the context of deep learning and recently released PDNET[3], an open-source framework for distance prediction. Different from the successful methods such as RaptorX, AlphaFold, and trRosetta, which predict distograms, ours is a deep learning method for predicting real-valued protein distances. This work is an extension of PDNET, and our deep ResNet model is trained using a representative set of 43 thousand protein chains and learns to predict only a real-valued distance map.

Methods

We generated MSAs using the DeepMSA[4] tool developed by the Zhang group. We augmented DeepMSA using metagenomic sequence databases from multiple sources. These databases are large in size, ranging from 50 GB to 450 GB when uncompressed. Since, running DeepMSA with these databases is slow with conventional hard-drives, we used solid state disks (SSDs). The common parameters for running alignment prediction tools such as coverage and e-value consider the entire sequence as input and are ineffective when the input sequence is long. Typical coverage parameter values such as 60% are not effective when we are searching for alignments for a long sequence. This is because the sequence hits for some subsequence of our input sequence with length lesser than the coverage parameter are not reported. Although structural domain prediction is a possible route to explore, previous CASP participants have reported that a failed domain splitting can badly hurt the precision. In this work, we evenly split the input sequence longer than 256 residues, into overlapping pieces of 256 residues long subsequences, with an overlap of 128 residues. For example, for a protein of length 500, we will have three subsequences: (1) 1 to 256 residues, (2) 128 to 384 residues, and (3) 244 to 500 residues. Then, we generate MSAs for all the subsequences including the original full input sequence. We found that merging these MSAs does not work well. Hence, we predict distances with all the MSAs for each subsequence and later merge the overlapping distance maps by selecting the minimum predicted distance at each pixel (shorter distances are more accurately predicted).

Current approaches to train a deep learning model either use a smaller set of a sequence similarity reduced database such as PISCES or a structural similarity reduced database such as CATH. For example, methods such as Raptor-X use a dataset reduced to have minimum sequence similarity, whereas more recent methods such as AlphaFold use a dataset reduced to have minimum structural similarity. A recent work by Kandathil et al.[5] group suggests that the future is in the use of structural similarity reduced datasets such as CATH or ECOD for training and evaluation of deep learning methods. While sequence similarity reduced databases ensure the representativeness of the sequences of known proteins, the later capture representativeness of the structural fold space. In this work, we were interested in training a deep learning model that learns from both, the protein sequence space and the structural space. To achieve this we merged the PISCES and CATH datasets. We used the May 2018 release of the PISCES dataset with the 27,832 chains curated using the following parameters: percentage identity cutoff = 70%, resolution cutoff = 3.0 Angstroms, R-factor cutoff = 1.0, and X-RAY structures excluded. The dataset is maintained by the Dunbrack Lab and is available at http://dunbrack.fccc.edu/. We further cleaned this list by removing chains that had large structural gaps after removal of non-standard amino acids or had lesser than 12 residues. Chains longer than 512 residues were trimmed by keeping only the first 512 residues. Our final PISCES set included 27,319 chains (set P). Similarly, we cleaned the v4.2 of CATH domains (released in April 2018) consisting of 31,289 structural domains, to obtain a final set consisting of 24,864 unique chains (set C). The CATH dataset is available at https://www.cathdb.info/. Finally, we merged the two sets to obtain a total of 43,071 unique protein chains (P U C). This final development set is lesser than the sum of the two because of a large number of overlapping protein chains. A random set of 200 chains IDs from the development set are selected as a validation set leaving the remaining as the training set.

Our network architecture is a variant of a standard residual network (ResNet). Each residual block in our network consists of a batch normalization layer followed by a exponential linear unit (ELU) activation, a 2D convolution layer consisting of 128 3x3 filters, a dropout layer with dropout rate of 20% followed by ELU activation, and finally a 2D convolution layer consisting of 128 filters that alternate between 3x3 and 1x5 kernels, and also at alternating dilations of 1, 2, and 4. In addition to the 128 residual blocks, the architecture has a 2D convolutional block to shrink the input volume (128x128x322) so the ResNet block receives 128 channel input, and a 2D convolutional block that receives the output of the ResNet block and shrinks the number of channels to one, effectively predicting real valued distances. With 128 residual blocks, and effectively 256+ convolutional layers, the model has 29.5 million parameters. We train our model at a fixed window of 128 x 128, i.e. in each model training/validation task, we only predict the distances between two sequence pairs each of maximum 128 residues long. It is counter-intuitive that such a setting does allow the model to learn the distances anywhere in the distance map for a protein of any length, and not just a 128 sequence window. With the batch size set to two, one epoch of training takes about 8 hours in a TITAN RTX GPU when the features and distance maps are all loaded from solid state disks.

1. Kukic, P., Mirabello, C., Tradigo, G. et al. Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. BMC Bioinformatics 15, 6 (2014).
2. Walsh, I., Baù, D., Martin, A.J. et al. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol 9, 5 (2009).
3. Adhikari, B. A fully open-source framework for deep learning protein real-valued distances. Sci Rep 10, 13374 (2020).
4. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics. 2020;36(7):2105-2112.
5. Kandathil, Shaun M., Joe G. Greener, and David T. Jones. "Recent developments in deep learning applied to protein structure prediction." Proteins: Structure, Function, and Bioinformatics 87.12 (2019): 1179-1189.

The PreferredFold pipeline: a top-down approach to predicting protein structure

Yibing Wu[1,2,3], Shin-ichi Maeda[2], Yi Ouyang[1,2], Nobuyuki Ota[1,2]
1. Preferred medicine, Burlingame, CA 94010 2. Preferred networks, Tokyo 100-0004 3. Department of Pharmaceutical Chemistry, UCSF, CA 94115
ota@preferred-medicine.com

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y. MetaG:N. Fragm:Y.3-9; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y.*

Together with residue co-evolution, deep convolutional neural nets have made significant progress in protein structure prediction[1-3]. However, dealing with a large protein or a modular protein still remains a big challenge. Here, we propose a top-down approach to predicting protein structure by following a divide and conquer strategy. This idea is aligned well with the cropping method used to train the deep neural nets to predict protein structures[1].

Methods

Similar to structure calculation approaches used in NMR community, our approach (Figure 1) mainly consists of three parts: constraint generation, structure calculation and structure refinement. While our own modules are being/will be developed for each part, the majority of the current version implements the start of the art algorithms widely accepted by the community.

Our top-down approach to predicting structure starts with multiple sequence alignment searching with the full-length sequence. deepMSA4 searches MSA against uniclust30_2018_08, uniref90 and metaclust_nr. The core of this pipeline relies on three CNN-based distance and torsion-angle predictors: RaptorX2, trRosetta3 and Prospr5 that is an implementation of AlphaFold1. The simple mean possibility will be calculated after obtaining the distance possibilities from all three. Then distance and torsional distributions will be converted to potential inputs to Xplor-NIH6 with a short MD or trRosetta that minimizes energy by gradient descent. The convergence will be evaluated by aligning the secondary structures of the protein core for the top 5 structures with lowest energy. RMSD 4Å is an empirical cutoff but it can be varied and judged by human experience. If the fold is not converged, the sequence is split into smaller domains/parts by inspecting the current fold. We repeat the approach for these individual domains/parts.

Besides the deep-learning based free modeling, template modeling with Muster[7] and Phyre2[8] are also performed. In addition, *ab initio* modeling is run with Rosetta[9] if the protein sequence size is less than 120 amino acids. The physic-based and statistics-based potentials by Rosetta is an excellent complement to the data-driven free and template modelling.

Finally, AIDA[10] and/or Modeller[11]/RosettaCM[12] are used to assemble the chopped domain/parts. To better pack the sidechains, GalaxyRefine[13] is applied for final MD refinement.

For future development, our on-going studies focus on algorithms using language models and deep reinforcement learning. Automating the new version to a full end-to-end pipeline is also one of our goals.

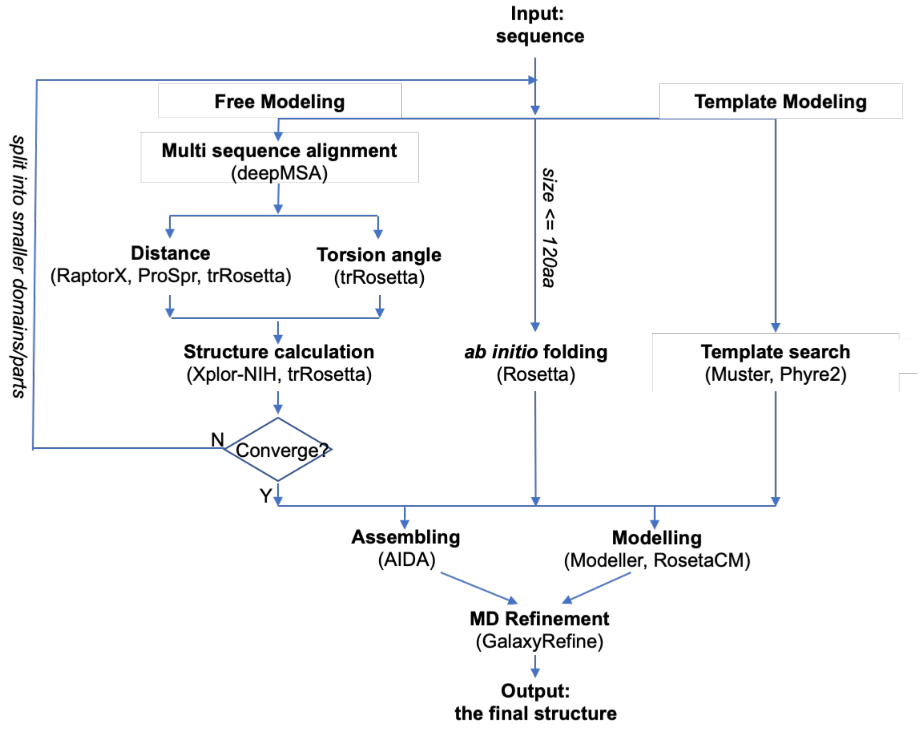


Figure 1: The PreferredFold protein structure pipeline

Availability

The detailed description for a publication will be prepared.

1. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710, doi:10.1038/s41586-019-1923-7 (2020).
2. Xu, J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A 116, 16856-16865, doi:10.1073/pnas.1821309116 (2019).
3. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A 117, 1496-1503, doi:10.1073/pnas.1914677117 (2020).
4. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 36, 2105-2112, doi:10.1093/bioinformatics/btz863 (2020).
5. Wendy M Billings, B. H., Todd Millecam, David Wingate, Dennis Della Corte. ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. bioRxiv https://doi.org/10.1101/830273 (2019).
6. Schwieters, C. D., Bermejo, G. A. & Clore, G. M. A three-dimensional potential of mean force to improve backbone and sidechain hydrogen bond geometry in Xplor-NIH protein structure determination. Protein Sci 29, 100-110, doi:10.1002/pro.3745 (2020).
7. Wu, S. & Zhang, Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72, 547-556, doi:10.1002/prot.21945 (2008).

8. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10, 845-858, doi:10.1038/nprot.2015.053 (2015).
9. Fleishman, S. J. et al. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6, e20161, doi:10.1371/journal.pone.0020161 (2011).
10. Xu, D., Jaroszewski, L., Li, Z. & Godzik, A. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. Bioinformatics 31, 2098-2105, doi:10.1093/bioinformatics/btv092 (2015).
11. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Protein Sci 86, 2 9 1-2 9 37, doi:10.1002/cpps.20 (2016).
12. Song, Y. et al. High-resolution comparative modeling with RosettaCM. Structure 21, 1735-1742, doi:10.1016/j.str.2013.08.005 (2013).
13. Heo, L., Park, H. & Seok, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. Nucleic Acids Res 41, W384-388, doi:10.1093/nar/gkt458 (2013).

Protein-blacksmith

Protein structure refinement: how frustration analysis assists principal component guided simulations

Shikai Jin, Mingchen Chen, Xingcheng Lin, Xun Chen, Wei Lu and Peter G. Wolynes[1]

[1] - Center for Theoretical Biological Physics, Rice University, Houston, 77005

pw8@rice.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y; Fragm:Y.9; Cont:Y; Dist:Y; Tors:N; DeepL:N; EMA:Y; MD:Y*

CASP13 has witnessed significant progress in structure refinement yielding moderate accuracy by means of all-atom molecular dynamics simulation[1]. It has been shown that principal component (PC) guided refinement can successfully search for refined structures that require large conformation changes[2]. This method has outperformed the state of the art methods in CASP12. Frustration analysis[3] augments this approach by finding those parts of the structure that already have high local accuracy. Combining these ideas can lead to the refinement of the protein structure with improved accuracy.

Methods

The starting structure for each refinement target was obtained as input for the simulation. A well-sampled ensemble using the AWSEM-Suite force field in OpenAWSEM based on a total of 60ns of unbiased simulation[4-5]. The Associative memory, Water mediated, Structure and Energy Model (AWSEM)-Suite, is a force field who has transferrable potentials that haven been optimized using the energy landscape theory of protein folding. From the sampled structures, principal component vectors were calculated based on the motion of the CA atoms in this simulation.

These principal components were used along with the initial starting structures to generate five windows of umbrella samplings using all-atom simulations that have been systematically perturbed to a reference points at -2.0, -1.0, 0, 1.0, and 2.0 times the standard deviation of the PC values from the AWSEM-Suite simulation[2]. The all-atom simulations were performed using the CHARMM36m force field and a time step of 2.0 fs in Gromacs 2018.6 patched Plumed 2.5.1[6-7]. The system was neutralized by adding Na and Cl ions to the 0.15 M final ionic concentration. The simulations were performed in the constant particle number, volume, and temperature ensemble. The umbrella potentials were gradually increased in strength over time by linearly increasing from 0.0 to a large spring constant of k = 200 kJ/mol within the first 1 ns and held constant at this large value for 2 ns. This strength was slowly decreased to k = 5.0 kJ/mol over the next 1 ns.

The final structures after the biased simulations were extracted and used for calculating the frustration pattern of each residue. Frustration indices are obtained by statistically analyzing the energy differences between the protein in its current conformation and a set of the randomly generated decoy states with locally different side chain packings and identities. The energy function for calculating the frustration was defined in the paper[3]. The residues that were least frustrated were chosen and were fixed during the next simulation. Another 47ns simulation with a constant bias to this frozen censored template k = 5.0 kJ/mol was performed to sample the conformations of the structure.

Every frame of the each of the 5 simulations was extracted. We first evaluated these structures with an algorithm that reported in our paper[3]. The top 500 structures were saved as the pool. Then the ClustQ algorithm was used to pick out the best 5 structures from the top 500 that calculated in the previous step[8]. The B-factors were calculated using the trajectory from the second half of the AWSEM-Suite simulation.

Availability

The source code for the AWSEM-Suite force field within the LAMMPS suite is available for download on Github (https://github.com/adavtyan/awsemmd). Other documentation and references can be found on this website: http://awsem-md.org. OpenAWSEM is available at https://github.com/npschafer/openawsem that implements the AWSEM-Suite force field in OpenMM.

1. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins. 87, 1011-1020.
2. Lin, X., Schafer, N. P., Lu, W., Jin, S., Chen, X., Chen, M., ... & Wolynes, P. G. (2019). Forging tools for refining predicted protein structures. Proc. Natl. Acad. Sci. U S A. 116, 9400-9409.
3. Chen, M., Chen, X., Jin, S., Lu, W., Lin, X., Wolynes, P. G. (2020) Protein Structure Refinement Guided by Atomic Packing Frustration Analysis. BioRxiv.
4. Jin, S., Chen, M., Chen, X., Bueno, C., Lu, W., Schafer, N. P., ... & Wolynes, P. G. (2020). Protein Structure Prediction in CASP13 using AWSEM-Suite. Journal of Chemical Theory and Computation. 16, 3977-3988.
5. Lu, W., Bueno, C., Schafer, N. P., Moller, J., Jin, S., Chen, X., Chen, M., Gu, X., Pablo, J. J., Wolynes, P. G. (2020) OpenAWSEM with Open3SPN2: a fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. BioRxiv.
6. Bonomi, M., Bussi, G., Camilloni, C., Tribello, G. A., Banáš, P., Barducci, A., .& Capelli, R. (2019). Promoting transparency and reproducibility in enhanced molecular simulations. Nat. methods, 16, 670-673.
7. Berendsen, H. J., van der Spoel, D., & van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. Computer physics communications, 91, 43-56.
8. Alapati, R., & Bhattacharya, D. (2018). clustQ: Efficient protein decoy clustering using superposition-free weighted internal distance comparisons. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 307-314).

QMEANDisCo – distance constraints applied on model quality estimation

G. Studer[1,2], C. Rempfer[1,2], A.M. Waterhouse[1,2], R. Gumienny[1,2], J. Haas[1,2] and T. Schwede[1,2]
[1]*Biozentrum, University of Basel, Basel 4056, Switzerland,* [2]*SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland*
gabriel.studer@unibas.ch

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

Assigning reliable estimates of overall, as well as per-residue qualities in 3D protein structure models is crucial to determine their utility and potential applications. Single model methods are capable of assessing individual models. In contrast, consensus methods exploit the variability of model ensembles for their predictions. QMEANDisCo[1] extends the single model composite score QMEAN[2] by introducing a consensus-based distance constraints (DisCo) score. QMEANDisCo is continuously benchmarked on the Continuous Automated Model EvaluatiOn platform (CAMEO)[3] and a previous version of it, FaeNNz[4], has successfully been tested in the CASP13 experiment.

Methods

QMEAN combines statistical potentials of mean force and the consistency of a model with structural features predicted from sequence to generate overall and per-residue quality estimates. QMEANDisCo complements the individual single model scores from QMEAN with a consensus based distance constraints score: DisCo. DisCo assesses the agreement of interatomic distances in a protein model with ensembles of constraints derived from experimentally determined protein structures that are homologous to the model being assessed. By directly using information derived from homologous template structures, QMEANDisCo avoids the requirement of an ensemble of models as input and can thus be considered a quasi-single model method. In case a large number of close homologues are detected, DisCo is likely to be highly reliable. However, low reliability is expected in case of few or no close homologues. In order to combine the ability of single model scores to assess individual models with the power of DisCo in cases of sufficient structural information, we use feed-forward neural networks to adaptively weigh the various components.

Availability

QMEANDisCo is available as a web-server at https://swissmodel.expasy.org/qmean. The source code can be downloaded from https://git.scicore.unibas.ch/schwede/QMEAN.

1. Benkert,P. et al. (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics, 27, 343-350.
2. Studer,G. et al. (2020) QMEANDisCo − distance constraints applied on model quality estimation. Bioinformatics, 36, 1765-1771.

3. Haas,J. et al. (2019) Introducing ‘best single template’ models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). Proteins, 87, 1378-1387.
4. Cheng,J. et al. (2019) Estimation of model accuracy in CASP13. Proteins, 87, 1361-1377.

QUARK

Protein 3D Structure Prediction by D-QUARK in CASP14

Chengxin Zhang[1], Yang Li[1], Wei Zheng[1], Eric Bell[1], Xiaoqiang Huang[1], Robin Pearce[1], Xiaogen Zhou[1], Yang Zhang[1,2]

1 - Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, 2 – Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091

yangzhanglab@umich.edu

***Key:** Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:Y.1-20; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Methods

The tertiary structure prediction of the QUARK group in CASP14 is based on D-QUARK (C Zhang et al, in preparation), an extension of QUARK and C-QUARK, which integrates deep-learning-based distance and torsion angle predictions with replica-exchange Monte Carlo fragment assembly simulations. The pipeline consists of four consecutive steps. First, starting from the query sequence, a set of multiple sequence alignments (MSAs) are created by DeepMSA[1] and its variants, by iteratively searching the query through whole-genome and metagenome sequence databases (Metaclust, BFD, Mgnify, and IMG/M), where the MSA with the highest accumulative probability obtained by the TripletRes top $10L$ predicted contacts[2] is selected for the next step of modeling.

In the second step, the selected MSA is used as the input for DeepPotential, a newly developed deep residual neural-network-based predictor (see DeepPotential Abstract), to create multiple spatial characteristics including (1) distance-maps for both Cα and Cβ atoms and (2) N-Cα-Cβ torsion angles. Considering that DeepPotential tends to have higher precision for the distance models with shorter distance cutoffs, four sets of distance profiles are generated with distance ranges from [2, 10], [2, 13], [2, 16] and [2, 20] Å, where the four ranges are divided into 18, 24, 30, and 38 distance bins, respectively. Only the distance profiles from lower distance cutoffs are selected, i.e., distances in [2-10) Å are selected from model Set-1, distances in [10-13) Å from Set-2, [13-16) Å from Set-3, and [16-20] Å from Set-4. In addition to DeepPotential, three deep-learning naïve Bayes classfier based contact predictors (TripletRes[2], ResPRE[3], and NeBcon[4]) are used to create Cα and Cβ contact-maps with a distance cutoff of 8 Å. Meanwhile, LOMETS3, a newly developed meta-server program containing both profile- and contact-based threading programs (see Zhang-TBM Abstract), is used to identify structural templates from a non-redundant PDB structural library. Based on the significance and consensus of the LOMETS3 alignments, the target is assigned to one of four categories (Trivial, Easy, Hard, and Very-Hard)[5].

In the third step, full-length structure models are constructed using replica-exchange Monte Carlo (REMC) simulations under the guidance of a composite force field: $E = E_{QUARK} + E_{contact} + E_{distance} + E_{torsion} + E_{LOMETS}$. Here, E_{QUARK} is the inherent QUARK potential containing multiple physics- and knowledge-based energy terms [6] and fragment-derived contact-maps[7]; $E_{contact}$ is a three-gradient potential that accounts for the contact-map prediction[8]; $E_{distance}$ and $E_{torsion}$ are the negative logarithm of the DeepPotential predicted probabilities for distance and torsion angle maps, respectively; E_{LOMETS} is the spatial restraints collected from the LOMETS3 templates. The LOMETS-based term was extended from the I-TASSER force field[9] but with newly added template-based Cα-Cβ torsion angles. Three types of REMC simulations (labeled as 'QE', 'QN' and 'QT') are run depending on a target's category, i.e., 'QE' runs the original QUARK protocol with initial conformations created from random fragment connection and without including the LOMETS-based restraints in the force field; 'QN' is similar to 'QE' but with the initial conformations created from the LOMETS templates; 'QT' is similar to 'QN' but with the LOMETS-based restraints included in the force field. 'QE' is run for Very Hard and Hard targets, 'QN' for Hard and Easy targets, and 'QT' for Easy and Trivial targets, respectively. For each pipeline, five REMC simulations are performed, where the structural decoys from the 10 lowest-temperature replicas are submitted to SPICKER[10] for structure clustering and model selection.

In the fourth step, the SPICKER clusters are refined at the atomic level using FG-MD[11] and ModRefiner[12] sequentially, followed by FASPR[13] for side-chain rotamer repacking. To select models generated from different pipelines, a set of six MQAP programs are implemented, including the D-QUARK confidence score, predicted contact-map satisfaction rate, structural consensus measured by pair-wise TM-score[14], and three statistical potentials (RW, RWplus[15], and Rotas[16]). A meta-MQAP consensus score is calculated as the sum of the rank of the six MQAP scores; the models with the lowest consensus MQAP scores are selected for submission.

For multiple-domain sequences, FUpred[17] and ThreaDom[18] are used to predict the domain boundaries and linker regions from the contact-maps and LOMETS threading alignments, respectively. Structural models are first predicted by D-QUARK for the individual domains separately, which are then assembled into full-length models for the whole chain using a rigid-body domain docking and assembly algorithm, DEMO[19], guided by the whole-chain D-QUARK models and structural analogs identified by TM-align[20]. The procedure is fully automated.

Availability

https://zhanglab.ccmb.med.umich.edu/QUARK

1. Zhang,C., Zheng,W., Mortuza,S.M., Li,Y. & Zhang,Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105-2112.
2. Li,Y., Zhang,C., Bell,E.W., Zheng,W., Zhou,X., Yu,D.J. & Zhang,Y. (2020). Deducing high-accuracy protein contact-maps from a triplet ofcoevolutionary matrices through deep residual convolutional networks. submitted.
3. Li,Y., Hu,J., Zhang,C., Yu,D.J. & Zhang,Y. (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **35**, 4647-4655.
4. He,B., Mortuza,S.M., Wang,Y., Shen,H.B. & Zhang,Y. (2017). NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics* **33**, 2296-2306.
5. Zhang,Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* **82 Suppl 2**, 175-87.
6. Xu,D. & Zhang,Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-35.
7. Xu,D. & Zhang,Y. (2013). Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* **81**, 229-39.

8. Zhang,C., Mortuza,S.M., He,B., Wang,Y. & Zhang,Y. (2018). Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins* **86 Suppl 1**, 136-151.
9. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. & Zhang,Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7-8.
10. Zhang,Y. & Skolnick,J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25**, 865-71.
11. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**, 1784-95.
12. Xu,D. & Zhang,Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* **101**, 2525-34.
13. Huang,X., Pearce,R. & Zhang,Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758-3765.
14. Zhang,Y. & Skolnick,J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.
15. Zhang,J. & Zhang,Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **5**, e15386.
16. Park,J. & Saitou,K. (2014). ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics* **15**, 307.
17. Zheng,W., Zhou,X., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* **36**, 3749-3757.
18. Wang,Y., Wang,J., Li,R., Shi,Q., Xue,Z. & Zhang,Y. (2017). ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. *Nucleic Acids Res* **45**, W400-W407.
19. Zhou,X., Hu,J., Zhang,C., Zhang,G. & Zhang,Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proc Natl Acad Sci U S A* **116**, 15930-15938.
20. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.

RaptorX

Improved Protein Contact and Structure Prediction by Deep Learning

Jinbo Xu[1], Fandi Wu[1] , Matthew Mcpartlon[1,2] and Jin Li[1,2]
[1] – *Toyota Technological Institute at Chicago,* [2] – *University of Chicago*
jinboxu@gmail.com

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:N*

Since CASP13 we mainly focuses on improving the RaptorX software package in the following aspects: 1) prediction of contact and distance distribution by improving the deep ResNet architecture, better feature design and better model training strategies; 2) building 3D models from predicted distance distribution by energy minimization; and 3) integrating template information by deep ResNet when good templates are available.

Methods

The key components of our method include 1) a deep ResNet that predicts distance probability distribution for three types of backbone atom pairs (Cb-Cb, Ca-Ca and N-O) and inter-residue orientation probability distribution; 2) a revised gradient-based energy minimization method that builds 3D models from predicted distance and orientation potential as well as backbone torsion angles.

Distance and orientation prediction. In addition to inter-residue distance distribution, we also employ deep ResNet to predict inter-residue orientation distribution. We have studied two types of inter-residue orientation and found out that the orientation defined in the trRosetta paper[1] works slightly better. We discretize distance into the following bins: 0-2, 2-2.4, 2.4-2.8, 2.8-3.2, …, 19.6-20, > 20 and the orientation angles uniformly with the bin width set to 12 degree. In addition, we use one label to indicate an “unknown” distance when at least one of the two atoms do not have valid 3D coordinates in the PDB file.

Network architecture. The overall deep ResNet architecture is similar to what has been used in CASP13[2,3], except that the ResNet used in CASP14 has a larger capacity. In particular, it has 100 2D convolutional layers and at each layer on average 120 filters. In contrast, the deep ResNet used in CASP13 has only 60 2D convolution layers and each layer has about 60-70 filters. We have also tested some attention methods, but not observed performance gain. We simultaneously predict distance and orientation angles to reduce both training and test time. This multi-tasking learning framework was implemented in CASP13, but not fully tested back then. We have not observed obvious contact precision gain by multi-task learning, though.

Model training. We train our deep ResNet by subsampling MSAs. That is, we randomly sample 50% of the sequence homologs from an MSA (when it has at least 2 sequences) and then derive all input features from the sampled MSA. We train our deep ResNet for 20 epochs and select the model with the minimum validation loss as our final model. We have trained several deep models and then use them as an ensemble to predict distance and orientation distribution.

Input features. In CASP13, we generated MSAs without using any metagenome data. In CASP14, when one MSAs is shallow (i.e., Meff<=6), we enhance it by metagenome data. Here Meff refers to the number of effective sequence homologs in an MSA. In CASP13, we only use the L*L co-evolution matrix generated by CCMpred where L is the protein sequence length. Here we also use the much larger L*L*21*21 matrix generated by CCMpred, which can result in slightly better performance. We have also tried the L*L*21*21 covariance matrix, but not observed performance gain.

Building protein 3D models from predicted constraints with PyRosetta. The 3D model building protocol consists of the following major steps: (1) convert predicted distance and orientation probability distribution into energy potential using the method developed in our previous work[4]. We have experimented with both DFIRE and DOPE reference states and found that on average the DFIRE reference is slightly better; (2) interpolate the discrete energy potential for each pair of atoms (residues) to a continuous curve using the Rosetta spline function; (3) minimize the energy potential by gradient-based energy minimization, i.e., the LBFGS algorithm implemented in Rosetta. The starting 3D model for energy minimization is sampled from our predicted phi/psi probability distribution. Since LBFGS may not converge to the global minimum, once it reaches a local minimum, we perturb backbone phi/psi angles by a small deviation and then apply LBFGS again to see if a conformation with a lower energy potential can be generated. This perturbation procedure is repeated up to three times.

Deep comparative protein modeling. When a good template is available, we predict inter-residue distance and orientation distribution from both evolutionary and template information. A template is good when the E-value returned by HHblits is less than 1E-10 and its sequence identity with the target is no more than 35%.

Model refinement. This module is not used in RaptorX server since we do not have sufficient computing resources to do server prediction on time. Feig group did refinement on our server models and submitted them in a separate human group. This human group did not take advantage of models submitted by other servers. Instead it only refined models submitted by RaptorX. Experimental results on the CASP13 FM targets indicate that Feig's refinement method can consistently improve the 3D models produced by RaptorX.

Results

Contact prediction accuracy. On the 31 CASP13 FM targets, the top L/5, L/2 and L long-range contact precision is around 0.80, 0.69 and 0.58, respectively, better than our CASP13 result (0.70, 0.58 and 0.45, respectively), which is also the best in CASP13. The corresponding F1 is 0.277, 0.451 and 0.521, respectively, also much better than our CASP13 result (0.233, 0.362 and 0.411, respectively).

3D modeling accuracy. On the 32 CASP13 FM targets (including T0950), the average TMscore of the first models is around 0.64, better than what has been reported in literature.

Deep comparative modeling. Tested on the CASP13 TBM hard and easy targets, deep comparative modeling indeed can produce models with higher quality than pure template-free modeling and pure template-based modeling (e.g., MODELLER). However, in CASP14 there are not many targets suitable for deep comparative modeling, i.e., HHblits E-value <1E-10 and sequence identity <35%.

Folding of human-designed proteins. Tested on 32 de novo proteins designed by two research groups without using any co-evolution information, our method can predict correct folds for all of them with an average TMscore above 0.75.

Availability

The web server is available at http://raptorx.uchicago.edu/ and the standalone software package is available at https://github.com/j3xugit/RaptorX-3DModeling.

1. Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. PNAS 2019.
2. Jinbo Xu. Distance-based protein folding powered by deep learning. PNAS 2019
3. Jinbo Xu and Sheng Wang. Analysis of distance-based protein structure prediction by deep learning in CASP13. PROTEINS 2019.
4. Feng Zhao and Jinbo Xu. A position-specific distance-dependent statistical potential for protein structure and functional study. Structure 2012.

RaptorX-QA

Improved protein model quality assessment by integrating sequential and pairwise features using deep learning

Xiaoyang Jing, Jinbo Xu
Toyota Technological Institute at Chicago, Chicago, IL 60637, USA
jinboxu@gmail.com

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Significant progress has been made in computational protein structure prediction, especially template-free protein modeling. To facilitate application of predicted 3D models, it is desirable to have an estimation of their local and global quality in the absence of experimental structures. This work focuses on local and global quality assessment (QA) when there are very few models built for a protein (and thus, consensus methods are not very effective).

Methods

Inspired by our successful application of 1D and 2D convolutional residual neural networks (ResNet) to protein contact/distance prediction and distance-based protein folding[1,2], we propose a new single-model-based QA method (ResNetQA) for both local and global QA by using a deep neural network composed of 1D and 2D ResNet. The 2D ResNet module extracts useful information from pairwise features such as model-derived distance maps, co-evolution information and predicted distance potential. The 1D ResNet is used to predict local (global) model quality from sequential features and pooled pairwise information generated by 2D ResNet.

Feature extraction: For each protein sequence, we run HHblits to build its multiple sequence alignment (MSA) and then derive three types of features: sequential features, coevolution information and predicted distance potentials. Sequential features include: one-hot encoding of primary sequence, the relative position of a residue in a sequence, PSSM, SS3 and ACC predicted by RaptorX-Property. Coevolution information includes the output generated by CCMPred and raw and APC-corrected mutual information (MI). Distance potentials ($C\beta$-$C\beta$) is derived from distance distribution predicted by RaptorX-Contact from MSA. From each 3D model, we derive the following structural features: 1) secondary structure (SS3) and relative solvent accessibility (RSA) calculated by DSSP; and 2) distance maps of three atom pairs ($C_\alpha C_\alpha$, $C_\beta C_\beta$ and NO).

Deep neural network architecture: Our deep network mainly consists of one 2D ResNet module and one 1D ResNet module. The 2D ResNet module extracts information from pairwise features (model-derived distance maps, co-evolution information and predicted distance potential). This module outputs a high-level 2D feature map, which is then converted to two 1D feature maps by row-wise and column-wise mean pooling, respectively, and fed into the 1D ResNet module together with the original sequential features (one-hot encoding, residue relative position, PSSM, SS3 and ACC predicted by RaptorX-Property, SS3 and RSA calculated by DSSP). The output of the 1D ResNet module is used to predict local and global model quality. To predict local quality, one fully-connected layer and one sigmoid layer are employed at each residue. To predict global quality, the output of the 1D ResNet module is converted to one vector by mean pooling and fed into one fully-connected layer and one sigmoid layer.

Model training: For local QA, our deep model predicts a residue-wise S-score defined by $S(d)=1/(1+(d/d_0)^2)$ where d is the distance deviation of one $C\alpha$ atom from its position in the experimental structure calculated by LGA. Here we set d_0 to 3.8Å instead of 5.0Å to yield accurate prediction for small d. We convert predicted S-score to predicted distance error (or deviation) by the inverse function of S(d). For global QA, our deep model predicts GDT_TS. The loss of our deep model is the MSE (Mean Square Error) between predicted quality and its ground truth. Our deep network is trained to simultaneously predict local and global quality combined by equal weight. Further, to reduce bias introduced by a small training dataset, we train our deep model using a large set of decoy models of more than 14,000 proteins, in addition to CASP and CAMEO models. In particular, we built both template-based and template-free 3D models for ~14,000 proteins randomly selected from the CATH dataset using our in-house structure prediction software RaptorX.

Results

The test results on the CASP12 (64 targets) and CASP13 (76 targets) datasets in Table 1 show that our method significantly outperforms others in terms of most evaluation metrics.

Table 1. Comparison of ResNetQA with other single-model methods on local and global QA

Dataset	Method	Local QA			Global QA		
		PCC[1]↑	ASE[2]↑	AUC[3]↑	PCC[1]↑	Diff[4]↓	Loss[5]↓
CASP12 Stage 2	**ResNetQA**	**0.5866**	**0.8515**	**0.8058**	**0.8109**	**0.0785**	**0.0612**(±0.0665[6])
	ProQ3	0.4542	0.7409	0.7517	0.6552	0.1104	0.0615(±0.0649)
	ProQ2	0.4364	0.6928	0.7434	0.6108	0.1338	0.0707(±0.0682)
CASP13 Stage 2	**ResNetQA**	**0.5539**	**0.8373**	**0.7901**	**0.8157**	**0.0861**	**0.0844**(±0.0713)
	ProQ3D	0.4230	0.7312	0.7384	0.6532	0.1060	0.0852(±0.0891)
	ProQ3	0.4134	0.7205	0.7455	0.5921	0.1198	0.0898(±0.0918)
	ProQ4	0.3873	0.6100	0.7235	0.7190	0.1371	0.0871(±0.0908)

1. Pearson correlation coefficient between predicted score and its ground truth (all models of a protein target are pooled together when calculating PCC of local QA). 2. averaged residue-wise S-score error. 3. area under curve, an accurate residue is the one with Cα atom deviates from its experimental position by no more than 3.8 Å. 4. mean absolute difference between predicted global quality and ground truth. 5. absolute quality difference between the predicted best model and the real best model. 6. standard deviation on the Loss metric.

Availability

The source codes of the ResNetQA for both local and global quality assessment are available at: https://github.com/AndersJing/ResNetQA. The predicted distance potential and other sequence features can be generated by the RaptorX web server (http://raptorx.uchicago.edu) or the new version of RaptorX software package (https://github.com/j3xugit/RaptorX-3DModeling).

1. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology. 2017;13(1):e1005324.
2. Xu J. Distance-based protein folding powered by deep learning. PNAS. 2019;116(34):16856-16865.

RBO-PSP-CP

Protein Structure and Distance Prediction by RBO in CASP 14

Kolja Stahl, Stefan Junghans, Mahmoud Mabrouk, Lukas Hönig, and Oliver Brock
Robotics and Biology Laboratory, Technische Universität Berlin

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:N*

RBO-PSP-CP is a protein structure and distance prediction server that leverages distance predictions for template retrieval. The pipeline supersedes and unifies RBO-Aleph[1] and RBO-EPSILON[2].

Methods

RBO-PSP-CP combines evolutionary and sequence-based information to predict distograms. The pipeline builds on the ideas presented in trRosetta[3] (prediction of interresidue orientation was included later) and AlphaFold[4]. Features include the amino acid composition and co-evolutionary features derived from CCMpred[5]. The MSA pipeline is based on DeepMSA[6]. In addition to distogram prediction, we also predict a mixture of Gaussians (n=3) for residue pairs (i,j). Gaussians represent different hypotheses. We can show that of the three Gaussians one Gaussian is often close to the native distance and identifying all correct Gaussians would allow solving the structure with gradient descent. The reduction to three Gaussians reduces the search space and facilitates structure prediction. We use map align[7] to find consistent subsets of the hypotheses encoded in the Gaussians in known structures for template retrieval. The final (model) prediction is a mix of templates and decoys generated with gradient descent. We use Rosetta[8] Energy to score and select the decoys.

Availability

The server will be made available here: https://compbio.robotics.tu-berlin.de/epsilon .

1. Mabrouk, M., Putz, I., Werner, T., Schneider, M., Neeb, M., Bartels, P., & Brock, O. (2015). RBO Aleph: leveraging novel information sources for protein structure prediction. Nucleic acids research, gkv357.
2. Stahl, Kolja, Michael Schneider, and Oliver Brock. "EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction." BMC bioinformatics 18.1 (2017): 303
3. Yang, Jianyi, et al. "Improved protein structure prediction using predicted interresidue orientations." Proceedings of the National Academy of Sciences 117.3 (2020): 1496-1503.
4. Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." Nature 577.7792 (2020): 706-710.
5. Seemayer, Stefan, Markus Gruber, and Johannes Söding. (2014) CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30.21 : 3128-3130.
6. Zhang, Chengxin, et al. "DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins." Bioinformatics 36.7 (2020): 2105-2112.

7. Ovchinnikov, Sergey, et al. "Protein structure determination using metagenome sequence data." Science 355.6322 (2017): 294-298.
8. Rohl, Carol A., et al. "Protein structure prediction using Rosetta." Methods in enzymology. Vol. 383. Academic Press, 2004. 66-93.

Risoluto

Holistic Approach to Integrate Template-based and Template-free Modeling

Y. Yamamori[1], M. Takemoto[2], R. Ishitani[2], K. Mizuno[2], Y. Tsuchiya[1], K. Oono[2] and K. Tomii[1]
[1]- National Institute of Advanced Industrial Science and Technology (AIST), [2]- Preferred Networks Inc.
k-tomii@aist.go.jp

Key*: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N

In CASP14, we made predictions by integrating both template-based and template-free approaches to cope with large proteins and protein complexes. We have developed modeling pipelines based on profile–profile alignments for template-based modeling and contact predictions using deep neural networks (DNNs) for free modeling. Accumulated protein sequence and structure data were analyzed using methods we have developed.

Methods

Template-based modeling and free modeling were performed for all regular targets. For template-based modeling, we generated 3D-models based on target-template alignments derived from our profile–profile aligner FORTE[1–3], using MODELLER[4]. Profiles of both targets and templates were prepared based on multiple sequence alignments (MSAs), which were obtained using three methods: i) a combination of SSEARCH with the MIQS[5] matrix and PSI-BLASTexB[6], followed by MAFFT[7], ii) DELTA-BLAST[8], and iii) HHblits[9]. In some cases, we used multiple templates that were selected manually to generate 3D-models.

For free modeling, we generated 3Dmodels based on predicted contacts of residue pairs using CONFOLD2[10]. Predicted contacts were obtained using three methods: i) DeepECA[11], an end-to-end learning framework of protein contact prediction that can effectively use information derived from either deep or shallow MSAs; ii) distance distribution prediction similar to AlphaFold[12]; and iii) consensus prediction of contacts derived from selected CASP-hosted predictions. We converted distance prediction results, derived from method ii), into contacts when we used them. When no suitable template was found for multimetric targets, we performed rigid-body docking of subunits.

To evaluate and rank the generated 3D-models described above, we mainly used three guidelines: i) the coverage of predicted contacts, derived from DeepECA, satisfied in a model; ii) similarity to a consensus model, selected based on the average TM-score[13] with other models, among models provided by CASP-hosted servers; and iii) Z-scores calculated using FORTE. For easy targets, we weighed classical 3D-scores such as Verify3D[14] and dDFIRE[15]. For multimetric targets, we also considered the stoichiometry of templates. Occasionally, we performed cluster analysis to evaluate and rank our generated 3D-models. Because these guidelines and scores used for ranking are not always consistent, final ranking was sometimes done with human intervention. For the data-assisted target S1063, we used two metrics to select 3D-models to compare the calculated SAXS profile of a model and the experimental SAXS profile: Chi^2 and the volatility of the ratio[16].

For refinement targets, we identified region(s) for emphasis on improvement in a starting model(s) using Verify3D. Subsequently, we sampled conformations for those regions from the generated 3D-models and models provided by CASP-hosted servers. Then we refined 3D models by simulated annealing using MODELLER.

Results

We submitted our prediction models for all CASP14 targets. Most of them, approximately 60%, are models obtained using template-based approaches. The remaining models were derived from template-free approaches or combinations of the two approaches (partially modeled by free modeling). Roughly speaking, these approaches functioned complementarily.

For data-assisted target S1063, we assume that we were able to improve our model based on the multimer shape derived from the experimental SAXS data provided in terms of the results from Chi^2 and the volatility of the ratio.

Availability

The codes of DeepECA and PSI-BLASTexB are available respectively at GitHub, https://github.com/tomiilab/DeepECA and https://github.com/kyungtaekLIM/PSI-BLASTexB. FORTE, MAFFT, and SSEARCH with MIQS are available at http://forteprtl.cbrc.jp/forte/ (under refit), https://mafft.cbrc.jp/alignment/server/ and http://csas.cbrc.jp/Ssearch/.

1. Tomii, K. & Akiyama, Y. (2004). FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics. 20, 594–595.
2. Tomii, K., Motono, C. & Hirokawa, T. (2005) Protein structure prediction using a variety of profile libraries and 3D verification. Proteins. 61, 114–121.
3. Nakamura, T., Oda, T., Fukasawa, Y. & Tomii, K. (2018). Template-based quaternary structure prediction of proteins using enhanced profile–profile alignments. Proteins. 86, 274–282.
4. Webb, B. & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 54, 5.6.1 – 5.6.37.
5. Yamada, K. & Tomii, K. (2014). Revisiting amino acid substitution matrices for identifying distantly related proteins. Bioinformatics. 30, 317–325.
6. Oda, T., Lim, K. & Tomii, K. (2017). Simple adjustment of the sequence weight algorithm remarkably enhances PSI–BLAST performance. BMC Bioinformatics. 18, 288.
7. Nakamura, T., Yamada, K.D., Tomii, K. & Katoh, H. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 34, 2490–2492.
8. Boratyn, G.M., Schaffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J. & Madden, T.L. (2012). Domain enhanced lookup time accelerated BLAST. Biol Direct. 7, 12.
9. Remmert, M., Biegert, A., Hauser, A. & Soding, J. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment. Nat Methods. 9, 173–175.
10. Adhikari, B. & Cheng, J. (2018). CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC Bioinformatics. 19, 22.
11. Fukuda, H. & Tomii, K. (2014). DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. BMC Bioinformatics. 21, 10.
12. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D.T., Silver, D., Kavukcuoglu, K. & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. Nature. 577, 706–710.

13. Zhang, Y & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucl Acid Res. 33, 2302–2309.
14. Luthy, R., Bowie, J.U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature. 356, 83–85.
15. Yang, Y. & Zhou, Y. (2008). Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci. 17, 1212–1219.
16. Hura, G.L., Budworth, H., Dyer, K.N., Rambo, R.P., Hammel, M., McMurray, C.T. & Tainer, J.A. (2013). Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J. 105, 962–974.

ropius0

ROPIUS0: Restraint-Oriented Protocol for Inference and Understanding of protein Structures

M. Margelevičius[1]

[1] – Institute of Biotechnology, Life Sciences Center, Vilnius University
mindaugas.margelevicius@bti.vu.lt

Key: *Auto:N; CASP_serv:Y; Templ:Y; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

Protein structure prediction has reached a level where structural models obtained using de-novo modeling can exhibit similar accuracy to those obtained by homology modeling, the most reliable approach when structurally similar templates can be identified by sequence homology. This effectively means that high accuracy of models can be achieved even in the absence of structural templates. With wide applicability in mind, we developed an initial version of a versatile protocol ROPIUS0 for protein structure prediction, both de novo and by homology. Its unified framework is based on predicting the distributions of distances between residues and estimating the accuracy of structural models.

Methods

The initial objective of ROPIUS0 is to distinguish between two modeling approaches. If available templates for the target sequence exist and are identifiable, homology modeling is used. Otherwise, de-novo structure prediction takes place.

Templates for homology modeling are searched for exclusively using the profile-profile alignment method COMER[1]. An informative and high-quality profile (see below) constructed for the target sequence and the abundance of statistically significant target-template alignments trigger a structure-based procedure for template selection and alignment adjustment. According to this procedure, the most reliable alignment regions between the target and a template result from high consistency between the alignments of all target-template pairs and the structural alignments[2] between the template and the other templates. Templates with extensive reliably aligned regions constitute primary candidates for protein structure modeling. Modeling was performed using RosettaCM[3] during the CASP14 season.

The absence of templates prompts the need for de-novo modeling. Constraints for modeling are generated using a residual encoder-decoder convolutional neural network[4] (REDCNN) with transposed convolution operations on the decoder side and a total of more than 35 million trainable parameters. The REDCNN was trained on a diverse set of 2001 high-resolution protein structures sharing less than 20% sequence identity to predict the distributions of distances between the CB atoms of the sequence under consideration. Two-dimensional input data for the REDCNN are generated solely by the COMER software from a multiple sequence alignment (MSA). The input consists of 519 channels, including the cross-covariance matrix between two positions.

The accuracy of the REDCNN output and the quality of the COMER profile depend on the information content of the input MSA. For obtaining an informative MSA, UniRef, metagenomics and viruses sequence databases are searched with the query sequence using the HHblits[5] and HMMER3[6] software, and their search results are optionally combined.

Distance distributions predicted by the REDCNN can be used as constraints for de-novo modeling. We took a different approach, though. Because of lack of time to tune protein structure modeling software to generate models of desired quality from constraints (preparation for the CASP14 season began in March 2020), we used the REDCNN to estimate the accuracy of structural models and rank CASP-hosted server predictions.

Given the structural model, its accuracy is estimated by evaluating how close the distances between CB atoms (CA for Gly) in the model match those predicted by the REDCNN. Confidence in estimates is gained by employing three REDCNN models, each trained independently for at least 300 epochs. Since confidence drops for inputs corresponding to target sequences for which few homologous sequences can be detected, two sets of settings for the REDCNNs apply. They represent the situations of abundance and scarcity of homologous sequences and differ only in the thresholds of distance and prediction probability at which the output of the REDCNNs is considered. Local and global structural model accuracy estimates result from combining the processed output of the three REDCNNs with four Rosetta energy terms. Estimates obtained in this way were used to rank models obtained by homology modeling in the second half of the CASP14 season and were also submitted in the category of model accuracy estimation. The combination of the outputs from the three REDCNNs was submitted in the category of contact and distance prediction.

Predictions in the topology (free modeling) category, when templates were unavailable, were obtained using the following procedure. The CASP-hosted server predictions were ranked by estimating their accuracy, using the REDCNNs as described above. The top five to eight models with the highest estimated accuracy were refined using the Rosetta software[7], and the accuracy of the resulting refined models was estimated again using the same algorithm as described above. The submission included models with the highest estimated accuracy.

Funding

This work was supported by the European Regional Development Fund [01.2.2-LMT-K-718-01-0028] under a grant agreement with the Research Council of Lithuania (LMTLT).

Availability

The REDCNN models and the ROPIUS0 protocol are not yet publicly available. Other software is available as indicated in the references.

1. Margelevičius,M. (2020). COMER2: GPU-accelerated sensitive and specific homology searches. *Bioinformatics* **36**, 3570-3572.
2. Holm,L. (2019) Benchmarking fold detection by DaliLite v.5. *Bioinformatics* **35**, 5326-5327.
3. Song,Y., DiMaio,F., Wang,R.Y-R., Kim,D., Miles,C., Brunette,T., Thompson,J. & Baker,D. (2013) High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735-1742.
4. Badrinarayanan,V., Kendall,A. & Cipolla,R. (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481-2495.
5. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173-175.

6. Eddy,S. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol*. **7**, e1002195.
7. Conway,P., Tyka,M.D., DiMaio,F., Konerding,D.E. & Baker,D. (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci*. **23**, 47-55.

ropius0QA

Ranking CASP-hosted server predictions by the accuracy estimation module of the ROPIUS0 protocol

M. Margelevičius[1]

[1] – Institute of Biotechnology, Life Sciences Center, Vilnius University
mindaugas.margelevicius@bti.vu.lt

Key: *Auto:N; CASP_serv:Y; Templ:N; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

The predictions of the ropius0QA group in the category of tertiary structure prediction were based on ranking CASP-hosted server predictions. Protein structure modeling was not used. The ropius0QA group was registered to test the performance of the model accuracy estimation module of the ROPIUS0 protocol for protein structure modeling. We, therefore, refer the reader to the extended abstract of the ropius0 group for a more detailed description of the ROPIUS0 protocol. This abstract provides a summary of the model accuracy estimation module.

Methods

The estimated accuracy of a structural model follows from the correspondence between CB (CA for Gly) atom distances in the model and predicted distances. Distances are predicted using a residual encoder-decoder convolutional neural network[1] (REDCNN) trained on 2001 high-resolution protein structures sharing less than 20% sequence identity. Input for the REDCNN is generated using the COMER software[2] and includes the cross-covariance matrix for each pair of positions of the target sequence.

Accuracy estimates result from processing the predictions of three independently trained REDCNN models and combining them with four Rosetta3 energy terms calculated for the structural model. The accuracy is estimated based only on those pairs of atoms for which distances are predicted to be less than a predefined threshold with probability greater than a certain threshold. These thresholds were determined on a small subset of CASP13 targets for two scenarios of abundance and scarcity of homologous sequences for the target sequence.

Availability

The REDCNN models are not yet publicly available. Other software is available as indicated in the references.

Funding

This work was supported by the European Regional Development Fund [01.2.2-LMT-K-718-01-0028] under a grant agreement with the Research Council of Lithuania (LMTLT).

1. Badrinarayanan,V., Kendall,A. & Cipolla,R. (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481-2495.
2. Margelevičius,M. (2020). COMER2: GPU-accelerated sensitive and specific homology searches. *Bioinformatics* **36**, 3570-3572.

3. Conway,P., Tyka,M.D., DiMaio,F., Konerding,D.E. & Baker,D. (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci*. **23**, 47-55.

Rostlab

EMBER: Predicting inter-residue distances using novel sequence embeddings

Konstantin Weißenow & Burkhard Rost
Technical University Munich
k.weissenow@tum.de

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

The recent success of Deep Learning (DL) methods in structural bioinformatics has improved the quality of predicted structures significantly. However, the vast majority of state-of-the-art predictors still rely on evolutionary information captured by multiple sequence alignments (MSAs), making structures of proteins with few evolutionary relatives tough to predict. Additionally, creating high-quality MSAs is not trivial: the parameters for the alignment process need to be chosen on an individual basis in order to add enough, yet diverse sequences. This is done with the goal of obtaining a rich set of sequences that model structural constraints, whilst avoiding the inclusion of sequences with diverging structure.

Methods

We present EMBER (EMBedding-based inter-residue distance predictor), a novel DL method used to predict inter-residue distance maps. EMBER enriches the traditional MSA based input by sequence embeddings, represented by the hidden states of Natural Language Processing (NLP) systems such as BERT[1] and ELMo[2] trained on protein sequence sets.

We use deep dilated residual convolutional networks with many layers, similar to Alphafold[3] and ProSPr[4]. We also used the previously successful approach of training on crops of 64x64 residues instead of full samples. This enabled data augmentation and allowed more efficient mini-batching. Evolutionary information from the MSA is represented as plmDCA[5] parameters obtained by CCMpred[6]. In addition to evolutionary information and embeddings, inputs include: the relative position of the crop w.r.t. the overall sequence, the normalized length of the sample and the log-normalized number of effective sequences as inputs. The latter is primarily intended to allow the model to learn how to weigh embeddings vs. DCA constraints based on alignment quality. Our training and validation sets were based on ProteinNet12[7], but we also added the available samples of the free-modeling category from CASP13 for further validation. Similar to other recent methods, we favored the prediction of distance probability distributions instead of binary contacts by using 42 bins representing distance intervals between 2 and 22 Angstrom.

Since EMBER was developed during CASP14, we submitted predictions from multiple models, which were based on slightly different input combinations. For some of the samples with very sparse MSAs, we submitted predictions from models trained exclusively on embeddings and without evolutionary information.

Availability

EMBER is still under development and will be released at a later point alongside a manuscript.

Acknowledgements

We explicitly want to thank the HPC facilities of the University of Luxembourg[8] (see https://hpc.uni.lu), Ian Fisk from the Flatiron Institute (see https://www.simonsfoundation.org/flatiron/) and OpenPower@TUM (see https://openpower.ucc.in.tum.de/) for their invaluable support during the development of EMBER.

1. Devlin,J., Chang,M., Lee,K. & Toutanova,K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL].
2. Peters,M.E., Neumann,M., Iyyer,M., Gardner,M., Clark,C., Lee,K. & Zettlemoyer,L. (2017). Deep contextualized word representations. arXiv:1802.05365 [cs.CL].
3. Senior,A.W., Evans,R., Jumper,J. et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710.
4. Billings,W.M., Hedelius,B., Millecam,T., Wingate,D. & Corte,D.D. (2019). ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. bioRxiv 830273; doi: https://doi.org/10.1101/830273.
5. Ekeberg,M., Hartonen,T., & Aurell,E. (2014). Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. Journal of Computational Physics. 276. 341-356
6. Seemayer,S., Gruber,M. & Söding,J. (2014). CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 30, Issue 21, 3128-3130
7. AlQuraishi,M. (2019). ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics 20, 311.
8. Varrette,S., Bouvry,P., Cartiaux,H. & Georgatos,F. (2014). Management of an Academic HPC Cluster: The UL Experience. Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014). 959-967

SASHAN, UOSHAN, KUHHAN (QA, TS)

Estimation of protein model accuracy by the predicted features, torsion potential, templates and clustering algorithm in CASP14

Kun-Sop Han
Center of Natural Science, University of Science, Unjong-District, Pyongyang, DPR Korea
hks1981@star-co.net.kp

We participated in QA category of CASP14 with 3 methods. All 3 methods first calculate local scores and then obtain global score from local scores, which is different from 2 methods[1, 2, 3] in CASP13 where global and local score are obtained separately. The first method $\mathrm{SART_{UP}}$ is an updated version of single model-based EMA method SART[1, 3]. The $\mathrm{SART_{UP}}$ is mainly based on the predicted features. The second method is $\mathrm{SARTclust_{UP}}$. The procedure of calculating clustering-based local score is the same as the one in CASP13. However, the performance of it is better than the one in CASP13. The third method $\mathrm{SART_{3D}}$ combines $\mathrm{SART_{UP}\_L}$ with local score based on comparison of the interested model with 3D protein models generated by MODELLER[7]. 36063 CASP9 and CASP11 server models are used as training set. For TS category, only one protein model is submitted. Selection of best protein model and assignment of local deviation to it is based on $\mathrm{SARTclust_{UP}}$.

Methods

1. Estimation of model accuracy

Single model-based score: $\mathrm{SART_{UP}}$ **.** We propose single model-based score $\mathrm{SART_{UP}}$ (group name: SASHAN), an updated version of single model-based EMA method SART[1, 3]. The local score $\mathrm{SART_{UP}\_L}$ is based on regression between S-score calculated from true distance deviation and 7 terms calculated from the sphere (radius 12 Å) centered on residue of interest. The terms include 6 agreement terms and one torsion potential-related term. The agreement terms are based on comparison of predicted values with real ones of several features, including secondary structure[4], solvent accessibility[4], residue contact and torsion angle.

Single model-based global score $\mathrm{SART_{UP}\_G}$ is obtained from $\mathrm{SART_{UP}\_L}$ by the following equation[5].

$$\mathrm{SART_{UP}\_G} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+(\frac{SART_{UP}\_L_i}{c})^2},$$

where, L is length of target sequence.

Clustering-based score: $\mathrm{SARTclust_{UP}}$. The procedure of calculating clustering-based local score $\mathrm{SARTclust_{UP}\_L}$ (group name: UOSHAN) is the same as the one in CASP13[2, 3]. Clustering-based global score $\mathrm{SARTclust_{UP}\_G}$ is obtained from $\mathrm{SARTclust_{UP}\_L}$ by using the same scheme as the $\mathrm{SART_{UP}\_G}$ is calculated.

Another single model-based score: $\mathrm{SART_{3D}}$ We also propose another new single model-based EMA method, $\mathrm{SART_{3D}}$ (group name: KUHHAN), which combines $\mathrm{SART_{UP}}$ with the scores calculated from comparison of the interested model with 3D protein models generated by MODELLER. In details, we identify templates by using PSI-BLAST[6] search on PDB database. The 3D protein models are generated by MODELLER. Then, the 3D models-based local score 3D_L is calculated from superposition of the 3D protein models and models to be assessed by the algorithm similar to $\mathrm{SARTclust_{UP}\_L}$. The final local score $\mathrm{SART_{3D}\_L}$ is an average of SARTup_L and 3D_L if the residue is modeled by MODELLER. If the templates are not identified from PDB database or the residue is not modeled by MODELLER, the SART$_{\mathrm{UP}}$_L is used as the final SART$_{\mathrm{3D}}$_L score. Single-based global score SART$_{\mathrm{3D}}$_G is obtained from local score SART$_{\mathrm{3D}}$_L by using the same scheme as the SART$_{\mathrm{UP}}$_G is calculated.

2. Selection of the 'best' protein model

For TS session (group name: UOSHAN), a protein model with the highest global score is selected by clustering-based EMA method SARTclust$_{\mathrm{UP}}$_G from stage2 dataset. Prediction of local deviation of the selected protein model is based on SARTclust$_{\mathrm{UP}}$_L.

Availability

Manuscript for SART series is in preparation.

1. Han,K-S, et al. (2018) Protein model quality estimation by single model-based method SART in CASP13. CASP13 Abstract, 156-158.
2. Han,K-S, et al. (2018) Protein model quality estimation by clustering-based method SARTclust in CASP13. CASP13 Abstract, 185-186.
3. Cheng,J., et al. (2019) Estimation of model accuracy in CASP13. Proteins. 87:1361−1377.
4. Cheng,J., et al. (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 33, W72−W76.
5. Liu,T., et al. (2016) Benchmarking Deep Networks for Predicting Residue-Specific Quality of Individual Protein Models in CASP11. Scientific Reports, 6, 19301.
6. Altschul,S.F., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
7. Sali,A. & Blundell,T.L., (1993) Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815

SASHAN, UOSHAN, KUHHAN (RR)

Prediction of residue conact and distance using deep learning in CASP14

Kun-Sop Han, Hyok Jang
Center of Natural Science, University of Science, Unjong-District, Pyongyang, DPR Korea
hks1981@star-co.net.kp

We participated in RR category of CASP14 with 3 methods. The first method (RRcon) predicts only residue contacts using deep convolutional network with 25 convolutional layers. The 497 input features and 1 output are derived from every pair of residues for contact prediction. The second method (RRdis) tries to predict residue-residue distance using deep convolutional network with 49 convolutional layers. In this case, 501 input features and 10 outputs are used for each pair of residues. The last one (RRmeta) produces prediction of contact and distance on the basis of comparison of RR server predictions with top TS server models selected by our EMA method SARTclust$_{UP}$.

Methods

RRcon: contact prediction method: RRcon (group name: KUHHAN) is a new deep learning-based contact prediction method. Multiple sequence alignment (MSA) is a critical source for prediction of residue contact and distance in proteins. We obtain MSAs by PSI-BLAST[1] search on Uniref50 database. For a target sequence of length L, total 497 × L × L inputs are extracted: the 441 × L × L covariance matrix derived from the multiple sequence alignments and 56 channels such as sequence profile, mutual information, entropy, secondary structure and solvent accessibility predicted from SCRATCH[2], and sequence separation.

The L × L outputs are derived as followings. If the distance of 2 residues is smaller than 16 Å, the corresponding element of the output matrix is calculated by the equation (1). If the distance is bigger than 16 Å, the element is 0.

$$\text{Output} = 10 \times \frac{1}{1+(\frac{d}{d_0})^2}, \; d_0 = 5 \qquad (1)$$

Of 4196 protein chains extracted from PDB, 4074 chains of them are used as training set and 122 chains as validation set. 497 × L × L inputs and L × L outputs are trained by convolutional neural networks (CNNs) with 25 convolutional layers.

RRdis: distance prediction method*:* For prediction of distance (group name: SASHAN), in addition to the features used for contact prediction, 2 predicted torsion angles are added, resulting in 501 channels. Output of the prediction model is 10 × L × L. The convolutional neural network with 49 convolutional layers is used for training of 5184 protein chains.

RRmeta: Server-based distance prediction: We propose a new server-based contact and distance prediction method (group name: UOSHAN). In this method, all TS server protein models are scored based on our EMA method SARTclust$_{UP}$[3] and 5 top-scoring protein models are selected. Then, all RR server predictions which made the distance prediction are compared with the contacts and distances of 5 top-scoring protein models. We choose top (3 ~ 5) RR server predictions which show the good performance[4] when comparing with 5 top-scoring protein models. Lastly, these top RR server predictions are averaged to produce the final prediction (RRmeta).

1. Altschul,S.F., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.
2. Cheng,J., et al. (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 33, W72−W76.
3. Han,K-S, et al. (2018) Protein model quality estimation by clustering-based method SARTclust in CASP13. CASP13 Abstract, 185-186.
4. Xu,J., et al. (2019) Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins. 87:1069−1081.

SBROD

Evaluation of coarse-grained scoring function SBROD in the QA category of CASP14

S. Grudinin[1] and M. Karasikov[2]

[1] - Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France, [2] - Department of Computer Science, ETH Zurich, Zurich, 8092, Switzerland

Sergei.Grudinin@inria.fr

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

Protein quality assessment (QA) is a crucial element of protein structure prediction, a fundamental and yet open problem in structural bioinformatics. QA aims at ranking predicted protein models to select the best candidates. Although consensus-model QA methods often outperform single-model ones, their performance substantially depends on the pool of available candidates. This makes single-model QA methods a particularly important research topic since these usually assist when sampling the candidate models. It is especially interesting to develop novel coarse-grained single-model methods that only assess positions of the backbone model atoms, as in the case of our method SBROD[1].

Methods

The SBROD (Smooth Backbone-Reliant Orientation-Dependent) method uses only the backbone protein conformation, and hence it can be applied to scoring coarse-grained protein models. We derived the SBROD scoring function from a training set of protein models, which were the server submissions from the previous CASP rounds. The SBROD scoring function is composed of four terms related to different structural features. These are the relative residue-residue orientations, the contacts between the backbone atoms, the hydrogen bonds, and the solvent-solvate interactions. The model is then trained using linear ridge regression to predict the global GDT-TS scores. The obtained scoring function is smooth with respect to atomic coordinates and thus is potentially applicable to continuous gradient-based optimization of protein conformations. Furthermore, it can also be used for coarse-grained protein modeling and computational protein design.

Results

For the original paper, we evaluated SBROD on diverse datasets (CASP11, CASP12, and MOULDER) and proved that it achieved state-of-the-art performance among single-model QA methods. For the CASP13 exercise, the server was not fully ready, and we had to adjust the pipeline on the fly. In the CASP14 exercise, we applied SBROD to the QA category of targets using a fully automated server. We should also mention that the method is sufficiently fast, especially when a large pool of models is to be evaluated. Indeed, SBROD was among the fastest servers in the QA category of CASP14.

Availability

The standalone application implemented in C++ and Python is freely available at https://team.inria.fr/nano-d/software/sbrod/ and https://gitlab.inria.fr/grudinin/sbrod for Linux, MacOS, and Windows.

1. Karasikov,M., Pages,G., & Grudinin,S. (2019). Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics, 35(16), 2801-2808.

Seok-assembly, Seok (Assembly)

Modeling oligomeric proteins by GALAXY in CASP14

Taeyong Park, Hyeonuk Woo, Jinsol Yang, Sohee Kwon, and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea
chaok@snu.ac.kr

Key: *Auto:N; CASP_serv:N; Templ:Y; Fragm:Y9; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y*

Seok-assembly is an automatic structure prediction server for oligomeric protein targets. Human intervention was made only when stoichiometry of target did not correspond to A_n or A_1B_1 or when additional information was provided. Human group Seok submitted structures that were manually predicted using GALAXY programs[1].

Methods

Two newly developed methods, GalaxyHomomer2 and GalaxyHeteromer, were tested on oligomeric protein targets in CASP14.

GalaxyHomomer2 is an upgraded version of GalaxyHomomer[2], which performs automatic homo-oligomer modeling by template-based and *ab initio* docking depending on availability of template. In GalaxyHomomer2, the tertiary structure prediction pipeline of Seok-server in CASP14 is used for monomer structure prediction. Homo-oligomer structure templates are then selected by rescoring top-ranking proteins by HHsearch[3] based on a monomer template score[4], an in-house target difficulty score, TM-score[5] of monomer structure, and TM-score of oligomer interface. Depending on the scores, homo-oligomer structures are built by superimposing the monomer model on the oligomer template or by restraint-based model-building method GalaxyCassiopeia[2]. If less than five models are generated by template-based modeling due to template unavailability, *ab initio* symmetric docking method GalaxyTongDock_C[6] is used to generate more models to get a total of five models.

GalaxyHeteromer performs automatic hetero-dimer modeling also by template-based and *ab initio* docking depending on availability of hetero-dimer templates. GalaxyHeteromer uses monomer structures generated by the tertiary structure prediction pipeline of Seok-server. Templates to be used for hetero-dimer modeling are then selected from HHsearch high-rankers including monomers and homo-oligomers as well as from a hetero-dimer structure database. The hetero-dimer structure database was prepared by compiling hetero-dimer structures with atomic contacts among protein chains and clustering with CD-HIT[7] and TM-align[5]. Hetero-dimer models generated by superimposing the monomer models on templates are ranked by monomer TM-scores, with additional consideration of the number of clashes, the number of contacting residue pairs, and interface area. If less than five models are generated due to template unavailability, *ab initio* asymmetric docking method GalaxyTongDock_A[6] is used to generate more models to get a total of five models.

When target's stoichiometry did not correspond to A_n or A_1B_1, GalaxyHomomer2 and GalaxyHeteromer were used together. For example, for the server prediction of H1072 (A_2B_2) GalaxyHeteromer was run to predict the structure of A_1B_1, and GalaxyHomomer was run to predict the structure of A_2 and B_2. A_2B_2 models were then generated by combining A_2, B_2, and A_1B_1 that have compatible interfaces. In the case of H1036 ($A_3B_3C_3$), B_1C_1 is an antibody. This information was used by predicting B_1C_1 structure, but not B_3 or C_3, and selecting only the models whose CDR regions of the antibody lie on the interface.

For human predictions, available information from the literature and human insight were employed for model generation and selection. Monomer models from other servers were also tried to predict oligomer structures.

Availability

GalaxyHomomer, GalaxyTongDock_C, and GalaxyTongDock_A are available as free web servers on the GalaxyWEB page (http://galaxy.seoklab.org).

1. Ko.J, Park.H, Heo.L, Seok.C. GalaxyWEB server for protein structure prediction and refinement. Nucleic Acids Res. 2012;40:W294-7.
2. Baek.M, Park.T, Heo.L, Park.C, Seok.C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. Nucleic Acids Res. 2017;45:W320-W4.
3. Soding.J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:951-60.
4. Ko.J, Park.H, Seok.C. GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. BMC Bioinformatics. 2012;13:198.
5. Zhang.Y, Skolnick.J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33:2302-9.
6. Park.T, Baek.M, Lee.H, Seok.C. GalaxyTongDock: Symmetric and asymmetric ab initio protein-protein docking web server with improved energy parameters. J Comput Chem. 2019.
7. Fu.L, Niu.B, Zhu.Z, Wu.S, Li.W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150-2.

Seok-refine (TS)

A Meta-server Utilizing Refinement and a *De Novo* Structure Prediction Server Performing Global Optimization of a Neural Network Energy Function

Jonghun Won, Sohee Kwon, Minkyung Baek, and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea
chaok@snu.ac.kr

Key: *Auto:Y; CASP_serv:Y; Templ:N; Fragm:Y3,9; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y*

Seok-refine submitted predictions for TS targets by refining structures selected among the CASP server models as model 1-3 (a.k.a. meta-server), and by *de novo* structure prediction guided by a neural network energy function as model 4 and 5.

Methods

Model 1-3: meta-server utilizing MSA-based scoring and GalaxyRefine2: All CASP server models are first scored by GalaxyQA, and top 24 models are re-ranked by the predicted distance distribution (CNN model of Seok-server). GalaxyQA, an energy-based, non-consensus model quality assessment method, ranks models based on a knowledge-based potential called KGB[1,2] after local optimization with the GalaxyRefine energy[3]. The top model is further refined using GalaxyRefine2[4] with an additional modification in the structure hybridization step in which the 24 models selected by GalaxyQA are used for hybridization. Three lowest-energy models are finally submitted as model 1-3.

Model 4 & 5: de novo modeling by global optimization of a neural network energy function*:* Models 4 and 5 of Seok-refine are generated by an MSA-free *de novo* protein structure prediction protocol. This protocol employs conformational space annealing (CSA)[5] to globally optimize a neural network energy function designed to mimic physical energy function. Initial pool of conformations is sampled by fragment assembly using Rosetta fragment library. The conformation pool is evolved by gradually decreasing conformational diversity and focusing on deeper energy minima. Torsion angle mutations and torsion angle crossovers between pool conformations and fragment insertions are repeatedly carried out to exploit the combinatorial characteristics of the conformational space. Local energy minimization is performed in the backbone torsion angle space and then in the Cartesian space. The neural network energy function takes amino acid types, atomic distances, and sequence distance as input to estimate per-residue energy. A set of 1,100 non-redundant, small (Nres < 150) monomeric proteins were used to train the parameters of the neural network energy function. Two models are chosen at two different extents of CSA optimization (more optimized: Model 4; less optimized: Model 5), followed by final refinement using GalaxyRefine.[3]

Availability

GalaxyRefine and GalaxyRefine2 are available as free web servers on the GalaxyWEB page (http://galaxy.seoklab.org). A standalone version of GalaxyRefine is also downloadable (http://seoklab.github.io/GalaxyRefine).

1. Heo,L. & Seok,C. A new statistical potential with consideration of solvation effects for protein simulations. *in preparation*.
2. Lee,G.R., Heo,L. & Seok,C. (2018). Simultaneous refinement of inaccurate local regions and overall structure in the CASP12 protein model refinement experiment. *Proteins*. **86**, 168-176.
3. Heo,L., Park,H. & Seok,C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384-W388.
4. Lee,G.R., Won,J., Heo,L. & Seok,C. (2019). GalaxyRefine2: Simultaneous refinement of inaccurate local regions and overall protein structure, Nucleic Acids Res. 47 (W1), W451-W455.
5. Lee,J., Scheraga,H.A. & Rackovsky,S. (1998). New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J Comp Chem.* **18**, 1222-1232.

Seok-server (Refinement) / Seok (Refinement)

Protein structure refinement by GALAXY in CASP14

Jonghun Won and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea
chaok@snu.ac.kr

Key: *Auto:Y; CASP_serv:N; Templ:N; Fragm:Yv; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y / Auto:N; CASP_serv:N; Templ:N; Fragm:Yv; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:Y*

In the CASP14 refinement experiment, we tested an automatic refinement method combining a physics-based conformational sampling and a neural network energy function that mimics physical energy function. Several geometric move sets such as anisotropic normal mode perturbation, secondary structure perturbation, and structure hybridization were applied iteratively for diverse conformational sampling.[1] Final models were selected based on a neural network energy function that was trained on a large set of native and non-native structures. We also tested a human prediction method in which more aggressive sampling strategies such as relieving restraints and introducing strong secondary structure perturbations were tried on manually selected targets to overcome local energy barriers of initial structures.

Methods

The overall procedure of the automatic server protocol follows that of GalaxyRefine2[1] except that a neural network energy function was used for final model selection starting from R1057. First, residue-level error estimation of the initial structure was performed to detect unreliable local regions by three different measures: fluctuation during short MD relaxation, consistency with fragment library, and PSSM-based score against putative structural templates. Diverse structures were then generated by conformational move sets such as loop modeling, anisotropic normal mode perturbation, structure hybridization, secondary structure perturbation, and sidechain perturbation. The regions predicted to be inaccurate were sampled more frequently than other regions. The generated structures were next subject to 3- or 1.2-ps molecular dynamics relaxations depending on the magnitudes of structural changes. The energy function employed for the relaxation was identical to that used in the CASP13 refinement protocol, in that Lorentzian function is used for restraints instead of harmonic function when the initial GDT_HA is less than 60 or unknown. Low-energy structures were selected and used as initial structures for the next conformational sampling round. After iterating this procedure, five models were selected by re-ranking all sampled conformations with a newly developed neural network energy function, trained on conformations generated by GalaxyRefine2[1] for a set of one thousand non-redundant proteins.

For human predictions, a more aggressive sampling that uses no restraints in the structure pool update stage for the next round of iteration was also attempted. A large degree of perturbations to relative orientations of secondary structure segments was conducted for R1029 and R1033 which were expected to have very low initial model accuracy. Disulfide bridges were considered when sulfur atoms of two cysteine residues were close. Unreliable loop regions were detected by local quality assessment assisted by human and were subject to intensive loop modeling by using GalaxyLoop.[2] For target R1902-D3, dimer environment was considered since C-terminal structure seemed unstable without inter-molecular interactions.

Availability
GalaxyRefine2 is freely available at http://galaxy.seoklab.org.

1. Lee,G.R., Won,J., Heo,L. & Seok,C. (2019). GalaxyRefine2: Simultaneous refinement of inaccurate local regions and overall protein structure, Nucleic Acids Res. 47 (W1), W451-W455.
2. Lee,G.R., Park,H., Heo,L & Seok,C. (2014). Protein loop modeling using a new hybrid energy function and its application to modeling in inaccurate structural environments. PLoS ONE 9 (11), e113811.

Seok-server (TS)

GALAXY in CASP14: Automated Protein Tertiary Structure Prediction

Jonghun Won, Jinsol Yang, Sohee Kwon, Taeyong Park, Hyeonuk Woo, Minjae Park,
Beomchang Kang, Sangwoo Park, Hwan Won Chung, Katsuhito Inui, Seho Lee, Hakjean Kim,
Jayoon Choi, Nuri Jung, Hyewon Chung, Minkyung Baek, and Chaok Seok

Department of Chemistry, Seoul National University, Seoul 08826, Republic of Korea
chaok@snu.ac.kr

Key: *Auto:Y; CASP_serv:N; Templ:N; Fragm:Y9; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Seok-server performed fully automated protein tertiary structure predictions for TS targets. The server utilized several deep learning-based tools, such as prediction of residue distance distributions and estimation of model accuracy in addition to template-based modeling employed in the previous CASPs.[1]

Methods

The protein tertiary structure prediction pipeline of Seok-server consists of the following steps: (1) modeling unit detection, (2-1) multiple sequence alignment (MSA) generation and distance-histogram prediction, (2-2) template search and sequence alignment, (3) structure building, refinement, and model selection, (4) prediction of domain orientation, and (5) final optimization.

For each target sequence, modeling units are detected by GalaxyDom[2] which runs HHsearch[3] against SCOP70[4] and PDB70. For each modeling unit, distance distributions of residue pairs are first estimated by a convolutional neural network (CNN) model similar to that of AlphaFold[5] from multiple sequence alignments generated by HHblits[6] on Uniclust[7] and BFD, a metagenome sequence database.[8] Then torsion-based conformational space annealing (CSA)[9] is performed to globally optimize initial conformations, which are randomly generated based on torsion prediction by the CNN model. The optimized models are then refined using GalaxyRefine[10] with restraints from predicted distance distributions. At the same time, HHsearch[3] and map_align[11] on PDB70 are performed for template search. Templates are selected by re-ranking the detected proteins using the scores of the search methods and a target difficulty score estimated by machine learning. Tertiary structures are built from sequence alignment between a given target and template(s) by PROMALS3D.[12] In this step, 24 models are constructed by short VTFM MD simulations with template-driven restraints and the CHARMM22 force field followed by short MD relaxations after repetitive side-chain perturbations.[10] The models are then refined using GalaxyRefine2.[13] The refined models from two different methods are ranked by a random forest classification model. For targets with multiple modeling units, orientations between the units are optimized by perturbing torsion angles of the linkers connecting the units to satisfy predicted distance distributions. Final models are subject to optimization in full-atom topology to improve stereochemical properties.

Availability

GalaxyTBM, GalaxyRefine, and GalaxyRefine2 are available as free web servers on the GalaxyWEB page (http://galaxy.seoklab.org). A standalone version of GalaxyRefine is also downloadable (http://seoklab.github.io/GalaxyRefine).

1. Ko,J., Park,H. & Seok,C. (2012). GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. *BMC Bioinformatics* **13**, 198.
2. Choe,K., Heo,L., Ko,J. & Seok,C. GalaxyDom: a method to detect modeling units for protein structure prediction. *submitted.*
3. Söding,J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.
4. Fox,N.K., Brenner,S.E. & Chandonia,J.M. (2013). SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acid Res*. **42**, D304-D309.
5. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., ... & Hassabis,D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710.
6. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **9**, 173-175.
7. Mirdita,M., Driesch,L., Galiez,C., Martin,M.J., Söding,J. & Steinegger,M. (2016). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acid Res*. **45**, D170-D176.
8. Steinegger,M., Mirdita,M. & Söding,J. (2019). Protein-level assembly increases protein sequence recovery from metagenomics samples manifold. *Nature Methods* **16**, 603-606.
9. Lee,J., Scheraga,H.A. & Rackovsky,S. (1998). New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *J Comp Chem.* **18**, 1222-1232.
10. Heo,L., Park,H. & Seok,C. (2013). GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384-W388.
11. Ovchinnikov,S., Park,H., Varghese,N., Huang,P.S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. & Baker, D. (2017). Protein structure determination using metagenome sequence data. *Science* **355**, 294-298.
12. Pei,J., Kim,B. & Grishin,N. (2008). PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res*. **36**, 2295-2300.
13. Lee,G.R., Won,J., Heo,L. & Seok,C. (2019). GalaxyRefine2: Simultaneous refinement of inaccurate local regions and overall protein structure, Nucleic Acids Res. 47 (W1), W451-W455.

ShanghaiTech

Contact map prediction for proteins with less effective sequence homologs

Xuyang Liu[1,2], Lei Jin[3], Shenghua Gao[3] and Suwen Zhao[1,2]

[1] - *iHuman Institute, ShanghaiTech University, Shanghai, 201210 China,* [2] - *School of School of Life Science and Technology, ShanghaiTech University, Shanghai, China,* [3] - *School of Information Science and Technology (SIST), ShanghaiTech University, Shanghai, China.*

gaoshh@shanghaitech.edu.cn and zhaosw@shanghaitech.edu.cn

A new data augmentation method and consistency learning were used in our contact map prediction which aims to improve the contact map precision for proteins with small Nf (normalized number of effective sequence) in multiple sequence alignment (MSA). MSA dropouts were generated from MSAs with big Nf, and were used as data augmentation. Features were learned from both original MSAs and MSA dropouts by a network branch called consistency learning. Improved contact predictions were realized for proteins with Nf < 40.

Methods

Multiple sequence alignment generation and dropout: The MSAs were generated using the Zhang lab's DeepMSA program[1]. DeepMSA is a MSA generation pipeline by combining HH-suite and HMMER program to search homology, which can be divided into three stages in databases Uniclust, UniRef and Metaclust respectively. In this work, we generated MSAs using databases Uniclust30_2018_08, UniRef90 in December 2019 and Metaclust50_2018_08. For each protein sequence in training and test datasets, the default search parameters in DeepMSA were used.

The sampled MSAs were also used for input feature generation. We randomly selected 20 homolog subsets from each original MSA and called them MSA dropouts. The sampled MSAs have small Nf values (Nf < 20) and were used as data argumentation.

Features generation: Input features are the same with the features used in RaptorX-contact[2].

Consistency learning: We observed that features learned with small Nf MSA inputs were not discriminative enough, i.e., the network can produce a much higher loss when given a hard case (which usually has small Nf) as input. What's more, small Nf MSA inputs might also mislead the network. To handle the above challenge, we encouraged our network to produce similar features for MSA inputs with big Nf (Nf >128) and small Nf. That is, we added a consistency loss beyond the standard cross-entropy loss. Such operation is commonly adopted in teacher-student networks for knowledge distillation. This consistency loss further helps us to learn a better feature.

Availability

Our standalone package will be available as soon as our paper is published.

Acknowledgements

We thank for the helpful discussions with Jinbo Xu and Sheng Wang. We also acknowledge the use of several tools and datasets (DeepMSA, HHBlits, JackHmmer, TGT_Package, CCMPred, Uniclust, UniRef, Metaclust, PDB).

1. Zhang, C.; Zheng, W.; Mortuza, S. M.; Li, Y.; Zhang, Y., DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020, 36 (7), 2105-2112.
2. Xu, J., Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A 2019, 116 (34), 16856-16865.

Shen-CDeep

Protein contact map prediction based on SE-net

Pei-Dong Zhang[1,2], Shi-Hao Feng[1,2], and Hong-Bin Shen[1,2]

[1] - *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University,* [2] - *Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China*

hbshen@sjtu.edu.cn

Recent CASP experiments have shown promising results of machine learning-based models, especially the deep learning-based models, for enhancing the protein residue contact map prediction[1-4]. In CASP 14, we have developed the SE-net based models for this task. Given a target protein sequence, we first searched the available sequence databases to generate multiple sequence alignment (MSA) using hhblits, hmmsearch and jackhammer[5-6]. Our prediction model was built based on Squeeze-and-Excitation Network (SE-net)[7]. It is composed of three SE-net with 29 residual[8] convolutional layers. Each network takes as input the 441 dimension intermediate results of PSICOV[9], EVFOLD[10], and CCMPRED[11]. Our local tests show that there are strong complementarities between these three methods and it has been shown that simultaneously employing them will lead to better performance. SE-net[7] is first proposed in image processing field to explicitly modeling the interdependencies between the channels of the convolutional features. We expect that through taking into consideration the relationship between the channels of the co-evolutional features, the performance of the deep learning model will be further improved. Thus, we employ the channel-level self-attention mechanism and expansive convolution technique in SE-net. The three SE-nets are trained separately. The outputs of them are further concatenated in channel and serve as the input of a fully connection layer, which gives the final prediction of the CASP14 model. Preliminary tests on 39 CASP13 targets, this SE-net-based deep model can achieve an overall accuracy of 71.7% on the top L/5 long-range contacts.

1. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology 2017;13(1):e1005324.
2. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018, 34(19):3308-3315.
3. Li Y, Hu J, Zhang C, Yu DJ, Zhang Y, ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks, Bioinformatics, 2019, 35: 4647-4655.
4. Yang J, Jin Q-Y, Zhang B, Shen H-B. R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. Bioinformatics 2016;32(16):2435-2443.
5. REMMERT M, BIEGERT A, HAUSER A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 2012, 9(2): 173-175.
6. SöDING J. Protein homology detection by HMM-HMM comparison. Bioinformatics, 2005, 21(7): 951-960.
7. Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
8. HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

9. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2011;28(2):184-190.
10. Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC bioinformatics 2014;15(1):85.
11. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.

Reduction of Local Steric Clash and Improved Phi/psi Energies by Assembly of Refined Fragments onto Server Models as Templates

D. Shortle[1]
[1]*Department of Biological Chemistry. The Johns Hopkins University School of Medicine*
dshortl1@jhmi.edu

Key: *Auto:N; CASP_serv:Y; Templ:Y; Fragm:Y; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

In his book "Statistical Mechanics of Chain Molecules[1]", Paul Flory emphasized the outsized importance of interactions between nearby monomers in determining the behavior/structure of polymers. To address the often neglected issue of the energy of local interactions between residues separated by less than 6 positions, we employ two strategies: (1) Steric clash is monitored by separate parameters that sum the atom-atom overlap linearly (i.e., not raised to a polynomial) over three different separation windows: namely i to i+1, i to i+2-5, and separations of six or more residues (2) For the remaining non-steric interactions involving changes in Coulombic and dispersion energies with changing phi/psi angles, we employ eight statistical potentials/probabilities[2] for monomers, dimers, and trimers, arguing that in the absence of more physics-based metrics this is the best one can do..

Methods

Our overall approach is to use server models provided by the Prediction Center as templates for fragment assembly, which leads to replacement of most, if not all, chain segments with fragments taken from PDB structures. In the previous CASPs, we used PDB fragments directly, but better results can be achieved by first refining fragments concatenated into full length chains to reduce the local overlap and lower the Ramachandran energies that result from altering the amino acid sequence.

If the server models are in reasonable agreement, we start by randomly recombining a set of 25 server models, yielding hybrid models comprised of roughly equal length segments from 3 different models. If there is little agreement among models, this same random recombination is carried out on an equimolar mixture of server models and refined full-length chains.

Subsequent refinement steps involve a multiplicity of scoring terms of different types. The principal global terms are: (1) the distribution of turn, core, and surface residues in 5 concentric shells of equal volume within an ellipsoid fit to the shape of the model being refined, but given the volume calculated for the folded protein; and (2) an estimate of side-chain solvation of dimers using stretched CB atoms and the propensity of each amino acid in the pair to have the measured number of neighbors in the three different secondary structures. To quantify the energetics of atom pairs we use statistical potentials and local/global probabilities based on 30 atom types and 49 atom types, and as with atom overlap, pair energies are calculated separately for interactions between residue i and i+1, i to i+2 through i+5, and separations of more than 5 residues.

Conformational search is driven by a genetic algorithm with various Monte Carlo moves and steps added. To reduce the problems of local minima and loss of structural diversity inherent to genetic algorithms, three separate scoring functions are constructed from sets of pseudo-energy terms: (1) a selection function determines which MC moves are accepted; (2) a survival function picks out the best conformation in the second half of the completed MC trajectory; and (3) a global survival function determines which conformations among the final 2N population are saved at the end of each generation, typically with rounds of 3 or 4 generations. Typically, scoring functions 2 and 3 include more atom-level terms than function 1. As refinement progresses, emphasis is shifted from predominantly local interactions to predominantly long range interactions.

In summary, the general method described above was used for all TS and refinement predictions, with extensive manual intervention to achieve a balanced reduction in values of the many pseudo-energy terms tracked during refinement, particularly atom overlap.

Results

For approximately half of the targets addressed, the values of local clash and phi/psi energies were reduced to values in the range observed in high resolution PDB structures. This emphasis on local interactions made achieving native-like values for long range clash and atom-atom interactions much more difficult. We have not yet analyzed the accuracy of submitted models.

Availability

All software used in this work was written in C++ by the group leader trying to conform to best programming practices. Since our computer programs were built around an old, proprietary object/template library purchased from RogueWave Software (Windows Version), several obstacles would have to be overcome to make it useable for other groups running Linux. In addition the source code needs extensive re-writing to make it understandable by other programmers. But given these caveats, it will be available to anyone up to the challenge. Preferable would be a collaboration in which central features of our approach are transferred to faster, more user-friendly code.

1. Flory, P.. (1969). Statistical Mechanics of Chain Molecules. Oxford University Press, New York. *Nucleic Acids Res.* **25**, 3389-3402.
2. Shortle, D. (2003) Propensities, probabilities, and the Boltzmann hypothesis. *Protein Science.* ***12***, 1209-1302.

Spider (assembly)

A Novel Statistical Energy Function and Effective Conformational Search Strategy based Protein Complex Structure Prediction

Avdesh Mishra[1], Md Wasi Ul Kabir[2], Md Tamjidul Hoque[2,*]

1 - Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX, USA

2 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA

* thoque@uno.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:Y*

In CASP14, we have used our developed protein structure prediction (PSP) method called 3DIGARS-PSP for the prediction of protein complex structure (or assembly prediction). 3DIGARS-PSP uses an effective statistical energy function, called 3DIGARS, and an advanced search algorithm called KGA. We refer to our assembly prediction method as 3DIGARS-PSP-ASSEMBLY. The 3DIGARS-PSP method employs a memory assisted genetic algorithm (GA) extended from KGA for the conformational sampling of the protein folding process. We design GA using two important operators: memory assisted crossover and mutation. These operators perform the important function of angle rotation and segment translation to assist in careful sampling. We also utilize the propensities of secondary structure and torsion angle to assist the search process. Through the memory assisted GA based sampling that minimizes the statistical energy function, a large-scale ensemble of decoys are generated. Finally, we select the top five models for each CASP14 assembly target by clustering the ensemble of decoys, and consequently, these models are submitted to CASP14.

Methods

The assembly targets have more than one subunit in CASP14, and each subunit has a corresponding fasta sequence. First, we combine the fasta sequences of the subunits by adding 20 Glycine (GLY or G) amino acids in between the fasta sequences to prepare a single fasta sequence. Glycine amino acid is used to combine the fasta sequences of the subunits because of its smallest size of the side chain among 20 standard amino acids. Then we use the I-TASSER[1] tool to obtain the predicted models using the combined fasta sequence. The prediction of the 3D structure of the assembly target starts by initializing some of the chromosomes of the GA population with the Cartesian coordinates of the backbone atoms of the models obtained from I-TASSER. The rest of the chromosomes are filled by single point torsion angle changes (rotation). To make a guided change of the torsion angles (Φ or Ψ), the occurrence frequency of 20 standard amino acids with different Φ-Ψ angle pairs are constructed from the 4,332 high-resolution experimental structures extracted in our previous work[2]. The Φ and Ψ angle range is divided into 120 bins with an interval of 3 degrees, and the frequency of the bins are updated based on the value of the Φ and Ψ angles of every amino acid in the protein to obtain the frequency of distribution of 20 standard amino acids. We further categorize the frequency distributions into zones by looking at the cluster of the frequency values. Then, the most probable torsion angle (namely, $p\Phi$ or $p\Psi$) of the zone is extracted using the roulette wheel selection method, and a random angle around this angle is selected as a new torsion angle.

Moreover, the propensities of secondary structure (SS) types of the amino acids are also extracted from the same experimental structures used above by running the DSSP program to guide the torsion angle rotation. The SS types given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The torsion angle pair and SS types of the amino acids in protein are used to obtain the SS distribution. Later, this distribution of SS is used such that the SS type, which has the largest frequency count, is assigned to the given amino acid having a certain Φ-Ψ angle. Furthermore, the Φ-Ψ angle pairs corresponding to the H and E types are grouped into helix and beta groups and are consequently used to update the Φ or Ψ angle that results in a clash within the structure.

The chromosomes (models) for the next generation of GA are obtained by two different types of structural change operators: *i)* angle rotation, and *ii)* segment translation. The mutation in GA involves torsion angle rotation, and crossover involves segment translation followed by torsion angle rotation at the crossover point. The torsion angle rotation technique is based on the principle of rotation about an arbitrary axis. On the other hand, crossover in GA performs segment translation where all the amino acid indexes that are not SS type E or B are considered as possible crossover points. This is done to avoid random changes in the beta-sheet region and make more appropriately guided change during the mutation operation. The children's structures in the crossover process are generated from two parent structures and a structure with the best fitness saved in the memory[3].

The decoys generated by the conformational change through memory assisted GA guided by the statistical energy function are then converted into the all-atom level by using Oscar-star software[4]. The large-scale pool of decoys are clustered into five different cluster groups, at least 5Å apart among each other based on the average root-mean-square deviation (RMSD). Then, we select the top five models in different clusters based on the 3DIGARS energy score ranking. The subunits of the top five models are further refined using the ModRefiner[5] software. Then, we use the ResQ[6] method to add B-factors to the subunits of the top five models. Finally, the models of the subunits are combined together in the CASP14 assembly format before submission.

Availability

Source code, manual, and example data of 3DIGARS-PSP for Linux are freely available, for non-commercial use, at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Lab, Z. I-Tasser Software, Vol. 2020, pp. http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. Mishra, A. & Hoque, M. T. (2017). Three-Dimensional Ideal Gas Reference State Based Energy Function. Current Bioinformatics 12, 171-180.
3. Hoque, M. T. & Iqbal, S. (2017). Genetic algorithm-based improved sampling for protein structure prediction. International Journal of Bio-Inspired Computation 9, 129-141.
4. Liang, S., Zheng, D., Zhang, C. & Standley, D. M. (2011). Fast and accurate prediction of protein side-chain conformations. Bioinformatics 27, 2913-2914.
5. Xu, D. & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophysical Journal 101, 2525-2534.
6. Yang, J., Wang, Y. & Zhang, Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. Journal of Molecular Biology 428, 693-701.

Spider (ab initio)

A Novel Statistical Energy Function and Effective Conformational Search Strategy based *ab initio* Protein Structure Prediction

Avdesh Mishra[1], Md Wasi Ul Kabir[2], Md Tamjidul Hoque[2,*]

1 - Electrical Engineering and Computer Science, Texas A&M University-Kingsville, Kingsville, TX, USA
2 - Computer Science, University of New Orleans, 2000 Lakeshore Drive, New Orleans, LA 70148, USA
* thoque@uno.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:Y*

In CASP14, we test our proposed novel *ab initio* protein structure prediction (PSP) method, called 3DIGARS-PSP. 3DIGARS-PSP method utilizes an effective statistical energy function, called 3DIGARS, and an advanced search algorithm called KGA. It employs a memory assisted genetic algorithm (GA) derived from KGA to sample the complex energy surface of the protein folding process. To address the critical search process, GA employs two effective operators: memory assisted crossover and mutation, which are decorated with angle rotation and segment translation features. Moreover, propensities of secondary structure and dihedral angle distribution are utilized to guide the conformational search. The GA based sampling that minimizes the statistical energy function generates a large-scale decoy pool. Finally, we collect the top five models for each CASP14 target by clustering the ensemble of decoys and consequently submit these models to CASP14.

Methods

Protein structure is primarily represented by backbone atoms N, Cα, C, and O in 3DIGARS-PSP. We first obtain the predicted models from I-TASSER[1] for each CASP14 targets. We start by initializing some of the chromosomes of the GA population with the Cartesian coordinates of the backbone atoms of the models from I-TASSER. Next, the remaining chromosomes are initialized by single point torsion angle changes (rotation). For an informed change of the torsion angles (Φ or Ψ), we utilize the frequency of occurrence of 20 different amino acids with different Φ-Ψ angle pairs, summarized from the 4,332 high-resolution experimental structures extracted in our previous work[2]. The range of both Φ and Ψ angles for every amino acids are divided into 120 bins with an interval of 3 degrees, and the frequency of the bins are updated based on the value of the Φ and Ψ angles. The frequency distribution obtained for each amino acid is further categorized into zones by looking at the cluster of the frequency values. Then, the roulette wheel selection approach is applied to select the most probable torsion angles (namely, $p\Phi$ or $p\Psi$) belonging to the zone. Next, a random Φ or Ψ (say, $r\Phi$ or $r\Psi$) between $p\Phi$-3 and $p\Phi$ or $p\Psi$ and $p\Psi$+3 is selected, and rotation of the current torsion angle is performed to achieve a new torsion angle, $r\Phi$ or $r\Psi$.

In addition, the change of the torsion angles is further guided by the propensities of secondary structure (SS) types of the amino acids extracted from the 4,332 high-resolution experimental structures by running the DSSP program. The eight different SS types (E, B, H, G, I, T, S, and U) given by DSSP are broadly categorized into four different SS types (H, G, and I = H; E and B = E; T and S = T; and U). The Φ-Ψ angle pair and SS types are used to obtain the index in the SS frequency table and increase the frequency count of the cell in the table by one. Later, the SS type, which has the largest frequency count, is assigned to the given amino acid having a certain Φ-Ψ angle. Additionally, we collect the Φ-Ψ angle pairs belonging to the H and E types and group them into helix and beta groups. We utilize the Φ-Ψ angle pairs belonging to the helix or sheet group to update the Φ or Ψ angle that results in the clash within the structure.

To generate new chromosomes (structural samples) for the next generation of GA, we apply two types of conformational change operators *i)* angle rotation; and *ii)* segment translation. The mutation operation involves phi or psi angle rotation, and crossover operation involves segment translation followed by phi or psi angel rotation at the crossover point. Rotation of phi and psi angles is based on an idea of rotation about an arbitrary axis. For segment translation, a set of possible crossover points are selected based on the secondary structure information. All amino acid indexes except the amino acids belonging to the beta-sheet secondary structure type (either E or B) are considered as possible crossover points. This is done to preserve beta-sheet regions in the structure from random changes during the crossover operation and perform more controlled changes of this region while performing mutation operation. We generate four children structures from two parent structures using the crossover process and a structure with the best fitness saved in the memory[3].

Using the statistical energy function, decoys are generated by minimizing the potential energy using associated memory GA discussed above. Each decoy generated by 3DIGARS-PSP is then converted into the all-atom level by using Oscar-star software[4]. Then a single-model based model quality assessment program Qprob[5], which predicts a model's quality by estimating the error of structural, physiochemical, and energy-based features using probability density distributions, is used to rank the decoys. Next, the MUFOLD-CL[6] method is used to cluster the decoys. Then, we select the top five models in different clusters based on their Qprob rankings. The top five models are further refined using ModRefiner[7] software. Then, we use the ResQ[8] method to add B-factors to the top five models before submission.

Availability

Source code, manual and example data of 3DIGARS-PSP for Linux are freely available to non-commercial use at http://cs.uno.edu/~tamjid/Software/ab_initio/v2/PSP.zip.

1. Lab,Z. I-Tasser Software, Vol. 2020, pp. http://zhanglab.ccmb.med.umich.edu/I-TASSER/.
2. Mishra,A. & Hoque,M.T. (2017). Three-Dimensional Ideal Gas Reference State Based Energy Function. *Current Bioinformatics* **12**, 171-180.
3. Hoque,M.T. & Iqbal,S. (2017). Genetic algorithm-based improved sampling for protein structure prediction. *International Journal of Bio-Inspired Computation* **9**, 129-141.
4. Liang,S., Zheng,D., Zhang,C. & Standley,D.M. (2011). Fast and accurate prediction of protein side-chain conformations. *Bioinformatics* **27**, 2913-2914.
5. Cao,R. & Cheng,J. (2016). Protein single-model quality assessment by feature-based probability density functions. *Scientific Reports* **6**, 23990.
6. Zhang,J. & Xu,D. (2013). Fast algorithm for population-based protein structural model analysis. *Proteomics* **13**, 221-229.
7. Xu,D. & Zhang,Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal* **101**, 2525-2534.
8. Yang,J., Wang,Y. & Zhang,Y. (2016). ResQ: An approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of Molecular Biology* **428**, 693-701.

SSThread

Template-free structure prediction using SSThread 2.2 and AlphaFold's distogram prediction

K. J. Maurice
kevin_maurice@hotmail.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

SSThread[1] works by first predicting the structure of pairs of contacting secondary structure elements derived from a database of known 3D structures. Then core structures are generated by assembling overlapping pair predictions. Loops and side chain conformations are then predicted. The final predictions are then refined. Predictions are scored using a coarse-grained Knowledge Based Potential (KBP), an all-atom KBP and deep learning prediction of several structure properties. Version 2.2 of SSThread uses distogram predictions from AlphaFold[2] and can use sparse NMR data.

Methods

Multiple sequence alignments are generated using a method similar to DeepMSA[3]. The coarse-grained KBP contains a contact term, a backbone torsion angle term, a half-sphere exposure[4] term and a compactness term. The secondary structure, torsion angles and half-sphere exposure are predicted by the deep learning package DeepPPP (unpublished).

The highest performing free modeling method in CASP13 was AlphaFold[5]. AlphaFold uses a deep neural network that predicts distograms (for distance histogram) of distances between the β-carbons of each pair of residues. Distogram predictions are used as one the scoring terms.

Loops are predicted by database search and Cyclic Coordinate Descent[6]. Side chains are predicted using TreePack[7]. A quick energy minimization is carried out using GROMACS[8] with the AMBER03[9] force field. The best predictions are selected using the above mentioned terms as well as the all-atom KBP RWplus[10] with clustering to reduce redundancy among the predictions. The predictions were then refined using GalaxyRefine[11].

SSThread can use sparse NMR data in three forms. DeepPPP can optionally use torsion angle restraints derived from chemical shifts as an input. Residual dipolar couplings can be used by SSThread. Information from ambiguous NOE restraints can be combined with distogram predictions from AlphaFold. For each pair of amino acid types and hydrogen types or group of hydrogen types distograms have been calculated from a set of known 3D structures using a 5 Å distance cutoff between hydrogen atoms. The distograms from NOE restraints are averaged with the distogram prediction of AlphaFold in which the NOE terms are weighted by their reliability according to the equation

$$w = \left(\frac{(1-p_f)f(d_s)}{(1-p_f)f(d_s)+p_f g(d_s)}\right)\left(\frac{\frac{f(d_s)}{g(d_s)}}{\sum_{i=1}^{n}\frac{f(d_{si})}{g(d_{si})}}\right)$$

where p_f is the probability that any given peak is a false positive, $f(d_s)$ is the distribution of distances in sequence for real NOEs, $g(d_s)$ is the distribution of distances in sequence for residues chosen at random and the summation is over all possible NOE restraints from the same peak. The $f(d_s)$ is skewed toward shorter distances compared to the $g(d_s)$.

Availability

The software can be downloaded from www.kjmaurice.com/downloads.html. DeepPPP is available there as well as software that makes AlphaFold's distogram prediction more accessible. SSThread version 2.2 will be released at a future date. DeepPPP and SSThread are free for non-commercial use and the distogram prediction software is open source.

1. Maurice,K.J. (2014). SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. J. Comput. Chem. 35, 644-656.
2. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Zidek,A., Nelson,A.W.R., Bridgland,A., Penedones,H., Petersen,S., Simonyan,K., Crossan,S., Kohli,P., Jones,D.T., Silver,D., Kavukcuoglu,K. & Hassabis,D. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710.
3. Zhang,C., Zheng,W., Mortuz,S.M., Li,L. & Zhang,Y. (2020) DeepMSA: constructing deep multiple sequence alignment to improve contact prediction an fold-recognition for distant-homology proteins. Bioinformatics 36, 2105-2112.
4. Hamelryck,T. (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 59, 38-48.
5. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Zidek,A., Nelson,A.W.R., Bridgland,A., Penedones,H., Petersen,S., Simonyan,K., Crossan,S., Kohli,P., Jones,D.T., Silver,D., Kavukcuoglu,K. & Hassabis,D. (2019) Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins 87, 1141-1148.
6. Canutescu,A.A. & Dunbrack,R.L. Jr. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci. 12, 963-972.
7. Xu,J. & Berger,B. (2006) Fast and accurate algorithms for protein side-chain packing. J. of the ACM 53, 533-557.
8. Berendsen,H.J.C., van der Spoel,R. & van Drunen,R. (1995) GROMACS: A message-passing parallel molecular dynamics implementation. Comp. Phys. Comm. 91, 43-56.
9. Duan,Y., Wu,C., Chowdhury,S., Lee,M.C., Xiong,G., Zhangh,W., Yang,R., Cieplak,P., Luo,R., Lee,T., Caldwell,J., Wang,J. & Kollman,P. (2003) A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phased Quantum Mechanical Calculations. J. Comp. Chem. 24, 1999-2012.
10. Zhang,J. & Zhang,Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One 5, e15386.
11. Heo,L., Park,H. & Seok,C. (2013) GalaxyRefine: Protein structure refinement driven by side-chain repacking. Nucleic Acids Res. 41, W384-W388.

Takeda-Shitaka-Lab

Prediction of Oligomeric Protein Structures based on Template-Based Docking

Yasuomi Kiyota, Shinpei Kobayashi, Yoshiki Harada and Mayuko Takeda-Shitaka

School of Pharmacy, Kitasato University, Tokyo, Japan
shitakam@pharm.kitasato-u.ac.jp

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

We participated in the assembly category of CASP14. We predicted both homo- and hetero-oligomeric protein structures according to the oligomeric state in the CASP14 target list. Our modeling procedure was based on template-based docking method.

Methods

Monomer Selection: We basically used CASP14 server models (Stage 2) as monomer models. We selected high quality monomer models using combined score of ProQ3[1], ProQ3D[2] and VoroMQA[3]. The score was adjusted to pick up high quality models. When we could not obtain high quality server models, we constructed the monomer models by MODELLER[4] based on alignments from HHalign[5].

Oligomeric Template Search: To find reliable oligomeric templates, we carried out two-step template search. Firstly, the oligomeric templates were searched by HHblits[5] against UniRef30 and PDB70 database. Secondly, to search oligomeric templates more widely, we ran PSI-BLAST[6] on PDBaa using HHblits hits as inputs. According to the results of two-step template search and information of biological unit, oligomeric templates were selected.

Oligomeric Model Construction: To construct oligomeric models, we performed template-based docking. We superposed the monomer models onto the oligomeric templates using TM-align[7] or CAB-align[8]. When we could not obtain the oligomeric templates, we used DOCKGROUND[9] database as templates. For the template free docking targets, we used SymDock2[10] to construct the oligomeric models.

Scoring and Refinement of Oligomeric Models: The quality of oligomeric models were assessed by combined score of VoroMQA and SOAP-PP[11]. Clash information at the interface was considered manually. The selected 5 models were refined using MODELLER to remove steric clashes.

1. Uziela,K., Shu,N., Wallner,B. & Elofsson,A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Scientific reports.* **6**, 33509.
2. Karolis,Uziela., David,Menéndez.Hurtado., Nanjiang,Shu., Björn,Wallner. and Arne,Elofsson. (2017). ProQ3D: Improved model quality assessments using Deep Learning. *Bioinformatics.* **33**, 1578-1580.
3. Olechnovič,K., Venclovas,Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins.* **85**, 1131-1145.
4. A,Sali. and T.L.Blundell. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779-815.

5. Steinegger,M., Meier,M., Mirdita,M., Vöhringer,H., Haunsberger,S.J., and Söding,J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *Bioinformatics*. **20**, 473.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.
7. Y,Zhang., J,Skolnick. (2005). TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Res.* **33**, 2302-2309.
8. Genki,Terashi. and Mayuko,Takeda-Shitaka. (2015). CAB-align: a flexible protein structure alignment method based on the residue-residue contact area. *PLoS One.* **10**, e0141440.
9. Kundrotas,P.J., Anishchenko,I., Dauzhenka,T., Kotthoff,I., Mnevets,D., Copeland,M.M., Vakser,I.A. (2018). DOCKGROUND: a comprehensive data resource for modeling of protein complexes. *Protein Sci.* **27**, 172-181.
10. S.S.Roy,Burman., R.A.Yovanno., J.J.Gray. (2019). Flexible backbone assembly and refinement of symmetrical homomeric complexes. *Structure.* **27**, 1041-1051.
11. Dong,G.Q., Fan,H., Schneidman-Duhovny,D., Webb,B., Sali,A. (2013). Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics.* **29**, 3158-3166.

Combining Deep Learning Enhanced Hybrid Potential Energy for Template-Based Modelling

L. Zheng, H. Lan, T. Shen, J. Wu, W. Liu, S. Wang#, J. Huang#
Tencent AI Lab
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

We introduce a new Comparative Modeling (CM) pipeline tFold, which employs a hybrid potential energy function to enhance the original Template-Based Modelling (TBM) energy function with deep learning (DL) based residue-residue C-beta distances constraints. The core ideas adopted in this hybrid potential energy function are (i) completing the missing distance information (i.e., those un-aligned regions in the TBM step) with the de novo predicted distances in the DL model, and (ii) applying additional constraints (i.e., predicted from the DL model) plus the original constraints (i.e., those aligned regions in the TBM step) for the residue-pairs in the CM procedure.

Methods

The tFold pipeline is mainly composed by the following steps: 1) template searching and 3D structure modelling; 2) DL based C-beta distance and distance deviation prediction; 3) hybrid potential energy guided structure refinements of the decoy from step 1; 4) decoys ranking and selection.

Firstly, given a query sequence, the most probable templates are firstly searched using TBM tool CNFpred[1]. The best single template, as well as the query-template sequence alignment, is then subjected to RosettaCM[2] to generate 300 decoys. The best decoy (TBM decoy) is then selected using structure clustering tool Spicker[3].

Secondly, a DL model developed in our group (more details in tFold@RR) is utilized for de novo C-beta distances, orientations and decoy distance deviation prediction. As shown in Figure 1, different from other distance prediction models[4], the model was re-trained to accept 3D structure decoys (generated by RosettaCM) as inputs, which is capable for the distance deviation prediction for each input decoy. The distance deviation of a decoy is defined as the absolute difference of C-beta distances between those extracted from the decoy and the ground truth C-beta distances for each residue pairs. Note that this step could further improve the C-beta distance prediction (more details in tFold-IDT@RR).

Thirdly, starting from the TBM decoy, along with the Rosetta Ref2015 energy[5], a hybrid potential energy function in consideration of the predicted C-beta distance is constructed to guide the TBM decoy refinement following a protocol similar to the 3D structure modelling used in trRosetta[4]. Specifically, in this potential energy function, for each residue-pair we have two different energy forms: (a) if both residues in the pair are in the alignment region, the energy function for this pair is in a harmonic form using the C-beta distance from the decoy as the minimum, with a residue-pair specific constraint factor determined by the distance deviation predicted in the DL model; (b) otherwise, the energy function of the residue pair is in a spline form transformed from the discrete distance probability distribution predicted in the DL model. We generate 50 decoys for the final decoy selection.

Multiple decoys generated in the previous step are then scored and ranked by a quality assessment (QA) model consists of the three popular statistical potential energies: DOPE[6], GOAP[7] and KORP[8]. More details could be found in tFold-IDT@3D.

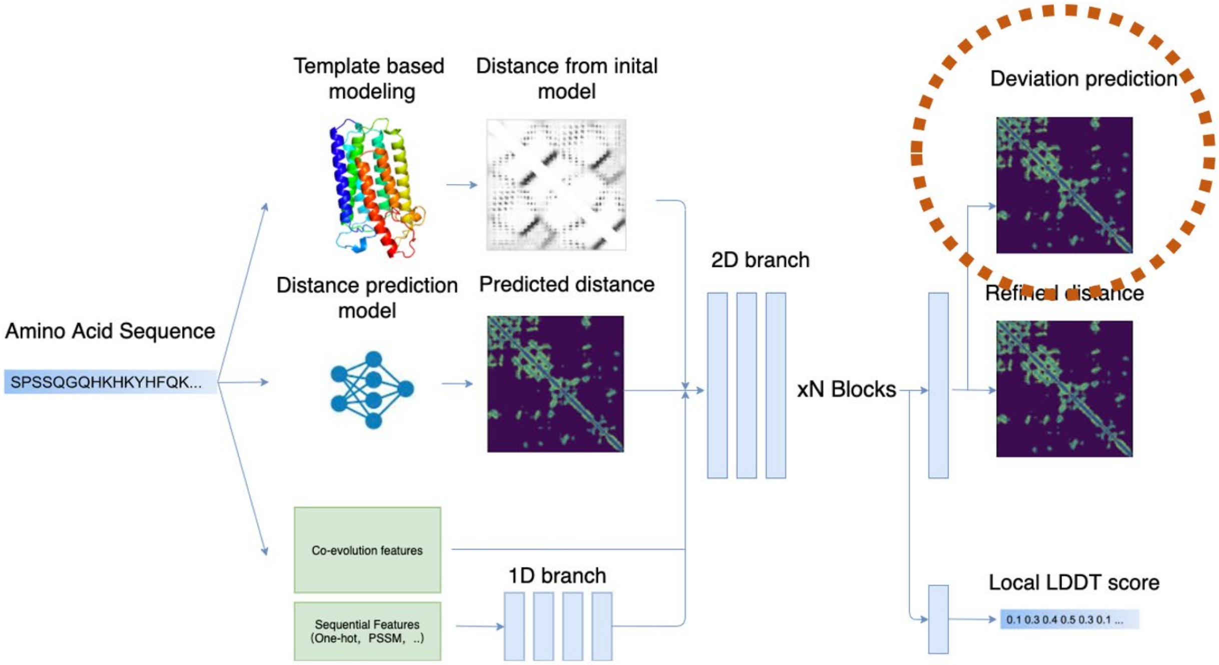


Figure 1. A scheme of our protocol to generate the predicted distance deviation.

1. Ma, J., Peng, J., Wang, S., & Xu, J. (2012). A conditional neural fields model for protein threading. Bioinformatics, 28(12), i59-i66.
2. Song, Y., DiMaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T. J., & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. Structure, 21(10), 1735-1742
3. Zhang, Y., & Skolnick, J. (2004). SPICKER: a clustering approach to identify near ‐ native protein folds. Journal of computational chemistry, 25(6), 865-871.
4. Yang, Y., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations, Proceedings of the National Academy of Sciences, 117: 1496-1503.
5. Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. (2017) doi:10.1021/acs.jctc.7b00125.
6. Shen, M. & Sali, A. Statistical potential for assessment and prediction of protein structures. Protein Sci. (2006) doi:10.1110/ps.062416606.
7. Zhou, H. & Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys. J. (2011) doi:10.1016/j.bpj.2011.09.012.
8. López-Blanco, J. R. & Chacón, P. KORP: Knowledge-based 6D potential for fast protein and loop modeling. Bioinformatics (2019) doi:10.1093/bioinformatics/btz026.

tFold @ QA

GQArank: Protein Model Quality Assessment via Neural Message Passing

J. Huang, T. Shen, N. Huang, H. Lan, W. Liu, S. Wang[#], J. Huang[#]
Tencent AI Lab
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

For quality assessment (QA) task in CASP14, we introduce GQArank by modifying a graph general architecture, Message Passing Neural Networks (MPNN)[1], to score the given decoys. GQArank combines information from the native and decoy and learns the interaction among residues to represent the similarity between the native and decoy. GQArank could not only predicts a global score for each decoy which represents how similar it is to native, it also could obtain the local score at each residue.

Methods

In the section, we show technical details of our GQArank method from two aspects: features and training strategy.

Features. In GQArank, features contain node features and edge features. (i) Node features are at residue-level. We first encode the residue identity into 22 dimensional one-hot code based on the corresponding amino acid. Next, we extract information of secondary structure by PDB_Tool[2] with the shape of L x 30. We also add a statistic about Multiple Sequence Alignment (MSA): PSSM features of shape L x 21. (ii) Edge features are the information between two residues. We use two MSA features: MRF (L x L x 441). Besides, we also use the in-house predicated distance map (see tFold@RR) of shape L x L x 37 (37 distance bins, defined according to trRosetta[3]).

Training strategy. During training, GQArank optimizes both global and local Mean Squared Error (MSE) losses. We choose global and local lDDT[4] scores computed from the native and decoy as global and local labels. The performance using both local and global loss is better than that only using global loss. Moreover, triplet loss is also optimized, which is used to make sure the large margin between scores of two decoys and improve ranking performance. Due to the absence or bad quality of features of some decoys, we design several data augmentations in the training process.

(a) Multi-MRF. We have 83 sources of MSA, thus get 83 MRF features (see tFold-CaT@RR). For each decoy in a training step, we randomly choose one of 83 MRF as input;

(b) Multi-distance. We have a variety of deep learning approaches to predict distance maps. Similarly, we randomly choose one of different distance maps as input in a training step;

(c) Feature-mask. The above-mentioned features are randomly masked with full-zero matrixes. Only one feature is masked in a training step;

(d) Distance/MRF-mask. We randomly mask part of distance and MRF features maps. Moreover, we observe that decoys exist the phenomenon of missing some residues, which would result in poor performance of GQArank. In order to alleviate this problem, we adopt node-drop strategy (i.e., randomly dropping several residues in a decoy) to process training samples.

1. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212.
2. https://github.com/realbigws/pdb_tool.
3. Yang, Y., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations, Proceedings of the National Academy of Sciences, 117: 1496-1503.
4. Mariani, V., Biasini, M., Barbato, A., & Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics, 29(21), 2722-2728.

tFold @ RR

Ultra-Deep Network for Distance Prediction with a Multi-input Multi-label Scheme under Criss-Cross Attention

T. Shen*, J. Wu*, H. Lan, L. Zheng, W. Liu, S. Wang#, J.Huang#
Tencent AI Lab
*equal contribution
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

To fully exploit the co-evolution information from a bunch of multiple sequence alignment (MSA) data during training and capture long-range interactions of residual pairs, here we present an ultra-deep convolutional network for distance prediction with a multi-input & multi-label scheme under criss-cross attention which achieves state of the art performance.

Methods

In this section, we describe the technical details of our distance prediction method from three aspects: 1) multi-input & multi-label scheme; 2) criss-cross attention module; 3) progressive training for the ultra-deep model.

***Multi-input & Multi-label Scheme*:** As shown in Figure 1, our network has multiple inputs which are features derived from the same protein sequence constructed under different sequence databases (UniClust30[1], UniRef90[2], NR90/NR70[3]). In details, given an amino acid sequence, our method runs HHblits[4] on UniClust30, JackHMMER[5] on UniRef90, PSI-BLAST[6] on NR90/NR70 to generate multiple MSAs. For each MSA, we extract Position Specific Scoring Matrix (PSSM) features of shape L x 21 and Markov Random Fields (MRF) features of shape L x L x 441, where L is the sequence length. The input features of each MSA is processed by stacked residual convolutional blocks, then the output features are concatenated to perform feature fusion and fed into the remaining blocks. Also, an additional aggregation branch is built with multiple aggregation blocks at each feature level. Each aggregation block takes features from the corresponding residual blocks of all branches and previous aggregation block, making the model capable of finding inter-database correlations from features at all levels. After the last residual block, the network branches out into 4 independent paths and simultaneously predict multi objectives: 1 distance histogram (*d* coordinate) and 3 angle histograms (ω, θ, and φ coordinates) as defined in trRosetta[7].

Criss-cross Attention Module: To capture long-range 2D contextual information more effectively, we adopted the criss-cross attention module[8] which considers a long but narrow kernel to expand the receptive field of CNN in fully attentional models. The key idea of this module is to separate 2D attention into two steps that apply 1D attention in the height and width axes sequentially. The efficiency of this approach enables attention over large regions, allowing our models to learn long-range, or even global, residue-residue interactions.

Progressive Training Schedule: We trained a network that is deeper than those in the published work and achieves better performance. The key idea is to grow the network progressively: 1) Start from a shallow model; 2) Concatenate new blocks to the model with identity mapping initialization when it converged. 3) Continue training and repeat 2)&3) if the performance improves. The network grows deeper as the progressive training. Specifically, we start from an 80 layers network and grow another 40 layers when the current model is converged. Following this progressive training schedule, we finally get a model which consists of over 600 layers.

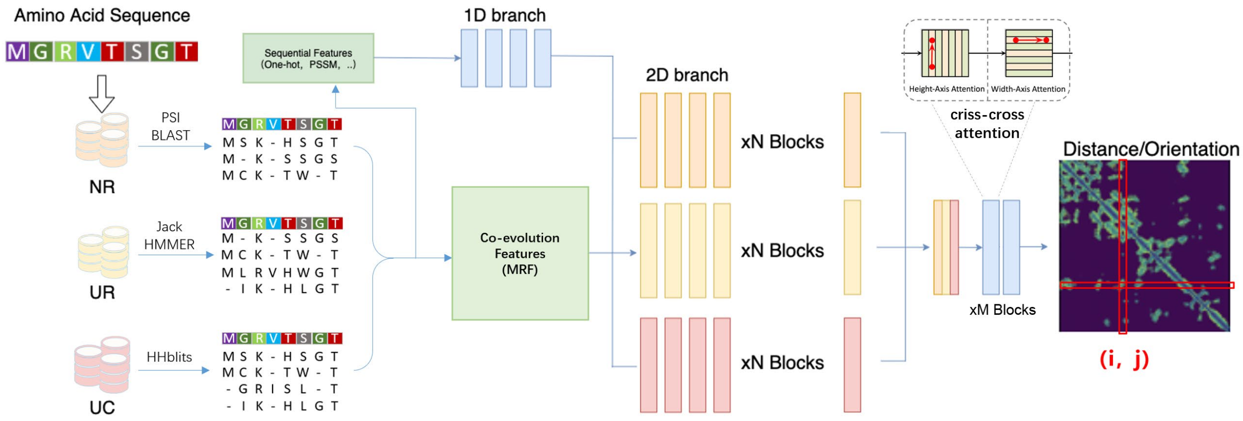


Figure 1. A scheme of our C-beta distance/orientation prediction protocol.

1. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., S?ding, J. and Steinegger, M., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research, 45(D1), pp.D170-D176.
2. UniProt Consortium, 2010. The universal protein resource (UniProt) in 2010. Nucleic acids research, 38(suppl_1), pp.D142-D148.
3. Coordinators, N.R., 2018. Database resources of the national center for biotechnology information. Nucleic acids research, 46(Database issue), p.D8.
4. Remmert, M., Biegert, A., Hauser, A. and Soeding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), p.173.
5. Finn, R.D., Clements, J. and Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39(suppl_2), pp.W29-W37.
6. Altschul, S.F., Madden, T.L., Sch?ffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.
7. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. and Baker, D., 2020. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3), pp.1496-1503.
8. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A. and Chen, L.C., 2020. Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. arXiv preprint arXiv:2003.07853.

tFold_human @ 3D

Protein Tertiary Structure Predictions by tFold Human Group in CASP14

T. Shen*, J. Wu*, J. Pei*, L. Zheng, H. Lan, W. Liu, S. Wang#, J. Huang#
Tencent AI Lab
*equal contribution
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP14, our tFold_human protein folding pipeline integrate multiple model quality assessment methods to rank and select CASP14 server models as the starting point. Along with our de novo approach, C-beta distance features are extracted from the top-ranking models to generate refined distance/orientation constraints using a deep learning model. Refined distance/orientations are generated with each selected sever model, and then fused with a three-stage pipeline.

Methods

The tFold_human pipeline is mainly composed by the following steps: 1) C-beta distance/orientation prediction from de novo approach; 2) CASP14 server models ranking; 3) C-beta distance/orientation refinement via features from server models; 4) The three-stage pipeline to automatically fuse such diverse distance refinements results.

Firstly, given a protein sequence, we extract MSA based features and perform C-beta distance/orientation prediction via our high precision deep neural network described in the tFold@RR abstract. Secondly, ranking methods from our groups (say, tFold@QA and tFold-CaT@QA) are applied to select top models from CASP14 server models. Thirdly, a deep learning model is developed to further optimize the C-beta distance/orientation prediction using both information from MSA based features and distance features from decoys (see tFold-IDT@RR). Refined distance/orientations are generated with each selected sever models. Finally, the three-stage pipeline described in the tFold-CaT@RR is applied, featuring hierarchical clustering and probability distribution based quality assessment, to generate the final distances/orientations.

After final distance/orientation is predicted, following trRosetta[1], we generated 3D structures from the refined distances/orientations using constrained minimization. Human interventions were made during the submission step.

1. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. and Baker, D., 2020. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3), pp.1496-1503.

tFold-CaT @ 3D

Protein Structure Prediction with Graph-based Decoy Ranking

J. Wu, T. Shen, L. Zheng, H. Lan, w. Liu, S. Wang#, J. Huang#
Tencent AI Lab
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

For the 3-D structure prediction task in CASP14, we adopt a mixture method with both template-based modeling and free modeling modules embedded. A graph-based decoy ranking method to proposed to identify the decoy that is the most likely to be the true 3-D structure of the given protein. This method captures both decoy-specific and inter-decoy features, and then combines them in a ranking model that is trained under a ranking-aware objective function.

Methods

In this section, we describe the two major components of protein structure prediction method with graph-based decoy ranking: decoy generation and decoy selection.

Decoy generation. We adopt two relatively independent pipelines for the structure prediction task in CASP14, one for template-based modeling targets and the other for free modeling targets. For each protein, we firstly run CNFpred[1] to check whether there exist sufficiently good templates to be used. If true, then it will be handled by the template-based modeling pipeline; otherwise, free modeling pipeline will be executed to generate candidate decoys.

a) Decoy generation with template-based modeling. If one target belongs to template-based modeling targets, then it indicates that CNFpred has already identified sufficiently good templates. With these templates, we run RosettaCM[2] to generate decoys and then apply Spicker[3] to cluster decoys and rank them according to their distance to the clustering centroid. For each template, we keep 300 decoys. For each target, we keep top 10 templates as selected by CNFpred.

b) Decoy generation with free modeling. If one target belongs to free modeling targets, then we must rely more on inter-residue distance and orientation predictions to provide extra guidance on the structure modeling process. An in-house ultra-deep network (see tFold@RR for more details) is trained to predict C-beta distance, and three inter-residue dihedral and plane angles, as defined in Yang et al.[4]. These predictions are converted into differentiable energy terms with spline interpolation, and then used to enhance Rosetta's statistics-based energy term in both centroid-mode and full-atom-mode optimization.

Decoy selection. Since candidate decoys may be generated with either template-based modeling or free modeling procedures, we need to develop a universal decoy ranking method that is applicable in both cases. We build our ranking strategy on a commonly used assumption by previous clustering-based quality assessment methods[5]: if one decoy is highly similar with most of candidate decoys, then this decoy is more likely to be a high-quality one. To fully exploit such information, we extract features to reflect the inter-decoy similarity, measured by various metrics (GDT-TS, TM-Score, and lDDT). Besides, decoy-specific features are also extracted, including statistics energies and consistency with predicted distance and orientation restraints.

Decoy-specific and inter-decoy features are combined within a graph-based model. Each decoy is treated as a node/vertex, and inter-decoy relationship corresponds to edges that connects two nodes. We set a similarity threshold to determine whether edge-features should be aggregated to the certain node, e.g. less similar decoys are omitted. With all the node features and aggregated edge features, we train a factorization machine model[6] to predict each decoy's ground-truth lDDT score. The loss function is a combination of mean-square-error loss and margin ranking loss, and we gradually increase the latter one's weighting coefficient to better preserve the ground-truth ranking order.

Results

We compare the graph-based decoy ranking method, namely "Rank-QA", with two baseline ranking methods, KORP and GraphQA. The test data consists of 74 targets selected from CAMEO's weekly release, ranging from 2019-12-07 to 2020-02-01. As depicted in Table 1, we discover that our proposed Rank-QA method outperform both baselines, regardless whether the pre-filtering strategy is enabled or not.

Table 1: Comparison on various decoy ranking methods, measured by averaged lDDT score of top-1 decoys on 74 CAMEO targets.

Method	w/o Pre-Filter	w/ Pre-Filter
KORP	0.6154	n/a
GraphQA	0.6217	0.6264
Rank-QA	**0.6256**	**0.6325**

1. Ma, J., Peng, J., Wang, S., & Xu, J. (2012). A conditional neural fields model for protein threading. Bioinformatics, 28(12), i59-i66.
2. Song, Y., DiMaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T. J., & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. Structure, 21(10), 1735-1742.
3. Zhang, Y., & Skolnick, J. (2004). SPICKER: a clustering approach to identify near-native protein folds. Journal of computational chemistry, 25(6), 865-871.
4. Yang, Y., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations, Proceedings of the National Academy of Sciences, 117: 1496-1503.
5. Han K.S., Choe M.H. (2018). Protein model quality estimation by clustering-based method SARTclust in CASP13. Abstract for CASP13:185-186.
6. Rendle, S. (2010). Factorization machines. IEEE International Conference on Data Mining: 995-1000.

Using Gradient Boosted Regression Trees for Assessing Local lDDT of Protein Model

N. Huang, J. Wu, L. Zheng, H. Lan, W. Liu, S. Wang[#], J. Huang[#]
Tencent AI Lab
shengwwang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

In CASP 14, we benchmarked our recently developed single model Quality Assessment (QA) method to predict local lDDT score of a protein model via Gradient Boosted Regression Trees (GBRT). Our prediction model could take input a variety of features, which originate from templates, predicted local properties, and some potential scores. In particular, template scores from Template-based Modeling (TBM) can provide us with reference information for the native structure of the protein model. Some accurately predicted local properties, such as secondary structure and solvent accessibility, could also bring us the knowledge about the quality assessment. Furthermore, we also selected some potential scores, such as Rosetta Ref2015[1] and QMEAN[2], as our input features for QA.

Methods

This section mainly introduces how to get the features of our model.

For the template scores, we use CNFpred[1] to search the template sequence, and the Multiple Sequence Alignment (MSA) used is searched in UniClust30[3] by HHblits[4]. The number of templates obtained from each CNFpred search may be different. Here we use the top 10 templates with the best quality (if there are less than 10, choose as many as possible), and use CD-HIT[5] to cluster the templates. After the clustering operation is completed, we will obtain several template clusters, and use the average similarity of the templates as the weight of each cluster. For each protein structure to be evaluated, after aligning with the template sequence, we calculate the Mean Absolute Error (MAE) of the C-alpha distance between all residues from the protein model and the corresponding template. The template of the same cluster should be multiplied by the weight of the cluster to obtain the template score, and finally the template score at that position is divided by the number of clusters.

For the QMEAN score, we use the following five potentials: (i) all-atom interaction potential, (ii) CB interaction potential, (iii) packing potential, (iv) reduced potential, and (v) torsion potential. We selected 8 scores from Rosetta Ref2015. For 1D prediction features, the secondary structure predicted by Porter5[6] and solvent accessibility predicted by ACCPro5[7]. Specially, we calculate the secondary structure score based on prediction confidence. The above features are extracted based on each residue, the average predicted score for each residue is treated as the global score prediction.

1. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., ... & Labonte, J. W. (2017). The Rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation, 13(6), 3031-3048.
2. Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics, 36(6), 1765-1771.

3. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., S?ding, J. and Steinegger, M., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research, 45(D1), pp.D170-D176.
4. Remmert, M., Biegert, A., Hauser, A. and Soeding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), p.173.
5. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28(23), 3150-3152.
6. Torrisi, M., Kaleel, M., & Pollastri, G. (2018). Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. bioRxiv, 289033.
7. Magnan, C. N., & Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics, 30(18), 2592-2597.

tFold-CaT @ RR

Multi-MSA Ensemble based Distance Prediction with Hierarchical Clustering and Quality Assessment upon Probability Distribution

J. Wu*, J. Pei*, H. Lan, T. Shen, w. Liu, S. Wang[#,] J. Huang[#]
Tencent AI Lab
*equal contribution
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

For the distance prediction task in CASP14, we adopt a multi-MSA ensemble based approach to fully exploit the co-evolution information from multiple sequence databases. An ultra-deep neural network is trained to predict inter-residue CB-CB distance, as well as several types of auxiliary inter-residue interactions. Distance predictions are generated with each MSA input independently, and then fused with a three-stage pipeline, featuring hierarchical clustering and quality assessment upon probability distribution, to generate the final ranked results.

Methods

In this section, we describe technical details of our multi-MSA ensemble based distance prediction method from three aspects: MSA search, network architecture, and multi-MSA ensemble.

MSA Search. We extract 83 groups of MSA data to fully exploit the co-evolution information, with 6 sequence databases and various MSA search hyper-parameters. Specifically, we run HHblits[1] on UniClust30[2] and UniRef30[2], JackHMMER[3] on UniRef90[4] (and its 2019 version), PSI-BLAST[5] on NR90/NR70[6] (and its 2019 version). Different combinations of E-value and number of iterations are used to introduce extra diversity in the multi-MSA data. For each MSA data, we extract PSSM features of shape L x 21 and MRF features of shape L x L x 441, where L is the sequence length. These are used as input features for the upstream distance prediction network.

Network architecture. We design an ultra-deep convolutional neural network (CNN) with 2D attention module to predict inter-residue CB-CB distance, which is similar as tFold@RR except for that the input data in this work only come from one MSA instead of multiple MSAs.

Multi-MSA ensemble. By feeding 83 groups of MSA data into the distance prediction network, we obtain 83 groups of distance predictions. Often, multiple unique patterns can be discovered from such prediction results. Simple averaging of all the distance predictions may hinder these unique patterns, and introduce unnecessary noise due to low-quality ones. On the other hand, if we feed all these 83 MSAs into the multi-MSA input network as described in tFold@RR, the calculation time will be extremely slow. Therefore, we develop a three-stage pipeline to automatically fuse such diverse distance predictions as shown in Figure 1.

Firstly, we perform hierarchical clustering to group distance predictions with similar patterns into the same cluster. To define the similarity between distance predictions, another CNN model is trained to predict the protein fold category from distance predictions. Afterwards, we extract the last layer's activations as the embedding vector for distance predictions, and compute the cosine similarity between embedding vectors for hierarchical clustering. After clustering, we filter out those clusters with only one member, and use the averaged distance predictions to represent each remaining cluster.

Secondly, we propose a probability distribution based quality assessment metric to identify high-quality distance predictions. Since the same network is used to generate all the distance predictions, if the distance distribution for some inter-residue pair shows a single high peak, then it implies that the model is highly confident in its predictions. Hence, we design a metric to measure the concentration degree of predicted distance distributions. This is defined as the AUC score of distance distribution's standard deviation vs. coverage ratio of inter-residue pairs. We filter out distance predictions with low assessment scores, which is more likely to be of low-quality.

Finally, we execute another round of hierarchical clustering on the remaining distance predictions, if necessary, and then calculate the averaged distance predictions within each cluster. These predictions are further ranked with the above assessment metric to identify the best one for the final submission.

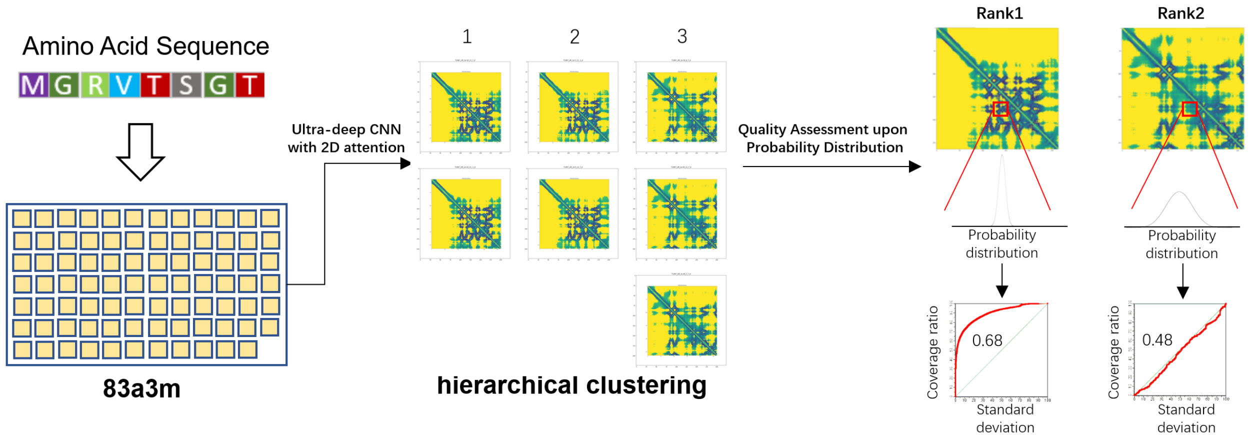


Figure 1. A scheme of our protocol to generate and rank C-beta distance prediction.

1. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., S?ding, J. and Steinegger, M., 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research, 45(D1), pp.D170-D176.
2. UniProt Consortium, 2010. The universal protein resource (UniProt) in 2010. Nucleic acids research, 38(suppl_1), pp.D142-D148.
3. Coordinators, N.R., 2018. Database resources of the national center for biotechnology information. Nucleic acids research, 46(Database issue), p.D8.
4. Remmert, M., Biegert, A., Hauser, A. and Soeding, J., 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 9(2), p.173.
5. Finn, R.D., Clements, J. and Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39(suppl_2), pp.W29-W37.
6. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.

tFold-IDT: Applying Statistical Potentials to Select Protein Decoys from Distance/Orientation-assisted De Novo Folding

J. Pei*, L. Zheng*, T. Shen, J. Wu, W. Liu, S. Wang#, J. Huang#
Tencent AI Lab
*equal contribution
shengwwang@tencent.com, joehhuang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

The 3D models submitted by tFold-IDT server were generated from the decoy ensembles by an automatic distance/orientation-assisted de novo folding pipeline, and followed by a statistical potential to rank and select. The distance/orientations used to generate the decoys came from two different sources: de novo prediction from the Markov Random Fields (MRF) features embedded in the Multiple Sequence Alignment (MSA), or integrating the structural features embedded in the initial pose from Template-Based Modeling (TBM).

Methods

Decoys generation: The predicted distance/orientations used in this server were generated by tFold@RR, tFold-IDT@RR, and tFold-CaT@RR. Once a predicted distance/orientation is selected, we follow the trRosetta[1] approach to generate the 3D decoys with several rounds of energy minimization in centroid mode and full-atom FastRelax, followed by minimization in terms of the Rosetta Ref2015[2] score. Afterwards, top 20 decoys generated by the previous two steps were selected with their corresponding Rosetta energy in consideration of the predicted distance/orientation as the constraints.

Decoys selection: We developed an integrated statistical potential for assessing the quality of the 3D decoys generated in the previous step. In detail, we collected a group of single-chain X-ray protein structures with resolution less than 2.0 angstrom and sequence identity less than 40%, the statistical energies (DOPE[3], GOAP[4] and KORP[5]) were calculated. Given a decoy with length L, the potential energies were also evaluated, and then the z-scores were calculated using the mean and stand deviation values from X-ray structures of similar sequence length (0.9*L to 1.1*L). We collected the decoys in CASP6-CAS12 as well as in CAMEO (before 2018) for training, and decoys in CASP13 and CAMEO (after 2018) for test. The z-scores of the decoys were computed and were fed to a linear regression model to approximate the GDT-TS (against the ground truth) score of the decoy. In CASP14, the decoy from the 5 highest scores were selected for final submission.

Results

In Figure 1, we present the GDT-TS score of the best submitted model for "all groups" and "server only" predictors participating the CASP14 tertiary structure prediction experiment for 21 currently releaed targets that could be identified in PDB. It shows that tFold-IDT server excellences at several targets, such as T1030, T1043, T1046s2, T1049, T1056, and achieves an average GDT-TS score of 0.52, which is ranked at 5th position among all participant servers.

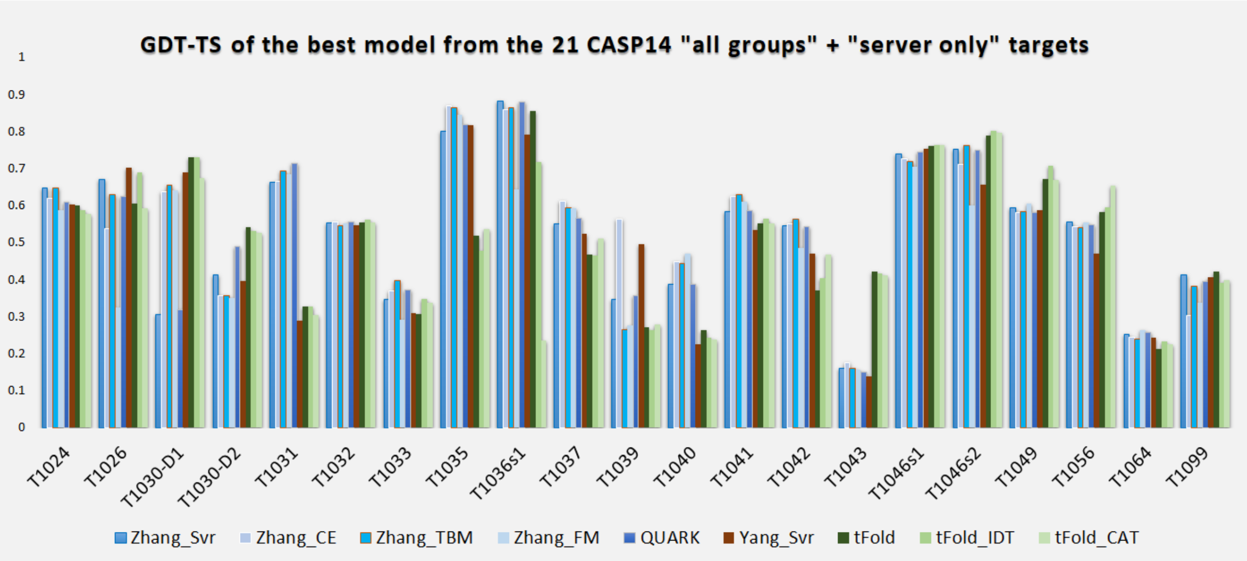


Figure 1. Performance of tFold-IDT in the tertiary structure prediction category for 21 CASP14 "all groups" + "server only" targets. Here we use GDT-TS score to measure the best submitted model with the native structure.

1. Yang, J. et al. trRosetta. Proc. Natl. Acad. Sci. U. S. A. (2020) doi:10.1073/pnas.1914677117.
2. Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. (2017) doi:10.1021/acs.jctc.7b00125.
3. Shen, M. & Sali, A. Statistical potential for assessment and prediction of protein structures. Protein Sci. (2006) doi:10.1110/ps.062416606.
4. Zhou, H. & Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys. J. (2011) doi:10.1016/j.bpj.2011.09.012.
5. López-Blanco, J. R. & Chacón, P. KORP: Knowledge-based 6D potential for fast protein and loop modeling. Bioinformatics (2019) doi:10.1093/bioinformatics/btz026.

tFold-IDT @ RR

Integrated approach for De novo folding with Template-based modeling

T. Shen, L. Zheng, H. Lan, J. Wu, W. Liu, S. Wang[#], J. Huang[#]
Tencent AI Lab
shengwwang@tencent.com

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

In CASP14, we proposed a new Integrated approach for De novo folding with Template-based modeling (IDT). Combining the distance prediction from de novo approach and the initial pose from template-based modeling (TBM) methods as the input, the predicted distance is further optimized using a deep neural network.

Methods

The IDT pipeline is mainly composed by the following steps: 1) C-beta distance/orientation prediction from de novo approach; 2) template searching and 3D structure modeling; 3) C-beta distance/orientation refinement via features from de novo approach and TBM methods;

Firstly, given a protein sequence, we extract multiple sequence alignment (MSA) based features and perform C-beta distance/orientation prediction via our high precision deep neural network described in the tFold@RR abstract.

Secondly, for the protein sequence, the most probable templates are searched using template-based modeling (TBM) tool CNFpred[1]. The best single template, as well as the query-template sequence alignment, is then used to generate 300 decoys following the standard RosettaCM[2] protocol. The best TBM decoy is then selected using structure clustering tool Spicker[3].

Finally, a deep learning model is developed to further improve the C-beta distance/orientation prediction using both information from MSA based features and templated based features. Specifically, the best TBM decoy is used to extract the templated alignment as well as C-beta distance as the input features, along with the MSA based features and the predicted distance in step 1), to be fed into the deep learning model.

This deep learning model is designed in a multi-task scheme and the output consists of three parts:(a) C-beta distance/orientation; (b) distance deviation; (c) local lDDT score of TBM decoy. The distance deviation of a decoy is defined as the absolute difference of C-beta distances (extracted from the decoy) and the ground truth. See Figure 1 in the abstract of tFold@3D.

Results

Our method can effectively integrate TBM features into distance/orientation prediction. As shown in Figure 1, experimental results on the CAMEO targets indicated that this method does not depend on the quality of the TBM decoy: high-quality templates will greatly improve the distance prediction while low-quality templates have almost no loss in performance. When the lDDT score of the TBM decoy is above 0.6, there is a high chance that the C-beta prediction could be greatly improved by our approach.

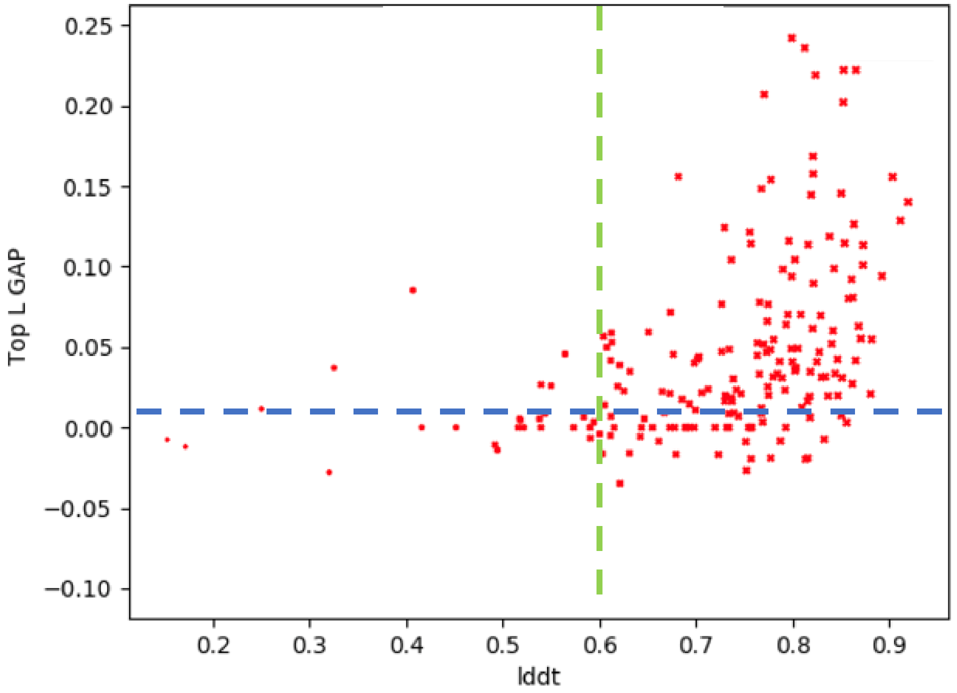


Figure 1. **The relationship between the model quality of the TBM decoy and the prediction enhancement of the C-beta distance on the CAMEO datasets**. The x-axis lDDT is the score for template modeling. The y-axis represents the enhancement of the C-beta distance prediction compared to the original input. It is indicated that the higher the lDDT, the greater improvement of the C-beta distance. However, the TBM decoy with low lDDT won't influence the prediction quality much.

1. Zhu, J., Wang, S., Bu, D. & Xu, J. Protein threading using residue co-variation and deep learning. in Bioinformatics (2018). doi:10.1093/bioinformatics/bty278.
2. Song, Y. et al. High-resolution comparative modeling with RosettaCM. Structure (2013) doi:10.1016/j.str.2013.08.005.
3. Zhang, Y. & Skolnick, J. SPICKER: A clustering approach to identify near-native protein folds. J. Comput. Chem. (2004) doi:10.1002/jcc.20011.

TOWER and FoldX

Protein Structure Prediction by Bridging TBM and FM Approaches

Jianwei Zhu, Tong Wang, Bin Shao and Tie-Yan Liu
Microsoft Research Asia, Beijing 100190, China
jianwzhu@microsoft.com

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:N*

The TOWER CASP14 submissions were automatically produced by a hybrid protein structure prediction system. TOWER combines the strengths of template-based modeling (TBM) and free modeling (FM). For proteins with template structures, we used template-based modeling method to select good templates and construct reliable structures based on the templates; For proteins without good templates, we used a distance-based free modeling method to construct their structures. The same distance-based free modeling approach also underpins the server FoldX.

Methods

Given a target protein sequence, we first construct MSA by running DeepMSA[1] against sequence databases and predict protein features and inter-residue distances from the generated MSA. Then we construct 3D structures using TBM approaches and FM approaches from the predicted features and distances. Finally, we select five models from the predicted decoys for submissions. The most important parts of our pipeline are inter-residue distance prediction, template-based modelling, and distance-based free modeling.

Inter-residue Distance Prediction. We trained a deep dilated residual neural network to predict the inter-residue distance from the raw MSA rather than handcrafted features such as profile, secondary structure, and solvent accessibility. Without using pre-calculated coevolutionary couplings, we can learn coevolution information from MSA directly and yield accurate distance distribution.

Template-based Modeling. We ran HHpred[2], SPARKS-X[3], DeepFR[4], and DeepThreader[5] independently, then select top 10 templates as the candidate templates for each software package. After template selection, we ran DeepThreader to build query-template alignments for the selected 40 templates and used MODELLLER[6] to build 3D structures for the 40 templates (except duplicates). Note that we used our own predicted inter-residue distances as described above instead of the distances predicted by RaptorX-Contact[7] for running DeepThreader. At last, we got about 40 candidate submissions for each target for the TBM targets.

Distance-based Free Modeling. We independently ran two distance-based FM approaches: our in-house distance-based method and trRosetta[8]. For each method, we generated 150 decoys for each target using different starting structures with random backbone torsion angles. Top five decoys for each method are ranked by the energy potential. At last, we got 10 candidate submissions for each target for the FM targets.

After model construction, five ranked models were submitted for each target by ProQ3D[9]. We chose the top two candidates from TBM models and three candidates from FM models by ProQ3D score. The model with the highest ProQ3D score was selected as the top one model

Results

We evaluated the performance of TOWER on 104 CASP13 human domains. For a fair comparison, the submissions for three top CASP13 human groups and three top serve groups were downloaded from the CASP13 data archive and were evaluated in the same way as TOWER. For all 104 domains of Human Group, TOWER achieved 0.727 average top-1 TM-score, which is higher than that of A7D (0.699). For the 31 FM targets of Human Group, which are usually considered more difficult than TBM targets, the average TM-score is 0.631, which outperforms A7D (0.580).

Methods	**All (104)**	**TBM (61)**	**FM/TBM (12)**	**FM (31)**
A7D	0.699/0.733	0.761/0.786	0.691/0.739	0.580/0.626
Zhang	0.692/0.719	**0.801**/0.816	0.605/0.665	0.509/0.549
MULTICOM	0.688/0.722	0.794/**0.817**	0.645/0.675	0.495/0.551
QUARK	0.672/0.699	0.786/0.808	0.589/0.648	0.479/0.503
Zhang-Server	0.671/0.699	0.787/0.807	0.593/0.627	0.475/0.514
RaptorX-DeepModeller	0.653/0.674	0.774/0.786	0.561/0.592	0.451/0.486
TOWER	**0.727/0.755**	0.780/0.807	**0.706/0.727**	**0.631/0.663**

1. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y., & Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*, 36(7), 2105-2112.
2. Söding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, 21(7), 951-960.
3. Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, 27(15), 2076-2082.
4. Zhu, J., Zhang, H., Li, S. C., Wang, C., Kong, L., Sun, S., ... & Bu, D. (2017). Improving protein fold recognition by extracting fold-specific features from predicted residue–residue contacts. *Bioinformatics*, 33(23), 3749-3757.
5. Zhu, J., Wang, S., Bu, D., & Xu, J. (2018). Protein threading using residue co-variation and deep learning. *Bioinformatics*, 34(13), 263-273.
6. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., ... & Sali, A. (2006). Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*, 15(1), 5-6.
7. Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34), 16856-16865.

8. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., & Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3), 1496-1503.
9. Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B., & Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, 33(10), 1578-1580.

trfold

trfold: deep neural network based de novo protein structure prediction

H.Miao[1], J.Wang[1], H.Sun[1], J.Li[1], Y.Tang[1], Q.Shen[1], and G.Xue[1]
[1] - *Tianrang Intelligence Inc Ltd,*
hj-miao@tianrang-inc.com

Key: *Auto:N; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

trfold CASP14 pipeline is a meta predictor that consists of three variants of independent deep neural network based de novo protein tertiary structure modelling approaches. The pipeline is designed to be fully automatic but was entered in CASP14 as a human group due to limited computing resources to meet CASP deadline for large targets. Full-chain models were constructed without domain segmentation and refinement was carried out with customised Rosetta relax protocols to optimise backbone conformation and to add in side-chains.

Methods

MSA construction: All three variants take the same set of multiple sequence alignment (MSA), which was constructed by progressively searching the latest UniRef and metagenome NR databases in a similar fashion as described in [1], but with slightly different parameters and filtering criteria. Although we do not carry out domain segmentation, each target sequence was sliced into fixed-length subsequences. MSAs for subsequences were constructed to overcome the problem of imbalance in MSA coverage, especially for multi-domain and long CASP targets.

DNN

All three variants are based on deep residue convolutional neural networks that predict the inter-residue distance and torsion angle distributions of target proteins, but they differ from several aspects: for the first, trRosetta[2] was used to predict distance and orientation angles together; for the second, two separate networks were trained on CATH domain database to predict distance and angle distributions and input features include sequence 1-hot encoding, HMM and Potts model; for the last, networks of similar architecture to the second were trained on full-chain PDB30, and input features that can be computed on the fly were included, i.e., sequence 1-hot encoding, PSSM and positional entropy, to allow MSA subsampling during training. Distance prediction from the third variant was submitted in the CASP contact/distance category.

If Neff values for subsequence MSAs differed significantly, distance and angles would be predicted for all subsequences as well as the full-length sequence. These predictions were then pasted together for model construction in the next stage.

A multi-head DNN with modified network architecture was trained on PDB100 at the end of CASP. This variant will be evaluated against the first three once CASP target structural information is released.

Model construction. Following the footsteps of [2,3], distance and torsion angle distributions were converted to potential scores and further smoothed to allow gradient decent (GD). van der Waals force, hydrogen bonding and several centroid scores were also used during GD. Side-chains were fitted using Rosetta FastRelax[4] with distance and angle score weights reflecting respective prediction confidence. To avoid being trapped in local minima, folding and refinement steps were carried out together. Optimisation of three variants was independently done for each.

Model selection. Five decoys of lowest potential from each variant were selected and scored by ProQ3D[5]. Top five scoring models formed our CASP14 submission and their per-residue error estimates (B-factor column) were converted from ProQ3D results. One model for T1061 was manually corrected due to steric clashes in this long target.

Results

All three variants in the trfold pipeline were benchmarked on CAMEO hard targets. Although average contact precision from three networks was similar, differences in distance distribution and subsequent constructed models were observed, especially for longer targets and targets with shallow MSAs. Thus, we pooled models from these variants to form our trfold CASP submission.

Availability

trfold pipeline will be available as a web server after formally benchmarking our CASP models against updated networks mentioned earlier.

1. Zhang, C., Zheng, W., Mortuza, S.M., Li, Y. and Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics, 36, 2105-2112.
2. Yang,J., Anishchenko,I., Park,H., Peng,Z., Ovchinnikov,S. and Baker,D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117, 1496-1503.
3. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Žídek,A., Nelson,A.W., Bridgland,A. and Penedones,H. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577, 706-710.
4. Tyka,M.D., Keedy,D.A., André,I., DiMaio,F., Song,Y., Richardson,D.C., Richardson,J.S. and Baker,D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. Journal of molecular biology, 405, 607-618.
5. Uziela,K., Menendez Hurtado,D., Shu,N., Wallner,B. and Elofsson,A. (2017). ProQ3D: improved model quality assessments using deep learning. Bioinformatics, 33, 1578-1580.

Protein structure prediction based on multi-objective optimization

Yu-Lin Wang[1,2], Hong-Bin Shen[1,2]

[1] - Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, [2] - Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

hbshen@sjtu.edu.cn

Protein structure prediction[1-6] is an important area for rapidly generating the structure models from amino acid sequences and many software has been developed, i.e., Rosetta[7], I-TASSER[8], and etc. Energy-based optimization is an important step in protein structure prediction and we try to test and dig out how the multi-objective optimization will affect the protein structure prediction, which will include more than 1 energy function in the protocol and optimized parallelly. A potential merit for parallelly optimizing more energy functions is it may jump out from the local optimal solution from only a single energy function. We have tested three energy functions in the optimization and model selection process, i.e., Rosetta energy function[6], CHARMM36[9] and RWplus[10]. We use three energy functions during the optimization iterations to evaluate solutions for finding the non-dominated solution particles. Then we perform a non-dominant sort[11] of set with crowding-distance calculation. We introduce elitism preserving strategy and randomly choose some of the first level structures as new swarm population and start the next iteration. Finally, we get many non-dominated solution models in the Pareto sets and the Knee[12] algorithm will be applied to rank the solutions and output the top 5 models. As for some very long sequences, we have applied the ThreaDom[13] to predict domain boundary and cut sequences into several parts for independent modeling. Our local tests on CASP13 targets have suggested that the three energy function-based non-dominated solution generation pipeline could be promising, which can be complementary to existing studies.

1. Klepeis JL, Wei Y, Hecht MH et al (2005) Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 58(3):560−570
2. Oldziej S, Czaplewski C, Liwo A et al (2005) Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proceedings of the National Academy of Sciences of the United States of America 102(21): 7547−7552
3. Jauch R, Yeo HC, Kolatkar PR et al (2007) Assessment of CASP7 structure predictions for template free targets. Proteins 69(Suppl 8):57−67
4. Kinch LN, Li W, Monastyrskyy B, et al. (2015). Evaluation of free modeling targets in CASP11 and ROLL. Proteins 84: 51−66
5. Taylor WR, Bartlett GJ, Chelliah V et al (2008) Prediction of protein structure from ideal forms. Proteins 70(4):1610−1619
6. Kim DE, Chivian D, Baker D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32, W526–31.

7. Bradley P, Malmstrom L, Qian B et al (2005a) Free modeling with Rosetta in CASP6. Proteins 61(Suppl 7):128−134

8. A Roy, A Kucukural, Y Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols, 5: 725-738 (2010)

9. Huang,J. et al. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat. Methods, 14, 71−73.

10. Zhang,J. and Zhang,Y. (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One, 5, e15386.

11. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, April 2002, doi: 10.1109/4235.996017.

12. Branke,J. et al. (2004) Finding knees in multi-objective optimization. In: International Conference on Parallel Problem Solving from Nature.

13. Z Xue, D Xu, Y Wang, Y Zhang. ThreaDom: Extracting Protein Domain Boundary Information from Multiple Threading Alignments. Bioinformatics, 29: i247-i256 (2013).

TripletRes

Inter-residue distance prediction by raw co-evolutionary features and joint training of multiple auxiliary tasks

Yang Li[1, 2], Chengxin Zhang[1], Wei Zheng[1], Xiaogen Zhou[1], Eric W. Bell[1], Dong-Jun Yu[2], and Yang Zhang[1]

[1] - *Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109*

[2] - *School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094;*

yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:N; MD:N*

TripletRes in CASP14 extends our previous version in CASP13 by contact confident score based Multiple Sequence Alignment (MSA) selection, raw co-evolutionary feature extraction and joint training of multiple inter-residue interaction tasks.

Methods

For a query sequence, 6 candidates MSAs are built by searching the query through different whole-genome and metagenome sequence databases. 3 MSAs can be generated by DeepMSA[1] which utilizes HHblits, Jackhmmer and HMMsearch to search against Uniclust30, UniRef90 and Metaclust. Another 2 MSAs are further generated by using HHblits3 and HMMsearch to search against BFD and Mgnify database respectively. IMG/M database is used for constructing the last MSA by HMMsearch. Note that searching results by Jackhmmer and HMMsearch are all built to custom HHblits format databases, which will be re-searched by HHblits.

A set of coevolutionary features will be extracted from each of the obtained MSAs. The raw coupling parameters of pseudolikelihood maximized (PLM) 22-state Potts model[2] and the raw mutual information (MI) matrix are the two major two-dimensional features in TripletRes. Here, the 22 states represent the 20 standard amino acids, the non-standard amino acid type and the gap state. The corresponding parameters for each residue pair in PLM and MI matrix are also extracted as additional features that measure query-specific coevolutionary information in an MSA. The field parameters and the self-mutual information are considered as the one-dimensional features, incorporated with HMM features. One-hot representation and the descriptors of MSAs, e.g., number of sequences in the MSA, are also considered.

The one-dimensional features and two-dimensional features are fed into deep convolutional neural networks separately. Each of them will go through a set of one-dimensional and two-dimensional residual blocks[3], respectively, and then tiled together. The feature representations are considered as the inputs of another fully residual neural networks which will output several inter-residue interaction terms. The Cα-Cα distances, the Cβ-Cβ distances, torsional angle terms, and H-bond geometry descriptors between residues are considered as prediction terms. The distance, torsion angles, and H-bond geometry values are discretized into binary descriptions and the neural networks are trained with cross-entropy loss.

TripletRes selects MSAs by the summation of the cumulative probability under 12 Å of top 10*L and predicts Cβ-Cβ distance distributions for all residue pairs. The Cβ-Cβ distance distribution based on the selected MSA with the highest probability is considered as the final prediction.

1. Zhang, Chengxin, Wei Zheng, S. M. Mortuza, Yang Li, and Yang Zhang. "DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins." Bioinformatics 36, no. 7 (2020): 2105-2112.
2. Ekeberg, Magnus, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models." Physical Review E 87, no. 1 (2013): 012707.
3. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

UNRES

Template-free prediction of protein structures with the coarse-grained UNRES force field and replica-exchange molecular dynamics

A. Antoniak[1], K.K. Bojarski[1], C. Czaplewski[1], P.A. Wesołowski[1,2], M. Maszota-Zieleniak[1], M. Kogut[1], A.G. Lipska[1], E.A. Lubecka[3], K. Zięba[1], A.K. Sieradzan[1], A. Giełdoń[1], M.M. Kogut[1], P. Krupa[4], M. Marcisz[1], S.A. Samsonov[1], R. Ślusarz[1], M.J. Ślusarz[1], and A. Liwo[1]

[1] – Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland,

[2] – Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, ul. Abrahama 58, 80-307 Gdańsk, Poland,

[3] – Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland,

[4] – Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46 Warsaw, PL-02668, Poland

adam.liwo@ug.edu.pl

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:Y; DeepL:N; EMA:Y; MD:Y*

We tested, with the CASP14 targets, our physics-based approach for protein-structure prediction, whose key component is the coarse-grained UNRES model of polypeptide chains[1] with multiplexed replica exchange molecular dynamics (MREMD)[2] as the main conformational-search engine. As opposed to most of the other approaches, our method does not use heavy knowledge-based input and the prediction candidates are selected according to the probabilities of the conformational ensembles they belong to. Thus, our approach is effectively database-independent. With respect to the last CASP, our methodology was extended to treat large proteins and protein complexes and to handle the data-assisted targets better.

Methods

UNRES is a heavily reduced model of polypeptide chains, in which a chain is represented by a sequence of a-carbon (C^a) atoms, connected by virtual bonds, with attached united side chains. Two interaction sites are assigned to each amino-acid residue: the united peptide group (p) located in the middle of two consecutive C^a atoms and the united side chain (SC). The C^a atoms are not interaction sites but serve to define chain geometry. The interactions are described by the UNRES potential-energy function, derived from the generalized cluster-cumulant expansion of the potential of mean force (PMF) of polypeptide chains in water, in which all degrees of freedom not present in the model are integrated out. Owing to the implementation of the scale-consistent approach developed in our earlier work[3], the averaged out degrees of freedom are embedded in the coarse-grained potentials, resulting in correct dependence of the effective energy terms on site orientation and enabling us to derive the analytical expressions for the multibody terms, which are necessary to reproduce regular protein structures at the coarse-grained level. We used the same version of the force field as in CASP13[4].

The structures of the target proteins were predicted by the following four-stage procedure[5]:

1. MREMD simulations or, for very large targets, canonical MD simulations (see Results), were run, which consisted of 48 replicas at 12 temperatures from 260 K to 370 K, for 20,000,000 4.89 fs steps/trajectory. To speed up the search for larger proteins, weak restraints were imposed on secondary structure based on secondary structure prediction by PSIPRED[6] and the conformational search was started from the top-quality server models, as determined by using the DeepQA method[7]. The initial structures of the assembly targets were generated, if possible, based on the HHPred[8] hints from homology-related assemblies. The raw PSIPRED output was used to generate bimodal restraints on backbone virtual-bond-dihedral angles, with one minimum in the a-helical and another one in the extended region, the well-depths depending on PSIPRED-determined probabilities[9]. The replicas were exchanged and the snapshots were collected every 10,000 MD steps.

2. The weighted-histogram analysis method (WHAM)[10] was used to calculate the relative free energy of each structure of the last section of MREMD simulation. The last 1,000 snapshots per trajectory were processed.

3. The conformational ensembles obtained in steps 2. and 3. were subjected to cluster analysis to generate 5 (for CASP) or 10 (for CAPRI) clusters (see refs 5, and 9 for details).

4. The conformations closest to the respective average structures corresponding to the found clusters, and ranked according to the free energies of the clusters, were converted to all-atom structures using the PULCHRA[11] and SCWRL[12] algorithms. These structures were subsequently refined by using the AMBER14 package[13] with the ff14SB force field and GBSA implicit-solvent model. 500 minimization steps followed by 3,000 steps of molecular dynamics with 0.1 fs step length (0.3 ps total), with restraints on the secondary structure and positional restraints from the parent UNRES structure, and 500 final minimization steps with imposing only restraints on secondary structure were run for each structure. The refined all-atom structures were submitted to CASP or CAPRI.

Anticipating large targets to appear in CASP14, we introduced the necessary modification into the UNRES conformational-search engine. These included (i) distance cut-off on long-range interactions with the interaction list distributed to slave tasks, (ii) transforming the inertia matrix in MD to a five-diagonal form and (iii) implementation of the LBFGS energy-minimization algorithm[14]. These modifications resulted in linear scaling of the CPU and memory requirements with system size, as opposed to scaling with the square of the system size in earlier versions of UNRES.

We also improved the penalty functions to handle the SAXS and ambiguous NMR data developed in our earlier work[9], by introducing the dependence of the site radii in SAXS on the number of neighbors to account for the hydration shell and by developing a procedure to compute approximate positions of the protons given the coarse-grained geometry. To handle the refinement targets, we used normal model analysis to determine the flexibility of the chain and impose restraints accordingly.

Results

We treated all regular, assembly, refinement, and data-assisted targets, following the procedure described in Methods. For the two large assembly targets: H1081 (a 20-mer, over 12,000 residues total) and T1099 (a 240-mer virus capsid, over 60,000 residues total), due to large resource requirements, we ran only canonical MD simulations. For T1099, a difficult issue was to construct the initial model without major overlaps. Assembling the server models of the monomer did not work and, therefore, we regenerated the C-terminal part of each monomer, subject to symmetry conditions and so as to avoid overlaps. An energy minimization followed by a canonical MD was run for each of the models generated. Because the minimal unit of the capsid containing the unique interface had to be submitted, we modified the cluster analysis to cut the final capsid structures into such units first and treat each unit as an independent structure in clustering.

We postpone the assessment of the approach until the official release of CASP14 results.

Availability

The standalone version of UNRES is available from www.unres.pl and the web server version is available at http://unres-server.chem.ug.edu.pl.

1. Liwo,A., Czaplewski,C., Ołdziej,S., Rojas,A.V., Kaźmierkiewicz,R., Makowski,M., Murarka, R.K. & Scheraga,H.A. (2008) Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In: *Coarse-Graining of Condensed Phase and Biomolecular Systems.*, ed. G. Voth, Taylor & Francis, Chapter 8, pp. 107-122.
2. Czaplewski,C., Kalinowski,S., Liwo,A. & Scheraga,H.A. (2009) Application of multiplexed replica exchange molecular dynamics to the UNRES force field: Tests with α and α+β proteins. *J Chem. Theory Comput.* **5**, 627-640.
3. Sieradzan,A.K., Makowski,M., Augustynowicz,A. & Liwo, A. (2017) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. *J. Chem. Phys.*, **146**, 124106.
4. Liwo,A., Sieradzan,A.K., Lipska,A.G., Czaplewski,C., Joung,I., Żmudzińska,W., Hałabis,A. & Ołdziej.S. (2019) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. *J. Chem. Phys.*, **150**, 155104.
5. Krupa,P., Mozolewska,M.A., Wiśniewska,M., Yin,Y., He,Y., Sieradzan,A.K., Ganzynkowicz,R., Lipska,A.G., Karczyńska,A. Ślusarz,M., Ślusarz,R., Giełdoń,A., Czaplewski,C., Jagieła,D., Zaborowski,B., Scheraga,H.A. & Liwo,A. (2016) Performance of protein-structure predictions with the physics-based UNRES force field in CASP11. *Bioinformatics*, **32**, 3270-3278.
6. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
7. Cao,R., Bhattacharya,D., Hou, J. & Cheng,J. (2016) DeepQA improving the estimation of single protein model quality with deep belief networks. *BMC Bioinf.* **17**, 495.
8. Söding,J., Biegert,A. & Lupas,A.N. (2005) The HHPred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244−W248.
9. Lubecka,E.A., Karczyńska,A.S., Lipska,A.G., Sieradzan,A.K., Zięba,K., Sikorska,C., Uciechowska,U., Samsonov,S.A., Krupa,P., Mozolewska,M.A., Golon,Ł., Giełdoń,A., Czaplewski,C., Ślusarz,R., Ślusarz,M., Crivelli,S.N. & Liwo,A. (2019) Evaluation of the scale-consistent UNRES force field in template-free prediction of protein structures in the CASP13 experiment. *J. Mol. Graph. Model.*, **92**, 154-166.

10. Kumar,S., Bouzida,D., Swendsen,R.H., Kollman,P.A. & Rosenberg,J.M. (1992) The weighted histogram analysis method for free energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011-1021.
11. Rotkiewicz,P. & Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.*, **29**, 1460-1465.
12. Wang,Q., Canutescu,A.A. & Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. *Nat. Protoc.* **3**,1832-1847.
13. Case,D.A. et al. (2014), AMBER 14, University of California, San Francisco.
14. Lui,D.C. & Nocedal,J. (1989) On the limited memory BFGS method for large scale optimization, *Math. Program.*, **45**, 503-528.

UNRES-contact

Contact-assisted structure prediction with the multiplexed replica exchange molecular dynamics in the coarse-grained UNRES force field

E.A. Lubecka[1], M. Kogut[2], M. Maszota-Zieleniak[2], A.G. Lipska[2], K.K. Bojarski[2], A. Giełdoń[2], S.A. Samsonov[2], P. Krupa[3], R. Ślusarz[2], M. A. Ślusarz[2], C. Czaplewski[2], A.K. Sieradzan[2], and A. Liwo[2]

[1] – Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland, [2] – Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland , [3] – Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46 Warsaw, PL-02668, Poland

emilia.lubecka@ug.edu.pl

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:Y; Dist:N; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Contact-assisted simulations, the contacts being predicted or determined experimentally, have become very important in the determination of the structures of proteins and other biological macromolecules. In this CASP experiment, we tested the effect of contact-distance restraints on the protein-structure prediction with the use of multiplexed replica exchange molecular dynamics (MREMD)[1] with the coarse-grained UNRES force field[2]. UNRES, like others physics-based approaches, so far, is less efficient for protein-structure prediction than the knowledge-based approaches; however, owing to its independence of structural databases, UNRES is ranked quite well in the free modeling category[3].

Methods

The UNRES model of polypeptide chains reduces each amino-acid residue to a united peptide group (p), a united side chain (SC), and the α-carbon (C^{α}) atom. The C^{α} atoms are connected by virtual bonds with the united side chains, and the peptide groups are positioned in the middle between the two consecutive $C^{\alpha}s$.[2] The united side chains and united peptide groups are the only interacting sites, while the C^{α}s only assist in geometry definition. The UNRES effective energy function originates from the potential of mean force (PMF) of polypeptide chains in water, which is expressed in terms of Kubo's cluster cumulant functions[4], which correspond to smaller sections of the system and can thus be identified with effective energy terms[5]. The solvent is implicit in the force field. We have used the same version of the force field as that used by the UNRES group[5]. This force field has been improved to enhanced capacity of modeling beta-sheet structures and was already tested in CASP13 competition[3].

To run production simulations we did restrained MREMD[1] simulations with the bounded contact-distance-restraint function introduced in our recent work[6,7]. This restraint function is derived from the Lorentzian function and does not generate a gradient when a restraint cannot be satisfied. Thus, the penalty terms do not force incompatible restraints (which usually correspond to wrongly predicted contacts), preventing a simulation from producing non-protein-like structures. Contact prediction, from which the distance restraints were derived, were carried out with DNCON2[8]. DNCON2 is an improved protein contact map predictor based on two-level deep convolutional neural networks. It consists of six convolutional neural networks: the first five predict contacts at 6, 7.5, 8, 8.5 and 10 Å distance thresholds, and the last one uses these five predictions as additional features to predict final contact maps. For simple/small or non-homologous proteins, we have generated contacts directly from the top-quality server models, as determined by using the DeepQA method[9].

The structures of the target proteins were predicted by the following four-stage procedure[10]:

1. MREMD simulations were run, which consisted of 48 replicas at 12 temperatures from 260 K to 370 K, for 20,000,000 4.89 fs steps/trajectory. To speed up the search for larger proteins, additionally weak restraints were imposed on secondary structure based on secondary structure prediction by PSIPRED[11] and the conformational search was started from the top-quality server models, as determined by using the DeepQA method[9]. The initial structures of the assembly targets were generated, if possible, based on the HHPred[12] hints from homology-related assemblies. The raw PSIPRED output was used to generate bimodal restraints on backbone virtual-bond-dihedral angles, with one minimum in the alpha-helical and another one in the extended region, the well-depths depending on PSIPRED-determined probabilities[10]. The replicas were exchanged and the snapshots were collected every 10,000 MD steps.

2. The weighted-histogram analysis method (WHAM)[13] was used to calculate the relative free energy of each structure of the last section of an MREMD simulation. The last 1,000 snapshots per trajectory were processed.

3. The conformational ensembles obtained in steps 2. and 3. were subjected to cluster analysis to generate 5 (for CASP) or 10 (for CAPRI) clusters (see refs 3, and 10 for details).

4. The conformations closest to the respective average structures corresponding to the found clusters, and ranked according to the free energies of the clusters, were converted to all-atom structures using the PULCHRA[14] and SCWRL[15] algorithms. These structures were subsequently refined by using the AMBER14 package[16] with the ff14SB force field and GBSA implicit-solvent model. 500 minimization steps followed by 3,000 steps of molecular dynamics with 0.1 fs step length (0.3 ps total), with restraints on the secondary structure and positional restraints from the parent UNRES structure, and 500 final minimization steps with imposing only restraints on secondary structure were run for each structure. The refined all-atom structures were submitted to CASP or CAPRI.

In data-assisted predictions (using NMR data) penalty terms were added to the target functions. We improved our penalty functions to handle the SAXS and ambiguous NMR data developed in our earlier work[3], by introducing the dependence of the site radii in SAXS on the number of neighbors to account for the hydration shell and by developing a procedure to compute approximate positions of the protons given the coarse-grained geometry.

Results

We treated regular, assembly and data-assisted targets, following the procedure described in Methods. We did not treat large assembly targets: H1081, T1099, H1097, H1060, H1044, due to very complex and complicated structures of these targets, and also large resource requirements and short time.

We postpone the assessment of the approach until the official release of CASP14 results.

Availability

The standalone version of UNRES is available from www.unres.pl and the web server version is available at http://unres-server.chem.ug.edu.pl.

1. Czaplewski,C., Kalinowski,S., Liwo,A. & Scheraga,H.A. (2009). Application of multiplexed replica exchange molecular dynamics to the UNRES force field: Tests with α and α+β proteins. J. Chem. Theory Comput. 5, 627-640.
2. Liwo,A., Czaplewski,C., Ołdziej,S., Rojas,A.V., Kaźmierkiewicz,R., Makowski,M., Murarka, R.K. & Scheraga,H.A. (2008). Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In: Coarse-Graining of Condensed Phase and Biomolecular Systems., ed. G. Voth, Taylor & Francis, Chapter 8, pp. 107-122.
3. Lubecka,E.A., Karczyńska,A.S., Lipska,A.G., Sieradzan,A.K., Zięba,K., Sikorska,C., Uciechowska,U., Samsonov,S.A., Krupa,P., Mozolewska,M.A., Golon,Ł., Giełdoń,A., Czaplewski,C., Ślusarz,R., Ślusarz,M., Crivelli,S.N. & Liwo,A. (2019) Evaluation of the scale-consistent UNRES force field in template-free prediction of protein structures in the CASP13 experiment. J. Mol. Graph. Model. 92, 154-166.
4. Kubo,R. (1962). Generalized cumulant expansion method. J. Phys. Soc. Japan 17, 1100-1120.
5. Sieradzan,A.K., Makowski,M., Augustynowicz,A. & Liwo,A. (2017) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. J. Chem. Phys. 146, 124106.
6. Sieradzan,A.K. & Jakubowski,R. (2017). Introduction of steered molecular dynamics into UNRES coarse-grained simulations package. J. Comput. Chem. 38, 553e562.
7. Lubecka,E.A. & Liwo,A. (2019) Introduction of a bounded penalty function in contact-assisted simulations of protein structures to omit false restraints, J. Comput. Chem. 40, 2164-2178.
8. Adhikari, B., Hou, J. & Cheng, J. (2018) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 34, 466-1472.
9. Cao,R., Bhattacharya,D., Hou, J. & Cheng,J. (2016) DeepQA improving the estimation of single protein model quality with deep belief networks. BMC Bioinf. 17, 495.
10. Krupa,P., Mozolewska,M.A., Wiśniewska,M., Yin,Y., He,Y., Sieradzan,A.K., Ganzynkowicz,R., Lipska,A.G., Karczyńska,A. Ślusarz,M., Ślusarz,R., Giełdoń,A., Czaplewski,C., Jagieła,D., Zaborowski,B., Scheraga,H.A. & Liwo,A. (2016) Performance of protein-structure predictions with the physics-based UNRES force field in CASP11. Bioinformatics, 32, 3270-3278.
11. McGuffin,L.J., Bryson,K. & Jones,D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics, 16, 404-405.
12. Söding,J., Biegert,A. & Lupas,A.N. (2005) The HHPred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244−W248.
13. Kumar,S., Bouzida,D., Swendsen,R.H., Kollman,P.A. & Rosenberg,J.M. (1992) The weighted histogram analysis method for free energy calculations on biomolecules. I. The method. J. Comput. Chem. 13, 1011-1021.
14. Rotkiewicz,P. & Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem., 29, 1460-1465.

15. Wang,Q., Canutescu,A.A. & Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. Nat. Protoc. 3,1832-1847.
16. Case,D.A. et al. (2014), AMBER 14, University of California, San Francisco.

UNRES-template

Template-assisted prediction of protein structures with the coarse-grained UNRES force field and replica-exchange molecular dynamics

P.A. Wesołowski[1,2], A. Antoniak[1], M. Kogut[1], M. Maszota-Zieleniak[1], K.K. Bojarski[1], E.A. Lubecka[3], A.G. Lipska[1], M.M. Kogut[1], I. Biskupek[1], A. Giełdoń[1], K. Zięba[1], A.K. Sieradzan[1], P. Krupa[4], M. Marcisz[1], S.A. Samsonov[1], R. Ślusarz[1], M.J. Ślusarz[1], A. Liwo[1], and C. Czaplewski[1]

[1] – Faculty of Chemistry, University of Gdańsk, Wita Stwosza 63, 80-308 Gdańsk, Poland, [2] – Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, ul. Abrahama 58, 80-307 Gdańsk, Poland, [3] – Faculty of Mathematics, Physics, and Informatics, University of Gdańsk, Wita Stwosza 57, 80-308 Gdańsk, Poland, [4] – Institute of Physics, Polish Academy of Sciences, Aleja Lotników 32/46 Warsaw, PL-02668, Poland

cezary.czaplewski@ug.edu.pl

Key: Auto:N; CASP_serv:Y; Templ:Y; MSA:N; Fragm:Y.9; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y

We tested, with the CASP14 targets, our hybrid approach for protein-structure prediction, which combines the physics-based coarse-grained UNRES model of polypeptide chains[1] with knowledge-based information from templates (selected CASP-hosted server predictions) and uses multiplexed replica exchange molecular dynamics (MREMD) [2] as the main conformational-search engine. The method implements restraints from the consensus fragments common to server models. With respect to the last CASP, we have updated the version of the UNRES force field [3,4] and extended our methodology to treat large proteins and protein complexes and to handle the data-assisted targets better.

Methods

UNRES is a heavily reduced model of polypeptide chains, in which a chain is represented by a sequence of a-carbon (C^a) atoms, connected by virtual bonds, with attached united side chains. Two interaction sites are assigned to each amino-acid residue: the united peptide group (p) located in the middle of two consecutive C^a atoms and the united side chain (SC). The C^a atoms are not interaction sites but serve to define chain geometry. The interactions are described by the UNRES potential-energy function, derived from the generalized cluster-cumulant expansion of the potential of mean force (PMF) of polypeptide chains in water, in which all degrees of freedom not present in the model are integrated out.

The structures of the target proteins were predicted by the following five-stage procedure[5]:

1. Top models from CASP-hosted server predictions (stage 2) were selected using DeepQA quality assessment[6]. The selected models were processed to extract the consensus (similar in geometry) fragments and, subsequently, to determine the geometry restraints from these fragments. The geometric restraints were imposed on the $C^{\alpha}\cdots C^{\alpha}$ distances, the backbone-virtual-bond angles θ, the backbone virtual-bond-dihedral angles γ, and the local coordinates of the side-chain-direction vectors. The restraint-penalty function consists of log-Gaussian quasi-harmonic terms.

In addition, pseudopotentials of the Dynamic Fragment Assembly (DFA) approach[7] derived from 9-residue fragments selected from the fragment library specific for the sequence under investigation were determined and added to the UNRES energy function.

2. MREMD simulations with the pseudoenergy function consisting of the UNRES force field, DFA pseudopotentials, and the restraint terms determined from the selected server models, were run using 48 replicas at 12 temperatures from 260 K to 370 K, for 20,000,000 4.89 fs steps/trajectory. The conformational search was started from the top-quality server models selected at stage 1. The initial structures of the assembly targets were generated, if possible, based on the HHPred[8] hints from homology-related assemblies. The replicas were exchanged and the snapshots were collected every 10,000 MD steps.

3. The weighted-histogram analysis method (WHAM)[9] was used to calculate the relative free energy of each structure of the last section of an MREMD simulation. The last 1,000 snapshots per trajectory were processed.

4. The conformational ensembles were subjected to cluster analysis to generate 5 (for CASP) or 10 (for CAPRI) clusters (see ref. 5 for details).

5. The conformations closest to the respective average structures corresponding to the found clusters, and ranked according to the free energies of the clusters, were converted to all-atom structures using the PULCHRA[10] and SCWRL[11] algorithms. These structures were subsequently refined by using the AMBER14 package[12] with the ff14SB force field and GBSA implicit-solvent model. 500 minimization steps followed by 3,000 steps of molecular dynamics with 0.1 fs step length (0.3 ps total), with restraints on the secondary structure and positional restraints from the parent UNRES structure, and 500 final minimization steps with imposing only restraints on secondary structure were run for each structure. The refined all-atom structures were submitted to CASP or CAPRI.

We introduced the necessary modification into the UNRES conformational-search engine to treat large proteins and protein complexes: distance cut-off on long-range interactions with the interaction list distributed to slave tasks, transforming the inertia matrix in MD to a five-diagonal form and implementation of the LBFGS energy-minimization algorithm[13]. These modifications resulted in linear scaling of the CPU and memory requirements with system size, as opposed to scaling with the square of the system size in earlier versions of UNRES.

We also improved the penalty functions to handle the SAXS and ambiguous NMR data developed in our earlier work[14,15], by introducing the dependence of the site radii in SAXS on the number of neighbors to account for the hydration shell and by developing a procedure to compute approximate positions of the protons given the coarse-grained geometry. To handle the refinement targets, we used normal model analysis to determine the flexibility of the chain and impose restraints accordingly.

Results

We treated all regular, assembly, refinement, and data-assisted targets, following the procedure described in Methods. For the two large assembly targets: H1081 (a 20-mer, over 12,000 residues total) and T1099 (a 240-mer virus capsid, over 60,000 residues total), due to large resource requirements, we ran only short MREMD simulations (H1081) or canonical MD simulations at 300K (T1099). We postpone the assessment of the approach until the official release of CASP14 results.

Availability

The standalone version of UNRES capable of doing template-assisted simulations is available from www.unres.pl.

1. Liwo,A., Czaplewski,C., Ołdziej,S., Rojas,A.V., Kaźmierkiewicz,R., Makowski,M., Murarka, R.K. & Scheraga,H.A. (2008) Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In: Coarse-Graining of Condensed Phase and Biomolecular Systems., ed. G. Voth, Taylor & Francis, Chapter 8, pp. 107-122.
2. Czaplewski,C., Kalinowski,S., Liwo,A. & Scheraga,H.A. (2009) Application of multiplexed replica exchange molecular dynamics to the UNRES force field: Tests with α and $\alpha+\beta$ proteins. J Chem. Theory Comput. 5, 627-640.
3. Sieradzan,A.K., Makowski,M., Augustynowicz,A. & Liwo, A. (2017) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. I. Backbone potentials of coarse-grained polypeptide chains. J. Chem. Phys., 146, 124106.
4. Liwo,A., Sieradzan,A.K., Lipska,A.G., Czaplewski,C., Joung,I., Żmudzińska,W., Hałabis,A. & Ołdziej.S. (2019) A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. J. Chem. Phys., 150, 155104.
5. Karczyńska,A.S., Zięba,K., Uciechowska,U., Mozolewska,M.A., Krupa,P., Lubecka,E.A., Lipska,A.G., Sikorska,C., Samsonov,S.A., Sieradzan, A.K., Giełdoń,A., Liwo,A., Ślusarz,R., Ślusarz,M., Lee,J., Joo,K., Czaplewski, C. (2020) Improved consensus-fragment selection in template-assisted prediction of protein structures with the UNRES force field in CASP13. J. Chem. Inf. Model.,60, 1844-1864.
6. Cao,R., Bhattacharya,D., Hou, J. & Cheng,J. (2016) DeepQA improving the estimation of single protein model quality with deep belief networks. BMC Bioinf. 17, 495.
7. Sasaki, T. N. & Sasai, M. (2004). A coarse-grained langevin molecular dynamics approach to protein structure reproduction. Chemical Physics Letters 402, 102-106.
8. Söding,J., Biegert,A. & Lupas,A.N. (2005) The HHPred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33, W244－W248.
9. Kumar,S., Bouzida,D., Swendsen,R.H., Kollman,P.A. & Rosenberg,J.M. (1992) The weighted histogram analysis method for free energy calculations on biomolecules. I. The method. J. Comput. Chem. 13, 1011-1021.
10. Rotkiewicz,P. & Skolnick,J. (2008) Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem., 29, 1460-1465.
11. Wang,Q., Canutescu,A.A. & Dunbrack,R.L. (2008) SCWRL and MolIDE: Computer programs for side-chain conformation prediction and homology modeling. Nat. Protoc. 3,1832-1847.
12. Case,D.A. et al. (2014), AMBER 14, University of California, San Francisco.
13. Lui,D.C. & Nocedal,J. (1989) On the limited memory BFGS method for large scale optimization, Math. Program., 45, 503-528.

14. Karczyńska,A.S., Mozolewska,M.A., Krupa,P., Giełdoń,A., Liwo,A., Czaplewski,C. (2018) Prediction of protein structure with the coarse-grained UNRES force field assisted by small X-ray scattering data and knowledge-based information. Proteins 86 (S1), 228-239.
15. Lubecka,E.A., Karczyńska,A.S., Lipska,A.G., Sieradzan,A.K., Zięba,K., Sikorska,C., Uciechowska,U., Samsonov,S.A., Krupa,P., Mozolewska,M.A., Golon,Ł., Giełdoń,A., Czaplewski,C., Ślusarz,R., Ślusarz,M., Crivelli,S.N. & Liwo,A. (2019) Evaluation of the scale-consistent UNRES force field in template-free prediction of protein structures in the CASP13 experiment. J. Mol. Graph. Model., 92, 154-166.

Modeling CAPRI Targets 164 – 181 and Oligomeric CASP14 Targets by Template-Based and Free Docking

Petras J. Kundrotas, Amar Singh, and Ilya A. Vakser
Computational Biology Program and Department of Molecular Biosciences, University of Kansas, Lawrence, KS
vakser@ku.edu and pkundro@ku.edu

***Key**: Auto:N; CASP_serv:Y; Templ:Y; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y*

Most proteins in interactome have to be models of often limited accuracy.[1] Thus, joint CASP-CAPRI rounds provide a unique opportunity to test the ability of the existing docking procedures to predict the structure protein-protein complexes in a broad range of structural accuracy by utilizing models generated by the CASP participants. The CASP14-CAPRI50 round involved 18 targets with different oligomeric states and different levels of modeling difficulty. In addition, 11 oligomeric targets were offered only through the CASP pipeline.

Methods

The initial target alignments were performed by HHpred.[2,3] Protein structures were generated by NEST from JACKAL package[4] or selected from CASP Stage 2 models. The template-based docking used TM-align[5] for structural alignment with the combined scoring.[6] The free docking was performed by FFT-based GRAMM,[7] followed by AACE18 scoring[8] and constraints generated by text-mining.[9] The structure refinement was performed by TINKER[10] with CHARMM22 forcefield.[11]

Results

At the time of the abstract submission the assessment results were not available. Thus, in this abstract we focused on the modeling protocols. If HHpred alignment of the protein-protein target had > 90% probability according to the HHpred and covered > 80% of the target sequence, we utilized NEST program from the JACKAL package to build the protein models for the docking. Otherwise, all CASP Stage 2 models of the individual proteins, except those with the loose packing, were used for the docking. The statistics on protein-protein targets with the HHpred templates is in Table 1. The CASP-only targets, generally, had less templates for modeling of the individual proteins. For the template-based docking, the structure alignment by TM-align was scored by a combination of structure similarity metrics, normalized AACE18 values for the interface, fraction of shared target/template contacts, target/template interface sequence identity, interface solvation score, and the extent of clashes in the unrefined predictions. The template free docking by GRAMM was performed at lower resolution (3.5 Å grid step) in order to accommodate the structural inaccuracies of the modeled proteins. The predicted matches were scored by the AACE18 potential. The automated text-mining procedure were used to identify the binding site residues, which served as additional docking constraints. All final predictions were minimized by TINKER.

Table 1. Statistics on CASP14-CAPRI50 targets: HHpred templates with probability > 90%.

CAPRI target	CASP target	Name of protein(s)	Organism	Assembly	Experimental method	Number of residues	Number of HHpred templates
T164	T1032	smchD1	Human	A2	X-ray	284	107
T165	H1036	Glycoprotein gB/antibody 93k	Varicella-zoster virus/human	A3B3C3	EM	622/128/106	9/250/250
T166	H1045	PEX4/PEX22	*Arabidopsis Thaliana*	A1B1	X-ray	157/173	22
T167	T1050	ATPase	*Bacteroides Ovatus*	A2	X-ray	779	250
T168	T1052	Tail spike protein	Salmonella phage epsilon15	A3	X-ray	832	250
T169	T1054	Outer-membrane lipoprotein	*Acinetobacter baumannii*	A2	X-ray	190	18
T170	H1060	tail subcomplex	T5 phage	A6B3C12D6	EM	464/298/140/204	11/1/4/6
T171	T1063	CCNB1IP1	Human	A4	X-ray	196	208
T172	H1066	CCPol/MP-2	-	A1B1	X-ray	366/123	11/0
T173	H1069	CCPol/MP-1	-	A1B1	X-ray	369/122	11/0
T174	T1070	Tail spike protein	Escherichia virus CBA120	A3	X-ray	335	2
T175	T1073	DUF4423	*Bdellovibrio bacteriovorus*	A4	X-ray	255	250
T176	T1078	Tsp1	*Trichoderma virens*	A2	X-ray	138	1
T177	H1081	Arginine decarboxylase	*Providencia stuarti*	A20	EM	758	250
T178	T1083	Nitro	*Nitrosococcus oceani*	A2	X-ray	98	0
T179	T1087	Tuna	*Methylobacter tundripaludum*	A2	X-ray	93	1
T180	T1099	Capsid protein	Duck hepatitis B virus	A?	EM	262	4
T181	H1103	Orf3a-HMOX1	SARS2-Human	A1B1	X-ray	275/288	1/37
CASP only oligomeric targets							
	T1034	BIL2	*Tetrahymena thermophila*	A4	X-ray	156	0
	T1038	TSWV glycoprotein	Tomato spotted wil virus	A2	X-ray	199	0
	H1047	FlgH-FlgI	*Shigella sonnei*	A1B1?	EM	232/365	1/0
	T1048	HD_1495	*Haemophilus ducreyi*	A4	X-ray	109	0

T1061	tail subcomplex	T5 phage	A3	EM	949	192
T1062	tail subcomplex	T5 phage	A3	EM	35	0
H1065	Cytosine Methyltransferase	*Serratia marcescens*	A1B1	X-ray	127/98	0/0
H1072	SYCE2/TEX12	Human	A2B2	X-ray	101/69	0/2
T1080	Bd3182	*Bdellovibrio bacteriovorus*	A3	X-ray	922	129
T1084	Meio	*Meiothermus silvanus*	A2	X-ray	73	0
H1097	AR9	*Bacillus phage PBS1*	ABCDE	EM	426/631/496/665/464	14/16/0/12/0

For *n*-homomeric targets, we performed spatial rearrangement of the target protein to match the monomers in co-crystallized complexes either from the full-structure template library[12] or from an *ad hoc* library generated from PDB for a particular target. The *ad hoc* library contained structures which (a) were identified by the HHpred as the likely templates (> 90% probability) and (b) had oligomeric state in the biounit corresponding to that of the target. For target T177/H1081, due to anticipated conformational changes and large size of the putative interface, the free docking of the two 10-mers was performed with the C^a atoms only.

For *n*-heteromeric targets, we looked for common HHpred templates, when the templates for the target monomers were identified either as interacting chains in a PDB entry, or non-overlapping parts of the same chain. If no reliable templates were found, we performed free docking, including cross-docking of all selected CASP stage 2 models.

Availability

The docking procedures and datasets used in this round are partially available at http://vakser.compbio.ku.edu/main/resources.php.

1. Anishchenko,I., Kundrotas,P.J., Vakser,I.A. (2017). Modeling complexes of modeled proteins. Proteins. 85, 470-478.
2. Remmert,M., Biegert,A., Hauser,A., Soding,J. (2011). HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 9, 173-175.
3. Soding,J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. 21, 951-960.
4. Petrey,D., Xiang,Z.X., Tang,C.L., et al. (2003). Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins. 53, 430-435.
5. Zhang,Y., Skolnick,J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309.
6. Kundrotas,P.J., Anishchenko,I., Badal,V.D., Das,M., Dauzhenka,T., Vakser,I.A. (2018). Modeling CAPRI targets 110-120 by template-based and free docking using contact potential and combined scoring function. Proteins. 86 Suppl 1, 302-310.
7. Vakser,I.A. (1995). Protein docking for low-resolution structures. Protein Eng. 8, 371-377.
8. Anishchenko,I., Kundrotas,P.J., Vakser,I.A. (2018). Contact potential for structure prediction of proteins and protein complexes from Potts model. Biophys J. 115, 809-821.
9. Badal,V.D., Kundrotas,P.J., Vakser,I.A. (2015). Text mining for protein docking. PLoS Comput Biol. 11, e1004630.
10. Ren,P., Wu,C., Ponder,J.W. (2011). Polarizable atomic multipole-based molecular mechanics for organic molecules. J Chem Theory Comput. 7, 3143-3161.
11. MacKerell,A.D., Banavali,N., Foloppe,N. (2000). Development and current status of the CHARMM force field for nucleic acids. Biopolymers. 56, 257-265.
12. Anishchenko,I., Kundrotas,P.J., Tuzikov,A.V., Vakser,I.A. (2015). Structural templates for comparative protein docking. Proteins. 83, 1563-1570.

Modeling of Protein Complexes in CASP14 and CAPRI Round 50

J. Dapkūnas, K. Olechnovič and Č. Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University
justas.dapkunas@bti.vu.lt, kliment.olechnovic@bti.vu.lt, ceslovas.venclovas@bti.vu.lt

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y; Fragm:N; Cont:Y; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y*

During the CASP14 experiment we continued exploring the capabilities and challenges in structural modeling of protein complexes[1]. To generate models for multimeric protein structures we applied a variety of template-based modeling and protein-protein docking methods, developed both in our laboratory and elsewhere.

Methods

In CASP14 we used the same general modeling workflow as in CASP13[1], introducing several improvements. We aimed to identify multimeric templates for every target. If traditional sequence-based search with PPI3D[2] and HHpred[3] web servers was not successful, we additionally employed structure-based searches by submitting CASP server models to the DALI server[4]. When templates were identified, structural models were generated using MODELLER plugin AltMod[5,6]. Coiled-coil targets were modeled by a threading procedure: structure models were automatically generated based on the same manually selected template and all possible target-template alignments, followed by model selection.

When no templates could be found for protein complexes or the identified templates were not reliable, free docking of top 5 selected monomeric CASP server models was done by Hex[7] for hetero-complexes and Sam[8] for homomultimers. For some larger target protein complexes templates were available only for some of the subunits or domains. In these cases a hybrid strategy was used, where homology models were generated for a part of the complex and other subunits were docked to it either simply using TM-align[9] or by free docking.

For model selection we utilized VoroMQA[10] taking into account both global and interaction interface scores as described previously[1] with some modifications: new VoroMQA-dark method for global structure evaluation and improved tournament-based ranking algorithm. Standard automated procedure was used to select the best template-based models and 10 best models in the CAPRI scoring challenge. In the cases of free docking the top 100-500 selected models were subsequently relaxed by a very short molecular dynamics simulation using OpenMM[11] and then re-ranked. Free docking models were also selected according to constraints obtained from literature search or CASP contacts prediction servers, if such data were available. All models, resulting from both template-based modeling and free docking, were visually inspected before submission and manual ranking adjustments were introduced, if necessary.

Results

Templates, using either sequence or structure search strategy, were identified for 14 of 30 CASP14 multimeric targets (11 of 18 CAPRI targets). Threading was used for 3 CASP targets. Free docking was applied for 7 targets (4 CAPRI targets). Large protein complexes (6 CASP, 3 CAPRI targets), for which only partial templates were available, were modeled by a combination of template-based and free docking.

Availability

The PPI3D web server is available at http://bioinformatics.ibt.lt/ppi3d/. The VoroMQA web application is available at http://bioinformatics.ibt.lt/wtsam/voromqa. Standalone VoroMQA software for Linux and macOS is included in the Voronota package available from http://bitbucket.org/kliment/voronota/downloads.

1. Dapkūnas,J., Olechnovič,K. & Venclovas,Č. (2019) Structural modeling of protein complexes: Current capabilities and challenges. Proteins 87, 1222−1232.
2. Dapkūnas,J., Timinskas,A., Olechnovič,K., Margelevičius,M., Dičiūnas,R. & Venclovas,Č. (2017) The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures. Bioinformatics 33, 935−937.
3. Zimmermann,L., Stephens,A., Nam,S.-Z., Rau,D., Kübler,J., Lozajic,M., Gabler,F., Söding,J., Lupas,A.N. & Alva,V. (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol. 430, 2237−2243.
4. Holm,L. (2020) DALI and the persistence of protein shape. Protein Sci. 29, 128−140.
5. Šali,A. & Blundell,T.L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. J. Mol. Biol. 234, 779−815.
6. Janson,G., Grottesi,A., Pietrosanto,M., Ausiello,G., Guarguaglini,G. & Paiardini,A. (2019) Revisiting the ‘satisfaction of spatial restraints’ approach of MODELLER for protein homology modeling. PLoS Comput. Biol. 15, e1007219.
7. Ritchie,D.W. & Kemp,G.J. (2000) Protein docking using spherical polar Fourier correlations. Proteins 39, 178−194.
8. Ritchie,D.W. & Grudinin,S. (2016) Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. J Appl Cryst 49, 158−167.
9. Zhang,Y. & Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302−2309.
10. Olechnovič,K. & Venclovas,Č. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins 85, 1131−1145.
11. Eastman,P., Swails,J., Chodera,J.D., McGibbon,R.T., Zhao,Y., Beauchamp,K.A., Wang,L.-P., Simmonett,A.C., Harrigan,M.P., Stern,C.D., Wiewiora,R.P., Brooks,B.R. & Pande,V.S. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. 13, e1005659.

VoroCNN-GEMME

Deep convolutional neural network built on 3D Voronoi tessellation of protein structures with additional sequence profile information

I. Igashov[1,2], K. Olechnovič[3], M. Kadukova[1,2], Č.Venclovas[3], E. Laine[4], and S. Grudinin[1]

[1] - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, [2] - Moscow Institute of Physics and Technology, 141701 Dolgoprudniy, Russia, [3] - Institute of Biotechnology Life Sciences Center Vilnius University, Sauletekio 7, Vilnius, LT 10257, Lithuania, [4] - Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB), 75005 Paris, France.

Sergei.Grudinin@inria.fr

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y; Y.MetaG: N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP14, we have tested several variations of our new pipeline VoroCNN[1]. VoroCNN is a single-model QA method based on a deep convolutional neural network that processes protein molecules represented as undirected weighted graphs. VoroCNN-GEMME operates on geometric information retrieved from 3D Voronoi tessellation of a protein model and also on additional information from protein sequence profiles.

Methods

Our method operates on a three-dimensional protein graph. In this graph, the nodes correspond to atoms, while the edges correspond to the covalent bonds and contact surface areas between the atoms computed using the Voronota[2] framework. Each node is associated with several geometric features that include the atom type, the volume of the corresponding Voronoi cell, the solvent-accessible surface area, and the buriedness (graph distance to the nearest solvent-accessible atom). Additionally, we compute a 20-dimensional (co-)evolutionary descriptor for each residue using the GEMME[3] tool. Briefly, GEMME predicts the extent to which a mutation at this sequence position to each of the other amino acids would be deleterious. The input for GEMME is a multiple sequence alignment (MSA) containing natural sequences sharing some similarity with the query sequence. The edges are split into several non-overlapping groups corresponding to the types of chemical covalent bonds, and also to sequence-separation values for the noncovalent edges.

To process these graphs, we constructed a deep convolutional neural network and trained it on local CAD-scores[4]. Our network consists of graph convolution layers and one pooling layer in the middle. The graph convolution layer is based on the message-passing concept[5]. The pooling layer converts the atom-level representation of the graph into the residue-level representation. After the pooling layer is applied, a vector of residues' evolution descriptors is stacked to the current feature matrix. To obtain a prediction of a model's global CAD-score, we average local scores predicted by the network.

We trained the VoroCNN-GEMME network on the CASP[8-11] datasets and validated it on the CASP12 dataset. For training, the data from CASP[8-11] was preliminarily refined: we removed excessive models' parts and filtered out targets of low quality (based on VoroMQA[6] predictions). To enrich the data with more near-native examples, we generated additional near-native conformations using the NOLB[7] library.

Results

In the CASP14 challenge, we applied VoroCNN-GEMME to the QA category of targets using a server model. Although all basic operations in the VoroCNN pipeline, such as building the graph, computing the geometric descriptors, and making a prediction, are automated, in the VoroCNN-GEMME modification, human intervention was involved. Specifically, we needed to manually run GEMME for computing the evolution descriptors on targets as soon as these targets were published on the CASP website. We will automate this operation in the next edition of our server.

Availability

The standalone application implemented in Python is freely available at https://gitlab.inria.fr/grudinin/vorocnn/ and supported on Linux and MacOS (10.15 and up).

1. Igashov,I., Olechnovič,K., Kadukova,M., Venclovas,Č., & Grudinin,S. (2020). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. bioRxiv.
2. Olechnovič,K., & Venclovas,Č. (2014). Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry, 35(8), 672-681.
3. Laine,E., Karami,Y., & Carbone,A. (2019). GEMME: a simple and fast global epistatic model predicting mutational effects. Molecular biology and evolution, 36(11), 2604-2619.
4. Olechnovič,K., Kulberkytė,E., & Venclovas,Č. (2013). CAD - score: a new contact area difference - based function for evaluation of protein structural models. Proteins: Structure, Function, and Bioinformatics, 81(1), 149-162.
5. Gilmer,J., Schoenholz,S.S., Riley,P.F., Vinyals,O., & Dahl,G.E. (2017). Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212.
6. Olechnovič,K., & Venclovas,Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins: Structure, Function, and Bioinformatics, 85(6), 1131-1145.
7. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. Journal of chemical theory and computation, 13(5), 2123-2134.

VoroCNN-select, Ornate-select, SBROD-select

Tertiary and quaternary structure prediction in CASP14 using a combination of physics-based approaches with machine learning

A. Karczynska[1] and S. Grudinin[1]
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
Sergei.Grudinin@inria.fr

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In the current CASP14 experiment, we participated as three human groups for the tertiary and quaternary structure prediction (the TS category of targets). These were VoroCNN-select, Ornate-select, and SBROD-select groups. Each of these groups followed the same protocol but used different quality assessment (QA) methods, VoroCNN-GEMME[1], Ornate[2], and SBROD[3], respectively. For the predictions of the assemblies, we extensively used the symmetry assembler SAM[4] and the binary docking method Hex[5].

Multimeric and one SAXS-assisted target also motivated us to develop novel methods specifically adapted to these targets. For example, we have extended Pepsi-SAXS[10] for rapid computation of scattering profiles of symmetric assemblies, we extended the SAM symmetry assembler for helical symmetries, we introduced new options into the symmetry analyzer AnAnaS[11,12], we developed a novel rigid-body replica-exchange Markov-chain Monte Carlo simulation technique, we introduced more options, specifically, symmetry constraints, into the interactive docking engine[13], and more.

Methods

For the monomeric targets, we rescored the stage-2 server models using the corresponding QA method, VoroCNN-GEMME[1], Ornate[2], and SBROD[3] and submitted the top-5 predictions from each group.

For the homo-oligomeric targets, we used the following general protocol. Firstly, we selected the 5 best server models of the monomers ranked by the corresponding QA method. Then, we ran the SAM symmetry assembler on each model specifying the desired symmetry. For example, for the tetramers, we examined both C4 and D2 symmetries. We generated about 500 docking poses per model, selected according to the SAM shape-complementarity score per monomer. Then we rescored all the oligomeric predictions using the corresponding QA model. Finally, we submitted the best assembly prediction per server-provided template for the monomer.

For the hetero-dimers, we used the following protocol. Firstly, we selected the 5 best server models of the monomers ranked by the corresponding QA method. Then, we docked each pair of the server models using Hex with shape complementarity as the energy function. We stored 2,000 of docking poses for the subsequent rescoring. After, we rescored all the dimers with the corresponding QA model. Finally, we submitted the best-ranked complexes for each pair of the server models.

For the SAXS-assisted S1063 tetrameric target we adopted the following protocol. Firstly, we exhaustively generated symmetric C4 and D2 assemblies for all 150 stage-2 server submissions using SAM symmetry assembler and storing the top 4,000 solutions for each of the monomers. This resulted in 1.2M of docking poses. Then, we rescored all of them with respect to the experimental SAXS profile. For this purpose, we extended our Pepsi-SAXS[10] method of the computation of scattering profiles for the cases of symmetrical assemblies. More specifically, we made use of the arithmetics in the Fourier space to rapidly scan many rigid-body transformations, and also introduced a rapid computational of the solvation shell. All this allowed us to compute and score 1.2M of scattering profiles in about 3 hours on a personal laptop. In these calculations, we used the polynomial expansion order of 25 and fixed values of the adjustable parameters, the excess density of the solvation shell of 5%, and no scaling of the atomic volumes. Then, we chose the 100 best assemblies (according to Chi2) and used Pepsi-SAXS again with the domain flexibility switched on. This allowed to locally adjust the positions of the monomers and further improve the goodness of fit. Group Ornate-select submitted top-5 C4 predictions after the refinement of the domain positions. Group SBROD-select submitted top-5 D2 predictions after the refinement of the domain positions. Group VoroCNN-select submitted the predictions prior to the refinement step.

For the H1036 target with the stoichiometry $(ABC)_3$, firstly we ran our trimeric docking algorithm DockTrina for the trimer ABC. Initial models for the subunit A were taken from the T1036s1 stage-2 server predictions ranked by the corresponding QA method. Subunits B and C were modeled with iTASSER[6] v.5.1. Then, we applied SAM symmetry docking using the C3 symmetry on top of the obtained DockTrina predictions to generate the required stoichiometry.

For the heteromeric targets beyond trimers, we developed a novel docking algorithm based on rapid rigid-body replica-exchange Markov-chain Monte Carlo (MCMC) simulations. These targets included H1044 (9 domains), and H1097 (5 domains). We used pairwise scores precomputed by Hex for the top-100K relative orientations of one domain with respect to another as the potential energy. We also used 8 replicas and several millions of MCMC steps for each of the simulations. For the H1044 target, we introduced additional distance restraints between the domains. Initial models for the domains were taken from the stage-2 server predictions ranked by the corresponding QA method. Final models of H1097 were ranked with the corresponding QA methods, otherwise, assemblies with the best MCMC energies were submitted.

For the H1047 and H1072 targets, we firstly applied the hetero-dimer docking protocol. Then, we additionally applied SAM symmetry assembler to the top-5 dimeric assemblies using C2, C3, C4, C5, C6, D2, and D3 symmetries for the H1047 target, and only C2 symmetry for H1072. Finally, we ranked the predictions on a per-monomer SAM score basis for H1047 and using the corresponding QA models for H1072.

The H1103 target was docked using an interactive in-house docking application with the principal normal mode (computed with NOLB[8]) activated for the HMOX1 dimer, PDB:1n3u, docked with the rigid Orf3a dimer, PDB:6xdc. We used precomputed initial positions of HMOX1 with respect to Orf3a, followed by local minimization with the KSENIA[9] potential and side-chain repacking. We ranked the predictions according to the KSENIA scores.

The H1099 target was modeled using an exhaustive scan of 150 stage-2 server submissions with the SAM symmetry assembler for the 2-fold and 3-fold symmetry axes in the asymmetric subunit. Then, the predictions were supplemented with 60 icosahedral symmetry operators between the asymmetric subunits, and a local optimization with the KSENIA potential, sidechain repacking, and an interactive in-house docking application applied[13]. We ranked the predictions according to the KSENIA scores.

We modeled the H1060 and H1081 targets starting from the general protocol for the homo-oligomers. To generate a monomeric subunit of H1081 we used Swiss-Model[7] (template 2VYC). We used experimental structures for subunits in rings A and D. We applied D5 symmetry to H1081, C3 symmetry to the A and B subunits of H1060, C12 symmetry to the C subunit of H1060, and C6 symmetry to the D subunit of H1060. We stacked the two D5 dimers of H1081 and the rings A and B of H1060 using SAM extended to helical symmetries. The other rings in H1060 were stacked along the symmetry axis using the Hex docking engine with only 2 degrees of freedom active, the translation between the subunits, and the twist angle between them. We ranked the H1060 models based on the Hex docking scores. The H1081 models were additionally optimized and ranked using KSENIA.

Availability

More information about our methods can be found at https://team.inria.fr/nano-d/software.

1. Igashov,I., Olechnovic,K., Kadukova,M., Venclovas,C., and Grudinin,S. (2020) VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. bioRxiv 2020.04.27.063586; doi: https://doi.org/10.1101/2020.04.27.063586.
2. Pagès,G., Charmettant,B., & Grudinin,S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics, 35(18), 3313-3319.
3. Karasikov,M., Pagès,G., & Grudinin,S. (2019). Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics, 35(16):2801-2808.
4. Ritchie,D.W., & Grudinin,S. (2016). Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. J. Appl. Cryst. 49, 158-167.
5. Ritchie,D.W., & Kemp,G.J. (2000). Protein docking using spherical polar Fourier correlations. Proteins, 39: 178-194.
6. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J., Zhang,Y. (2015). The I-TASSER Suite: Protein structure and function prediction. Nature Methods, 12: 7-8.
7. Waterhouse,A., Bertoni,M., Bienert,S., Studer,G., Tauriello,G., Gumienny,R., Heer,FT, de Beer,TAP, Rempfer,C., Bordoli,L., Lepore,R., Schwede,T. (2018). SWISS-MODEL: homology modeling of protein structures and complexes. Nucleic Acids Res. 46 (W1), W296-W303.
8. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. Journal of chemical theory and computation, 13(5), 2123-2134.
9. Popov,P., & Grudinin,S. (2015). Knowledge of Native Protein−Protein Interfaces Is Sufficient To Construct Predictive Models for the Selection of Binding Candidates. J. Chem. Inf. Model. 55, 10, 2242−2255.
10. Grudinin,S., Garkavenko,M., & Kazennov,A. (2017). Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta Crystallographica Section D: Structural Biology, 73(5), 449-464.
11. Pagès,G., Kinzina,E., & Grudinin,S. (2018). Analytical symmetry detection in protein assemblies. I. Cyclic symmetries. Journal of Structural Biology, 203(2), 142-148.
12. Pagès,G., & Grudinin,S. (2018). Analytical symmetry detection in protein assemblies. II. Dihedral and cubic symmetries. Journal of Structural Biology, 203(3), 185-194.
13. Grudinin,S., & Redon,S. (2010). Practical modeling of molecular systems with symmetries. Journal of Computational Chemistry, 31(9), 1799-1814.

Deep convolutional neural network built on 3D Voronoi tessellation of protein structures

I. Igashov[1,2], K. Olechnovič[3], M. Kadukova[1,2], Č.Venclovas[3], and S. Grudinin[1]

[1] - Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, [2] - Moscow Institute of Physics and Technology, 141701 Dolgoprudniy, Russia, [3] - Institute of Biotechnology Life Sciences Center Vilnius University, Sauletekio 7, Vilnius, LT 10257, Lithuania

Sergei.Grudinin@inria.fr

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:Y; EMA:Y; MD:N*

In CASP14, we have tested several variations of our new pipeline VoroCNN[1]. It is a single-model quality assessment (QA) method based on a deep convolutional neural network that processes protein molecules represented as undirected weighted graphs. VoroCNN and VoroCNN-GDT operate only on geometric information retrieved from 3D Voronoi tessellation of a protein model.

Methods

Our method operates on a three-dimensional protein graph. In this graph, the nodes correspond to atoms, while the edges correspond to the covalent bonds and contact surface areas between the atoms computed using the Voronota[2] framework. Each node is associated with several geometric features that include the atom type, the volume of the corresponding Voronoi cell, the solvent-accessible surface area, and the buriedness (graph distance to the nearest solvent-accessible atom). The edges are split into several non-overlapping groups corresponding to the types of chemical covalent bonds, and also to sequence-separation values for the noncovalent edges.

To process these graphs, we constructed a deep convolutional neural network and trained it on local CAD-scores[3]. Our network consists of graph convolution layers and one pooling layer in the middle. For the VoroCNN-GDT network, we added at the end an additional 1D-convolution layer. The graph convolution layers are based on the message-passing concept[4]. The pooling layer converts the atom-level representation of the graph into the residue-level representation. The 1D-convolution layer at the end of the VoroCNN-GDT network was added in order to achieve a better smoothness of the local quality predictions along the protein sequence. To obtain a prediction of a model's global CAD-score, we average local scores predicted by the network.

We trained both VoroCNN and VoroCNN-GDT networks on the CASP[8-12] datasets and validated them on the CASP13 dataset. Prior to training, we refined the data from CASP[8-12] as follows: we removed excessive models' parts and filtered out targets of low quality (based on VoroMQA[5] predictions). To enrich the data with more near-native examples, we generated additional near-native conformations using the NOLB[6] library.

We should mention that as VoroCNN is based on learning on 3D Voronoi tessellations with graph convolutional networks, it differs significantly from the recently developed QA methods that use 3D convolutional networks on regular volumetric representations, e.g. Ornate[7]. Indeed, graph-based methods have fewer trainable parameters and a more rigorous definition of the topological and spatial relationship between the protein residues.

Results

In the CASP14 challenge, we applied VoroCNN and VoroCNN-GDT to the QA category of targets using two separate servers. All the server operations were fully automated, and the VoroCNN servers were among the fastest in the QA category of CASP14.

Availability

More details about VoroCNN can be found at https://team.inria.fr/nano-d/software/vorocnn/. The standalone application implemented in Python is freely available at https://gitlab.inria.fr/grudinin/vorocnn and supported on Linux and MacOS (10.15 and up).

1. Igashov,I., Olechnovič,K., Kadukova,M., Venclovas,Č., & Grudinin,S. (2020). VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures. bioRxiv.
2. Olechnovič,K., & Venclovas,Č. (2014). Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. Journal of computational chemistry, 35(8), 672-681.
3. Olechnovič,K., Kulberkytė,E., & Venclovas,Č. (2013). CAD ‐ score: a new contact area difference ‐ based function for evaluation of protein structural models. Proteins: Structure, Function, and Bioinformatics, 81(1), 149-162.
4. Gilmer,J., Schoenholz,S.S., Riley,P.F., Vinyals,O., & Dahl,G.E. (2017). Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212.
5. Olechnovič,K., & Venclovas,Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins: Structure, Function, and Bioinformatics, 85(6), 1131-1145.
6. Hoffmann,A., & Grudinin,S. (2017). NOLB: Nonlinear rigid block normal-mode analysis method. Journal of chemical theory and computation, 13(5), 2123-2134.
7. Pagès,G., Charmettant,B., & Grudinin,S. (2019). Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics, 35(18), 3313-3319.

VoroMQA-dark

Model Quality Assessment Using VoroMQA-dark

K. Olechnovič and Č. Venclovas
Institute of Biotechnology, Life Sciences Center, Vilnius University
kliment.olechnovic@bti.vu.lt

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

We participated in CASP14 with an automated model accuracy estimation server VoroMQA-dark, which employed a new unpublished method, also called VoroMQA-dark, that is partially based on the previously published VoroMQA[1] method (which will be referred to as VoroMQA-light).

Methods

VoroMQA-dark uses a neural network (NN) trained to predict local (per-residue) CAD-score values. The training was done using CASP 8-13 models. The targeted NN output for each residue consists of three CAD-score[2] values: CAD-score-level0, based on all the inter-residue contacts involving the central residue; CAD-score-level1, based on all the inter-residue contacts involving at least one residue from the first layer of neighbors (the direct neighbors) of the central residue; CAD-score-level2, based on all the inter-residue contacts involving at least one residue from the first two layers of neighbors (the direct neighbors and the neighbors of the direct neighbors) of the central residue.

The NN input vector for each residue is computed from the Voronoi tessellation-based contact areas and the corresponding contact potential values (the same as in VoroMQA-light). The NN input vector is assembled from several levels of residue neighborhood descriptors, containing summed contact areas and pseudo-energy values. The descriptors are computed using tessellation breadth-first search and accumulating convolution operations. As the input vector is "pre-convoluted", convolutional layers were not used in the NN architecture, only one fully-connected hidden layer was used in the final version.

The VoroMQA-dark server in CASP14 reported predicted CAD-score-level0 values (converted to distances) as per-residue accuracy estimates, and average predicted CAD-score-level0 values as global accuracy estimates.

Availability

The VoroMQA-dark method will be included in the Voronota[3] package freely available from bitbucket.org/kliment/voronota/downloads.

1. Olechnovic, K., and Venclovas, C. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins, 85(6), 1131-1145.
2. Olechnovic, K., Kulberkyte, E., and Venclovas, C. (2013) CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins, 81(1), 149-162.
3. Olechnovic, K., and Venclovas, C. (2014) Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. J Comput Chem, 35(8), 672-681.

Model Quality Assessment Using VoroMQA-light

K. Olechnovič and Č. Venclovas
Institute of Biotechnology, Life Sciences Center, Vilnius University
kliment.olechnovic@bti.vu.lt

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

We participated in CASP14 with an automated model accuracy estimation server VoroMQA-light, which employed the latest published version of VoroMQA[1] ("Voronoi diagram-based Model Quality Assessment"), our method for the estimation of protein structure quality that combines the idea of statistical potentials with the advanced use of the Voronoi tessellation of atomic balls.

Methods

Given a protein structure, it can be represented as a set of atomic balls, each ball having a van der Waals radius corresponding to the atom type. A ball can be assigned a region of space that contains all the points that are closer (or equally close) to that ball than to any other. Such a region is called a Voronoi cell. Two adjacent Voronoi cells share a set of points that form a surface called a Voronoi face. A Voronoi face can be viewed as a geometric representation of a contact between two atoms. The Voronoi cells of atomic balls may be constrained inside the boundaries defined by the solvent accessible surface (SAS) of the same balls. The procedure to construct the described surfaces is implemented as part of Voronota software[2].

In VoroMQA, inter-atomic and solvent contact areas are used to evaluate the quality of protein structural models by employing the idea of a knowledge-based statistical potential. The VoroMQA scoring function produces quality scores at different levels including atoms, residues and the full structure. The VoroMQA scoring function was not optimized or trained in any way to correspond to any reference based model quality-assessment scores: unsupervised learning was performed using experimentally determined structures of protein biological assemblies as input.

Availability

The VoroMQA web application is available at bioinformatics.ibt.lt/wtsam/voromqa. VoroMQA software for Linux is included in the Voronota package freely available from bitbucket.org/kliment/voronota/downloads.

1. Olechnovic, K., and Venclovas, C. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins, 85(6), 1131-1145.
2. Olechnovic, K., and Venclovas, C. (2014) Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. J Comput Chem, 35(8), 672-681.

VoroMQA-select

Model Selection Using VoroMQA-dark

K. Olechnovič and Č. Venclovas

Institute of Biotechnology, Life Sciences Center, Vilnius University
kliment.olechnovic@bti.vu.lt

Key: *Auto:N; CASP_serv:Y; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:N*

We participated in CASP14 with a model selection method, VoroMQA-select, registered as a regular tertiary structure prediction group.

Methods

The VoroMQA-select method employed a new unpublished method, VoroMQA-dark, that is partially based on the previously published VoroMQA[1] method. Please, see the VoroMQA-dark server abstract for more details.

VoroMQA-dark was used to score and rank the CASP-hosted server predictions. Also, a variant of VoroMQA was used to determine if model structures contained unstructured N-terminal or C-terminal regions that needed to be removed prior to evaluation: this tails-removal procedure was semi-automatic and required manual confirmation of trimming positions.

Availability

The VoroMQA-dark method will be included in the Voronota[2] package freely available from bitbucket.org/kliment/voronota/downloads.

1. Olechnovic, K., and Venclovas, C. (2017) VoroMQA: Assessment of protein structure quality using interatomic contact areas. Proteins, 85(6), 1131-1145.
2. Olechnovic, K., and Venclovas, C. (2014) Voronota: a fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls. J Comput Chem, 35(8), 672-681.

Model Quality Assessment Using VoroMQA-stout

K. Olechnovič and Č. Venclovas
Institute of Biotechnology, Life Sciences Center, Vilnius University
kliment.olechnovic@bti.vu.lt

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:N; Fragm:N; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:N; MD:N*

We participated in CASP14 with an automated model accuracy estimation server VoroMQA-stout, which used the same underlying method as our VoroMQA-dark server, but reported different scores. The VoroMQA-stout server output was based on the CAD-score-level1 predictions, while the VoroMQA-dark server output was based on the CAD-score-level0 predictions. Please, see the VoroMQA-dark server abstract for more details.

xianmingpan

Template-based Protein Structure Prediction based on PairThreading and PBEscore

Xiang Chen[*,1], Lu-Yun Wu[*,2] and Zhi-Xin Wang[1] and Xian-Ming Pan[2]

** - Equal contribution, 1 - Key Laboratory of Ministry of Education for Protein Science, School of Life Sciences, Tsinghua University, Beijing 100084, China., 2 - Key Laboratory of Bioinformatics, Ministry of Education, School of Life Sciences, Tsinghua University, Beijing 100084, China.*

pan-xm@mail.tsinghua.edu.cn

Key: *Auto:N; CASP_serv:N; Templ:Y; MSA:Y; Fragm:Y.3,9; Cont:N; Dist:N; Tors:N; DeepL:N; EMA:Y; MD:Y*

Alignment-based remote protein homology detection methods are the cornerstone of protein structure prediction, especially for template-based protein structure prediction. This is because these methods can not only recognize remoter homologous proteins as templates but also generate alignments to model. After constructing hundreds of 3D-models, energy scoring function is used to calculate the energy of these models so as to facilitate the location the lowest energy basin and the native structure on its energy landscape to select the final Top 5 models. We construct 3D-models based on PairThreading to find templates and generate alignments, RosettaCM [1] or MODELLER [2] to model 3D structures, PBEscore to select nearly native models, and NAMD [3] to refine models. PairThreading and PBEscore used here are developed by our and unpublished.

Methods

Our pipeline to construct 3D-models for a target sequence has 3 steps as fellow.

1. Remote protein homology detection by alignment-based PairThreading. There are many alignment methods for remote protein homology detection, but these methods are based on the assumption that the types of residues at different positions are independent of each other. We break this assumption and propose a method, PairThreading, based on residue pair substitution information. PairThreading obtains position-specific residue pair substitution information indirectly from the position-specific score matrices (PSSMs) rather than directly from the multiple sequence alignments (MSAs) to avoid statistical non-convergence problem. Thus, PairThreading can detect more remote homologous proteins and can generate more accurate alignments.

2. Prediction 3D-models by RosettaCM [1] or MODELLER [2]. We used RosettaCM [1] or MODELLER [2] to construct 3D-models based on the single or multiple templates and alignments generated by PairThreading with scores large than 2. RosettaCM and MODELLER are comparative modeling methods, and can model structures according to multiple templates.

3. Ranking 3D-models by PBEscore. PBEscore is a novel knowledge-based-energy scoring function, simply considering the interactions of peptide bonds, rather than, as conventionally, the residues or atoms as the most important energy contribution. It achieves the accuracy of 85% on the CASP5-8 dataset while only using 19s.

4. Refinement 3D-models by NAMD [3]. We ran the molecular dynamics to refine the Top 1 ranked models. The program CHARMM22 was used to add hydrogen atoms, N- and C-terminal patches to the Top 1 model [4]. The generated models were solvated and neutralized in a box with TIP3P water at a minimum of 13 Å between the model and the wall of the box. All simulations were run using NAMD 2.9 with periodic boundary conditions (PBC) applied. The temperature was held at 300 K while the pressure was controlled at 1 atm. The time step was set to 2 fs and the particle mesh Ewald method was applied to model the electrostatics and the van der Waals interactions cutoff was set at 12 Å. All simulations followed a three-step pre-equilibration totaling 600 ps, the last snapshots of which were chosen as the starting structures for 50 ns productive simulations without constraints.

Availability

PairThreading is available at: http://spg.med.tsinghua.edu.cn/FoldRecognition/, and PBEscore is available at: http://166.111.152.74:8888/pbe_score/.

1. Song, Y. et al. High-resolution comparative modeling with RosettaCM. Structure 21, 1735-1742, doi:10.1016/j.str.2013.08.005 (2013).
2. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology 234, 779-815, doi:10.1006/jmbi.1993.1626 (1993).

3.. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. J Comput Chem 26, 1781-1802, doi:10.1002/jcc.20289 (2005).

4 MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. The journal of physical chemistry. B 102, 3586-3616, doi:10.1021/jp973084f (1998).

Yang-Server (QA), Yang_TBM (QA)

Improved quality estimation of protein structure models using predicted interresidue distance

Lisha Ye[1], Peikun Wu[1], Jianzhao Gao[1], Jianyi Yang[1,*]
[1] *School of Mathematical Sciences, Nankai University, Tianjin, 300071, China*
yangjy@nankai.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:N; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

The QA submissions in CASP14 by Yang-Server and Yang_TBM were based on the interresidue distances predicted by our recent deep learning-based structure prediction algorithm trRosetta [1]. The major difference between the submissions of Yang-Server and Yang_TBM is the prediction methods adopted, i.e., linear regression and deep neural networks, respectively.

Methods

Our method has two modules, one is single-model based and the other is clustering-based. In the singe-model based module, we designed several distance scores to assess the agreement between the model's interresidue distance matrix and the predicted distance matrix. In addition to potential scores, other single QA methods were also included to improve the performance. When multiple models with apparent structure similarity are available (i.e., in QA stage2), clustering-based features are designed in the clustering-based module. Based on the above features, two different methods including linear regression and deep neural networks, were used to predict the global quality, from which the local quality was subsequently predicted. The submissions for stage1 and stage2 models were from the single-model based module and cluster-based module, respectively.

Results

Tests on 80 CASP13 targets suggest that our method outperforms other global QA methods in CASP13, as measured by the three standard metrics, Best_difference, GDT_Loss, and Pearson Correlation Coefficient. The local QA tests also show the advantage of our method. The ablation analysis suggests that the improved QA predictions are attributed to the predicted interresidue distance.

1. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D., Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences of the United States of America 2020, 117, 1496-1503.

Yang-Server (TS), Yang_FM, Yang_TBM (TS), CATHER

Improved protein structure prediction by restraints from deep learning and templates

Zongyang Du[1], Hong Su[1], Hong Wei[1], Jianyi Yang[1,]*
[1] *School of Mathematical Sciences, Nankai University, Tianjin, 300071, China*
yangjy@nankai.edu.cn

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:N; MD:N*

In CASP14, both *de novo* modeling and template-based modeling were used for our submissions. Interresidue contacts, distances and orientations were predicted by deep residual neural networks, which were used to build 3D structures based on restraints-guided energy minimization through the trRosetta package [1], and submitted by Yang-FM. The predicted contacts and distances were also used to improve remote-homology template detection [2] (CATHER). Template-based modeling was done based on the I-TASSER Suite [3] (Yang-TBM). The Yang-Server submissions were generated based on energy minimization with restraints from deep learning and templates.

Methods

To predict the interresidue contacts, distances and orientations, multiple MSAs were generated from different sequence databases and alignment algorithms, as detailed in the trRosetta [1] and the MapPred papers [4]. The optimal MSA was selected based on the average probability of the top L predicted long+medium-range contacts. The predicted distances and orientations were used to build 3D structure models through the trRosetta package. A total of 100 centroid models were built at different of starting structures and restraints selected at varied probability thresholds. The top 10 models with the lowest energy scores were selected for relaxation and the top 5 relaxed models were submitted by Yang-FM.

The contacts-assisted threading algorithm CATHER was improved further by replacing predicted contacts by distances. MODELLER was then used to build full-length models based on the threading alignments. The top 5 models ranked based on Z-score were submitted by CATHER. The templates detected by the original CATHER and the improved CATHER were added into the template pool of I-TASSER Suite. In addition, the Ca-Ca distances predicted by deep learning were used as additional restraints in the I-TASSER simulations to improve the template-based modeling. The resulted models were submitted by Yang-TBM.

The Yang-Server submissions were from an enhanced trRosetta energy minimization with restraints from both deep learning and templates. Two different methods were tried: one is to convert the template information as additional features and fed them into the deep neural networks to predict interresidue distances and orientations; another is to convert the template information into distance restraints to guide the energy minimization together with the default restraints from deep learning.

Results

Benchmark tests on the CASP13 datasets show that the template detection by the new CATHER was improved by about 5% in TM-score. The combination of the templates and deep learning restraints in Yang-Server yields TM-score improvement by about 10% over the default trRosetta models.

Availability

The trRosetta server and the standalone package for *de novo* protein structure modeling are available at: https://yanglab.nankai.edu.cn/trRosetta/.

1. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences of the United States of America 2020;117(3):1496-1503.
2. Du Z, Pan S, Wu Q, Peng Z, Yang J. CATHER: a novel threading algorithm with predicted contacts. Bioinformatics 2020;36(7):2119-2125.
3. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature methods 2015;12(1):7-8.
4. Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J. Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics 2020;36(1):41-48.

Protein 3D Structure Prediction by Zhang Human Group in CASP14

Yang Li[1], Wei Zheng[1], Chengxin Zhang[1], Eric Bell[1], Xiaoqiang Huang[1], Robin Pearce[1], Xiaogen Zhou[1], Yang Zhang[1,2]

1 - Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, 2 – Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091

yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:Y.v; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

The tertiary structure prediction of the Zhang human group in CASP14 is based on the D-I-TASSER pipeline, which is identical to that used by the Zhang-Server group (see Zhang-Server Abstract), except that the whole set of structure models generated by the CASP servers, instead of the in-house LOMETS templates[1], were used as the starting models of the D-I-TASSER pipeline. In addition, a bug in the MSA generation pipeline, which affected the first 22 targets in Zhang-Server (and QUARK server), was identified and corrected in the Zhang human group.

The Zhang human group structure prediction pipeline consists of four consecutive steps. First, a set of multiple sequence alignments (MSAs) are created for the target sequence by DeepMSA[2] and its variants, by iterative sequence and sequence-profile search through whole-genome and metagenome sequence databases (Metaclust, BFD, Mgnify, and IMG/M). The MSA with the highest cumulative TripletRes probability for the top $10L$ contacts (see TripletRes Abstract) is selected for the next step of modeling.

In the second step, the selected MSA is used by DeepPotential, a newly developed deep residual neural-network based predictor (see DeepPotential Abstract), to create multiple geometry restraints including (1) distance-maps for both Cα and Cβ atoms; (2) Cα-based hydrogen-bonding networks [3]; (3) Cα-Cβ torsion angles. Considering that DeepPotential tends to have higher accuracy for the distance models with shorter distance cutoffs, four sets of distance profiles are generated with distance ranges of [2, 10], [2, 13], [2, 16], and [2, 20] Å, where the four ranges are divided into 18, 24, 30, and 38 distance bins, respectively. For each distance range, only the distance profiles from lower distance cutoffs are used, i.e., distances from [2-10) Å are generated from model Set-1, distances from [10-13) Å from Set-2, [13-16) Å from Set-3, and [16-20] Å from Set-4. In addition to DeepPotential, Cα and Cβ contact-maps with a distance cutoff of 8 Å are created by three deep-learning and naïve Bayes classifier based contact predictors (TripletRes[4], ResPRE[5], and NeBcon[6]). Meanwhile, LOMETS3, a newly developed meta-server program containing both profile- and contact-based threading programs (see Zhang-TBM Abstract), is used to identify structural templates from a non-redundant PDB structural library. Based on the significance and consensus of the LOMETS3 alignments, the target is assigned to one of four categories (Triv, Easy, Hard and Very-Hard)[7].

In the third step, replica-exchange Monte Carlo (REMC) simulations for full-length protein folding were performed, guided by a composite force field of $E = E_{I-TASSER} + E_{contact} + E_{distance} + E_{HB} + E_{torsion}$. Here, $E_{I-TASSER}$ is extended from I-TASSER[3] and contains a set of statistical energy terms and spatial restraints from LOMETS3; $E_{contact}$ is a three-gradient potential to satisfy the predicted contact restraints; $E_{distance}$, E_{HB} and $E_{torsion}$ are the negative logarithm of the DeepPotential predicted probabilities for distance, hydrogen-bonding, and torsion angle maps, respectively. Three types of REMC simulations (labeled as 'A', 'M' and 'F') are run depending on a target's category, i.e., 'A' keeps all Cα atoms on-lattice with the REMC simulations starting from random conformations; 'M' freely rotates and translates fragments excised from the threading alignments; and 'F' keeps the threading-aligned fragments frozen with changes only to the unaligned regions. 'M' and 'F' are implemented only for Trivial and Easy targets whose templates have a higher confidence, since the local structures (and the global topology for 'F') in these runs are kept ragid in these runs. Five REMC simulations are performed for each simulation type. The structural decoys from 8 (or 3 for Hard and Very-Hard targets) lowest-temperature replicas are submitted to SPICKER[8] for structure clustering and model selection.

In the last step, the SPICKER cluster centroids are refined at the atomic level by fragment-guided molecular dynamic (FG-MD) simulations[9] and their side-chains are repacked by FASPR[10]. A set of six MQAP schemes, including the D-I-TASSER C-score (Zheng et al, in preparation), contact satisfaction rate, structural consensus measured by pair-wise TM-score[11], and three statistical energy functions (RW, RWplus[12], and Rotas[13]), are used to select models from the simulation results, where a meta-MQAP consensus score is calculated as the sum of the rank of the six MQAP scores. Top-five models with the lowest consensus MQAP scores are selected for submission.

For multiple-domain sequences, FUpred[14] and ThreaDom[15] are used to predict the domain boundaries and linker regions from the contact-maps and LOMETS threading alignments, respectively. Structural models are first predicted by D-I-TASSER for the individual domains separately, which are then assembled into full-length models using a rigid-body domain docking and assembly algorithm, DEMO[16], guided by the whole-chain D-I-TASSER model and structural analogs identified with TM-align[17]. The procedure is fully automated.

Availability

https://zhanglab.ccmb.med.umich.edu/I-TASSER

1. Zheng,W., Zhang,C., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2019). LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res* **47**, W429-W436.
2. Zhang,C., Zheng,W., Mortuza,S. M., Li,Y. & Zhang,Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105-2112.
3. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. & Zhang,Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7-8.
4. Li,Y., Zhang,C., Bell,E.W., Zheng,W., Zhou,X., Yu,D.J. & Zhang,Y. (2020). Deducing high-accuracy protein contact-maps from a triplet ofcoevolutionary matrices through deep residual convolutional networks. submitted.
5. Li,Y., Hu,J., Zhang,C., Yu,D.J. & Zhang,Y. (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **35**, 4647-4655.
6. He,B., Mortuza,S.M., Wang,Y., Shen,H.B. & Zhang,Y. (2017). NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics* **33**, 2296-2306.

7. Zhang,Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* **82 Suppl 2**, 175-87.
8. Zhang,Y. & Skolnick,J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25**, 865-71.
9. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**, 1784-95.
10. Huang,X., Pearce,R. & Zhang,Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758-3765.
11. Zhang,Y. & Skolnick,J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10.
12. Zhang,J. & Zhang,Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* **5**, e15386.
13. Park,J. & Saitou,K. (2014). ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics* **15**, 307.
14. Zheng,W., Zhou,X., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* **36**, 3749-3757.
15. Wang,Y., Wang,J., Li,R., Shi,Q., Xue,Z. & Zhang,Y. (2017). ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. *Nucleic Acids Res* **45**, W400-W407.
16. Zhou,X., Hu,J., Zhang,C., Zhang,G. & Zhang,Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proc Natl Acad Sci U S A* **116**, 15930-15938.
17. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-9.

Zhang_Ab_Initio

Ab Initio Protein Folding Guided by Deep Learning Predicted Distance and Orientations

Chengxin Zhang[1], Yang Li[1,2], Xiaogen Zhou[1], Wei Zheng[1], and Yang Zhang[1]

[1] - Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, [2] - School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:N; Templ:N; MSA:Y.MetaG; Fragm:Y.1-20; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

Zhang_Ab_Initio is an extension of our CASP13 "QUARK" server[1], but with its predicted contact replaced by new deep learning predicted inter-residue distances and torsion angles.

Methods

Multiple sequence alignments (MSAs) for the query sequence are generated by three approaches (DeepMSA, qMSA, and mMSA), using four metagenome sequence databases (Metaclust, BFD, Mgnify, and IMG/M) and two whole-genome sequence databases (Uniclust30 and UniRef90) (see Figure 1). Here, DeepMSA[2] is our previous MSA construction program developed in CASP13. In the three stages of DeepMSA, HHblits2, Jackhmmer and HMMsearch were used to search the query against Uniclust30 (version 2017_04), UniRef90 and Metaclust, respectively. In Stage 2 and 3, homologs identified by Jackhmmer and HMMsearch, respectively, are constructed into a custom HHblits format database, which will be searched through by HHblits2 using the MSA input from the previous stage to generate new MSAs. As an extension of DeepMSA, qMSA (standing for "quadruple MSA") has four stages to perform HHblits2, Jackhmmer, HHblits3, and HMMsearch searches against Uniclust30 (version 2020_01), UniRef90, BFD, and Mgnify, respectively. Similar to DeepMSA Stage 2 and 3, the sequence hits from Jackhmmer, HHblits3 and HMMsearch in Stage 2, 3 and 4 of qMSA are converted into HHblits format database, against which the HHblits2 search based on MSA input from the previous stage is performed. In mMSA (or "multi-level MSA"), the qMSA Stage 3 alignment is used as a probe by HMMsearch to search through the IMG/M database and the resulting sequence hits are converted into a sequence database. This mMSA database is then used as the target database, which is searched through by three programs (the final stage by DeepMSA, the full qMSA pipeline with all four stages, and a reduced qMSA pipline with only Stage 1, 2, and 4), to derive three new MSAs. These steps result in 10 MSAs in total (i.e., 3 from DeepMSA, 4 from qMSA, and 3 from mMSA), which are scored by TripletRes contact prediction[3], where the MSA with the highest probabilities for top 10L (L is the sequence length) all range contacts (Cβ-Cβ distances<8Å) will be selected.

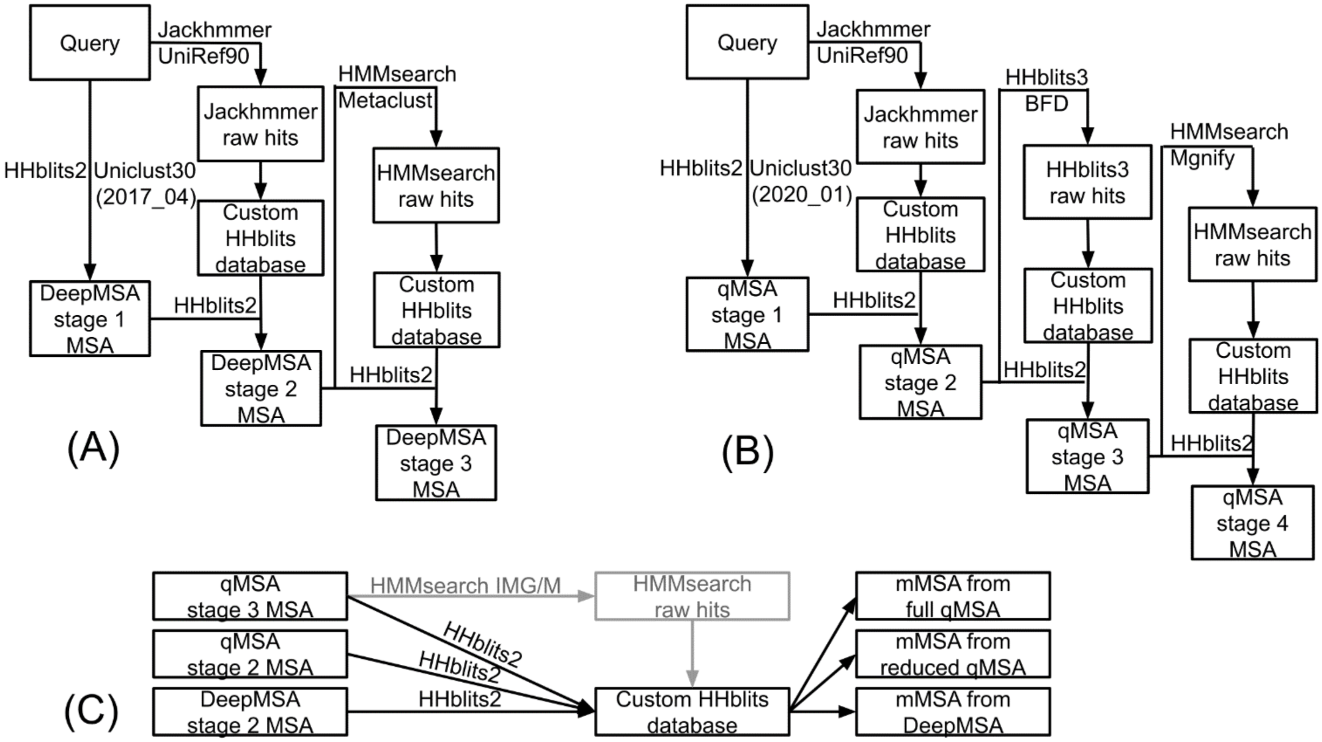


Figure 1. MSA generation by **(A)** DeepMSA, **(B)** qMSA, and **(C)** mMSA, which generates 3, 4 and 3 MSAs, respectively.

The selected MSA is used by the full-version TripletRes program to predict a set of spatial restraints, including the Cα-Cα distance, Cβ-Cβ distance and inter-residue torsional angles. The distances are predicted in the form of 38 distance bins (1 bin for <2Å, 36 bins for 2 to d_{cut}=20Å with bin width 0.5Å, and 1 bin for ≥20Å), while torsional angles are predicted with bin width of 15° plus an additional bin for no interaction (i.e. Cβ-Cβ distance ≥20Å). These distance and torsion angle restraints are used to construct initial conformations by L-BFGS gradient descent with the following distance (*d*) and torsion (*t*) potentials:

$$E(d) = -\log\left(\frac{P(d)+\epsilon}{P(d_{cut})+\epsilon}\right) + 1.57 \cdot \log\left(\frac{d}{d_{cut}}\right) \qquad (1)$$

$$E(t) = -\log\left(\frac{P(t)+\epsilon}{\epsilon}\right) \qquad (2)$$

$\epsilon = 1e - 4$ is a pseudo-count to avoid division by or logarithm of zero. Using different cutoffs ranging from to 0.55 to 0.95 for the probability of no interaction, 30 gradient descent runs were performed to generate 30 initial conformations. From these 30 conformations, continuous fragments ranging from 1 to 20 residues are extracted.

The fragments are assembled by a replica-exchange Monte Carlo (REMC) simulation in QUARK[4] guided by the same torsion angle potential shown in Eq (2), but with a different distance potential:

$$E(d) = \begin{cases} 1 \\ 1 - \frac{1}{2}\left(\left|\frac{d-\mu}{\sigma}\right| - 1\right)^2 \\ \frac{1}{2}\left(\left|\frac{d-\mu}{\sigma}\right| - 1\right)^2 \\ 0 \end{cases} \qquad (3)$$

μ and σ are the mean and standard deviation, respectively, of the distance prediction, calculated by fitting a Gaussian distribution to the predicted distance bins. REMC typically generates approximately 25,000 decoy conformations, which will be clustered by SPICKER[5]. The cluster centroids from the five largest clusters are refined by ModRefiner[6] and FG-MD[7] to get the five final models, which are ranked in descending order of the cluster size.

For multi-domain targets, the full length sequence is split into domains according to consensus domain boundaries predicted by ThreaDom[8] and FUpred[9]. Structure models are predicted for both the full length and the individual domain sequences. The full length structure is used as the template in DEMO[10] to assemble individual domains into the final model.

1. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins 2019;87(12):1149-1164.
2. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020;36(7):2105-2112.
3. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins 2019;87(12):1082−1091.
4. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge ‐ based force field. Proteins 2012;80(7):1715-1735.
5. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near ‐ native protein folds. J Comput Chem 2004;25(6):865-871.
6. Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. Biophys J 2011;101(10):2525-2534.
7. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19(12):1784-1795.
8. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 2013;29(13):i247-i256.
9. Zheng W, Zhou X, Wuyun Q, Pearce R, Li Y, Zhang Y. FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics 2020;36(12):3749-3757.
10. Zhou X, Hu J, Zhang C, Zhang G, Zhang Y. Assembling multidomain protein structures through analogous global structural alignments. Proc Natl Acad Sci USA 2019;116(32):15930-15938.

Zhang-Assembly

Template-based Protein Complex Structures Assembly using Restraints by TripletRes

Xiaogen Zhou[1], Zi Liu[1,2], Yang Li[2,1], Wei Zheng[1], Chengxin Zhang[1] and Yang Zhang[1]

[1] - Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 48109 , [2] - School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

yangzhanglab@umich.edu

Key: *Auto:N; CASP_serv:Y; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:N; DeepL:Y; EMA:Y; MD:Y*

Methods

The protein complex structures assembly of the Zhang-Assembly human group in CASP14 is based on an extended version of DEMO[1], which we originally developed for multi-domain protein structure assembly from individual domain models. Starting from the query sequence, protein complex templates are first identified by the multiple-chain threading algorithms SPRING[2] and COTH[3]; meanwhile inter-chain distances and contacts are predicted by TripletRes[4], a new deep residual convolutional neural-network based contact and distance predictor, through merging all monomeric sequences into a single chain. Next, replica-exchange Monte Carlo (REMC) simulations are performed to assemble the monomer structure models predicted by either our server groups or I-TASSER. The movement of the REMC simulations contains rigid-body rotation and translation of one of the monomers; which are guided by a composite force field consisting of templates restraints, inter-chain contacts, inter-chain distances, inter-chain steric clashes, inter-chain distance profiles extracted from top templates, and the inherent knowledge-based DEMO energy terms[1]. Finally, models with lowest energy are selected as final models.

For most targets in CASP14, we used the monomer structure models predicted by our server groups Zhang-Server, QUARK, Zhang-TBM, Zhang-CEthreader, and Zhang_Ab_Initio. When the server models are not available, the monomer models are generated by I-TASSER[5,6]. For several targets (e.g., H1045, H1072, H1081, and H1103), we also considered templates detected by the structural alignment using TM-align[7] if target monomers have high similar topologies with high TM-scores[8] to templates. In the templates searching, target monomers are structurally aligned with a template in the dimer library using TM-align to get TM-scores between the template and monomers. The harmonic mean of TM-scores of all monomers is defines as the score of a template, and the top 10 templates are selected to guide the assembly. For the multi-domain protein (i.e., H1044), domain boundaries are firstly predicted by FUpred[9]. Domain models are then modelled by I-TASSER or picked up from our server groups. Finally, DEMO is used to assemble all domains into full-length model, before performing the multi-chain structural assembly.

Availability

The DEMO server is available at https://zhanglab.ccmb.med.umich.edu/DEMO/
The SPRING server is available at https://zhanglab.ccmb.med.umich.edu/spring/
The COTH server is available at https://zhanglab.ccmb.med.umich.edu/COTH/

1. Zhou, X., Hu, J., Zhang, C., Zhang, G. & Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. Proceedings of the National Academy of Sciences 116, 15930-15938 (2019).
2. Guerler, A., Govindarajoo, B., Zhang, Y. & modeling. Mapping monomeric threading to protein–protein structure prediction. Journal of chemical information 53, 717-725 (2013).
3. Mukherjee, S. & Zhang, Y. Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19, 955-966 (2011).
4. Li, Y., Zhang, C., Bell, E.W., Yu, D.J. & Zhang, Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact‐map prediction in CASP13. Proteins: Structure, Function, Bioinformatics 87, 1082-1091 (2019).
5. Zheng, W., Zhang, C., Bell, E.W. & Zhang, Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. Future Generation Computer Systems 99, 73-85 (2019).
6. Zhang, Y. I-TASSER server for protein 3D structure prediction. BMC bioinformatics 9, 40 (2008).
7. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research 33, 2302-2309 (2005).
8. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics 57, 702-710 (2004).
9. Zheng, W. et al. FUpred: Detecting protein domains through deep-learning based contact map prediction. Bioinformatics (2020).

Zhang-CEthreader

DEthreader: protein folding using distance-guided threading and REMC simulation

Wei Zheng[1], Yang Li[1, 2], Xiaogen Zhou[1], Chengxin Zhang[1], Eric W Bell[1] and Yang Zhang[1]
1 - Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, 2 - School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094
yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

The Zhang-CEthreader server in CASP14 is based on DEthreader (Zheng et al, in preparation), a distance-guided threading program extended from the CEthreader method[1]. The pipeline includes four stages: (i) multiple sequence alignment (MSA) generation and inter-residue contact/distance prediction, (ii) contact-based and distance-based threading for template identification by CEthreader and DEthreader, (iii) a Replica Exchange Monte Carlo (REMC) simulation guided by deep-learning based residue-residue distance and hydrogen-bond network prediction, (iv) a deep-learning based QA method for ranking models.

Methods

DeepMSA[2] and qMSA are used to generate 7 multiple sequence alignments (MSAs). DeepMSA[2] is our previous MSA construction program developed in CASP13, which uses HHblits2[3], Jackhmmer and HMMsearch[4] to search the query sequence against the Uniclust30[5], UniRef90[6] and Metaclust[7] databases in three stages, respectively. qMSA is an extended version of DeepMSA, which uses HHblits2, Jackhmmer, HHblits3 and HMMsearch to search through the Uniclust30, UniRef90, BFD[8] and Mgnify[9] databases in four stages, respectively. Thus, we have five different MSAs generated from stages 1-3 of DeepMSA and stages 3-4 of qMSA, since the first two stages for both DeepMSA and qMSA generate identical MSAs. Furthermore, the MSA from qMSA stage 3 (the MSA from the BFD database) is used as a starting point for HMMsearch to search through the IMG/M[10] database, which contains more sequences than Metaclust, BFD and Mgnify. The resulting sequence hits are converted into a sequence database, which in turn is used for DeepMSA stage 3 and qMSA stage 4 to generate two additional MSAs. Finally, the 7 MSAs are put into TripletRes (Li et al, in preparation) to generate 7 sets of contact-map models. The MSA with the highest sum of top 10L (L is the length of protein) predicted contact probability is selected as the final MSA, and the corresponding TripletRes contact prediction will be selected as inputs for CEthreader and DEthreader. TripletRes will also predict the residue-residue distance distribution, torsional angles and hydrogen-bond network in addition to the contact-map. Here, the Cα-Cα and Cβ-Cβ distance distributions are predicted in the form of 38 distance bins (1 bin for <2Å, 36 bins for 2 Å to 20Å with bin width 0.5Å, and 1 bin for ≥20Å).

To create structural templates, the residue-residue contacts from TripletRes will be used for CEthreader to detect homologous templates against a non-redundant library containing ~80,000 structures from the PDB. The top 5,000 templates ranked by contact map overlap (CMO) score[1] of CEthreader will be selected as a template pool for DEthreader. The algorithm of DEthreader is extended from CEthreader, in which we added a distance-map based energy term to guide the template search through dynamic programming. Here, the predicted distance-map for the query is estimated from the TripletRes residue-residue distance distribution. For residue pair (i,j), the centroid distance value of the bin with largest predicted probability will be the estimated distance between residue i and residue j. Only the distances with the largest predicted probability that locate in 2-16Å will be used for DEthreader distance-map generation. After collecting the distances, the distance-map will be normalized by 16Å, resulting in a 0-1 real value symmetric matrix.

To perform distance-based alignments, Eigendecomposition of the distance-map is conducted, followed by the selection of the largest K Eigenvalues and corresponding Eigenvectors to calculate the K-dimensional distance Eigenvector sequences. A scoring function combining distance Eigenvectors, contact Eigenvectors, secondary structure, and profile terms is then utilized by a semi-global dynamic programming algorithm to compare the query sequence with a given template in the pool of 5,000 templates. For DEthreader, the absolute error between the predicted distance-map of the query sequence and the aligned templates is used for ranking templates. The top 40 templates detected by the four approaches (10 from each individual approach), including CEthreader with Cα-Cα contacts, CEthreader with Cβ-Cβ contacts, DEthreader with Cα-Cα distance, and DEthreader Cβ-Cβ distance, will be selected as the initial conformations for the follow-up REMC folding simulation. For extremely hard proteins where no good templates are detected (CMO<0.2), 10 additional conformations from L-BFGS gradient descent guided by the following distance (d) and torsion angle (o) potentials are added to the initial conformation set:

$$E(d) = -\log\left(\frac{P(d)+\epsilon}{P(20)+\epsilon}\right) + 1.57 \cdot \log\left(\frac{d}{20}\right) \tag{1}$$

$$E(o) = -\log\left(\frac{P(o)+\epsilon}{\epsilon}\right) \tag{2}$$

The REMC folding simulation used in the Zhang-CEthreader pipeline is based on the C-I-TASSER[11] simulation plus the newly added distance restraints, torsion angle restraints, and hydrogen-bond network restraints[12] predicted from TripletRes. For each target, 10,000 decoys are generated from the REMC simulation and are clustered by SPICKER[13]. The backbone atoms of each cluster centroid will be added by REMO[14], followed by side-chain packing by FASPR[15] to produce full atomic models. The models are further refined by FG-MD[16] to remove steric clashes. Finally, the top-five models selected by ResQA, a new deep-learning based model quality estimation method trained on the TripletRes restraints, are submitted.

For multi-domain proteins, ThreaDom[17] and FUpred[18] are used to detect the domain boundaries. Both the full-length structure model and the individual domain models are predicted, and the final model is assembled from the domain models by DEMO[19] with the full-length model as the template.

1. Zheng,W., Wuyun,Q., Li,Y., Mortuza,S.M., Zhang,C., Pearce,R., Ruan,J. & Zhang,Y. (2019). Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Computational Biology* **15**, e1007411.
2. Zhang,C., Zheng,W., Mortuza,S.M., Li,Y. & Zhang,Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105-2112.
3. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173-175.

4. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. & Finn,R.D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research* **46**, W200-W204.
5. Mirdita,M., Von den driesch,L., Galiez,C., Martin,M.J., Söding,J. & Steinegger,M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170-D176.
6. Suzek,B.E., Wang,Y., Huang,H., Mcgarvey,P.B., Wu,C.H. & Uniprot,C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932.
7. Steinegger,M. & Söding,J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542.
8. Steinegger,M., Mirdita,M. & Söding,J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods* **16**, 603-606.
9. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale, V., Potter, S.C., Richardson, L.J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya,O., Lapidus,A. & Finn,R.D. (2020). MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research* **48**, D570-D578.
10. Chen,I.M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R., Seshadri,R., Smirnova,T., Kirton,E., Jungbluth,S.P., Woyke,T., Eloe-Fadrosh,E.A., Ivanova,N.N. & Kyrpides,N.C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**, D666-D677.
11. Zheng,W., Li,Y., Zhang,C., Pearce,R., Mortuza,S.M. & Zhang,Y. (2019). Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* **87**, 1149-1164.
12. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. & Zhang,Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7-8.
13. Zhang,Y. & Skolnick,J. (2004). SPICKER: a clustering approach to identify near-native protein folds. *Journal of computational chemistry* **25**, 865-871.
14. Li,Y. & Zhang,Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins: Structure, Function, and Bioinformatics* **76**, 665-676.
15. Huang,X., Pearce,R. & Zhang,Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758-3765.
16. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure* **19**, 1784-1795.
17. Xue,Z., Xu,D., Wang,Y. & Zhang,Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **29**, i247-i256.
18. Zheng,W., Zhou,X., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* **36**, 3749-3757.
19. Zhou,X., Hu,J., Zhang,C., Zhang,G. & Zhang,Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. *Proceedings of the National Academy of Sciences* **116**, 15930.

Zhang-Server

Protein 3D Structure Prediction by D-I-TASSER in CASP14

Yang Li[1], Wei Zheng[1], Chengxin Zhang[1], Eric Bell[1], Xiaoqiang Huang[1], Robin Pearce[1], Xiaogen Zhou[1], Yang Zhang[1,2]

1 - Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, 2 – Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091

yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:Y.v; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:Y; MD:Y*

The tertiary structure prediction of the Zhang-Server group in CASP14 is based on the D-I-TASSER pipeline, which is an extension of I-TASSER and C-I-TASSER that integrates deep-learning-based distance and hydrogen-bonding network models with iterative threading assembly simulations. The pipeline consists of four consecutive steps. First, starting from the query sequence, a set of multiple sequence alignments (MSAs) are created by DeepMSA[1] and its variants, by iteratively searching the query through whole-genome and metagenome sequence databases (Metaclust, BFD, Mgnify, and IMG/M), where the MSA with the highest accumulative probability obtained by the TripletRes top 10L predicted contacts[2] is selected.

In the second step, the selected MSA is used as the input for DeepPotential, a newly developed deep residual neural-network-based predictor (see DeepPotential Abstract), to create multiple spatial restraints including (1) distance-maps for both Cα and Cβ atoms; (2) Cα-based hydrogen-bonding networks[3]; (3) Cα-Cβ torsion angles. Considering that DeepPotential tends to have higher confidence for distance models with shorter distance cutoffs, four sets of distance profiles are generated with distance ranges from [2, 10], [2, 13], [2, 16], and [2, 20] Å, where the four ranges are divided into 18, 24, 30, and 38 distance bins, respectively; only the distance profiles from lower distance cutoffs are selected, i.e., distances from [2-10) Å are selected from model Set-1, distances from [10-13) Å from Set-2, [13-16) Å from Set-3, and [16-20] Å from Set-4. In addition to DeepPotential, three deep-learning and naïve Bayes classifier based contact predictors (TripletRes[2], ResPRE[4], and NeBcon[5]) are used to create Cα and Cβ contact-maps with a distance cutoff of 8 Å. Meanwhile, LOMETS3, a newly developed meta-server program containing both profile- and contact-based threading programs (see Zhang-TBM Abstract), is used to identify structural templates from a non-redundant PDB structural library. Based on the significance and consensus of the LOMETS3 alignments, the target is assigned to one of four categories (Trivial, Easy, Hard and Very-Hard)[6].

In the third step, full-length structure models are constructed using replica-exchange Monte Carlo (REMC) simulations under the guidance of a composite force field: $E = E_{I-TASSER} + E_{contact} + E_{distance} + E_{HB} + E_{torsion}$. Here, $E_{I-TASSER}$ is extended from I-TASSER[3] which contains an optimized knowledge-based energy term plus spatial restraints from LOMETS3; $E_{contact}$ is a three-gradient potential that accounts for contact-map prediction; $E_{distance}$, E_{HB} and $E_{torsion}$ are the negative logarithm of the DeepPotential-predicted probabilities for distance, hydrogen-bonding, and torsion angle maps, respectively. Three types of REMC simulations (labeled as 'A', 'M' and 'F') are run depending on a target's category, i.e., 'A' keeps all Cα atoms on-lattice with the REMC simulations starting from random conformations; 'M' freely rotates and translates fragments excised from the threading alignments; and 'F' keeps the threading-aligned fragments frozen with changes only to the unaligned regions. 'M' and 'F' are implemented only for Trivial and Easy targets whose template alignments have a higher confidence. For each pipeline, five REMC simulations are performed, where the structural decoys from 8 (or 3 for Hard and Very-Hard targets) low-temperature replicas are submitted to SPICKER[7] for structure clustering and model selection.

In the fourth step, the SPICKER clusters are refined at the atomic level using fragment-guided molecular dynamic (FG-MD) simulations[8], with the side-chain rotamer structures repacked by FASPR[9]. To select models generated from different pipelines, a set of six MQAP programs, including the D-I-TASSER C-score (Zheng et al, in preparation), the satisfaction rate of predicted contact-maps, structural consensus measured by pair-wise TM-score[10], and three statistical potentials (RW, RWplus[11], and Rotas[12]), are implemented, where a meta-MQAP consensus score is calculated as the sum of the rank of the six MQAP scores. Top-five models with the lowest consensus MQAP scores are selected for submission.

For multiple-domain sequences, FUpred[13] and ThreaDom[14] are used to predict the domain boundaries and linker regions from the contact-maps and LOMETS threading alignments, respectively. Structural models are first predicted by D-I-TASSER for the individual domains separately, which are then assembled into full-length models for the whole chain using a rigid-body domain docking and assembly algorithm, DEMO[15], guided by the whole-chain D-I-TASSER models and structural analogs identified by TM-align[16]. The procedure is fully automated.

Availability

https://zhanglab.ccmb.med.umich.edu/I-TASSER

1. Zhang,C., Zheng,W., Mortuza,S.M., Li,Y. & Zhang,Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 36, 2105-2112.
2. Li,Y., Zhang,C., Bell,E.W., Zheng,W., Zhou,X., Yu,D.J. & Zhang,Y. (2020). Deducing high-accuracy protein contact-maps from a triplet ofcoevolutionary matrices through deep residual convolutional networks. submitted.
3. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. & Zhang,Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nature Methods 12, 7-8.
4. Li,Y., Hu,J., Zhang,C., Yu,D.J. & Zhang,Y. (2019). ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 35, 4647-4655.
5. He,B., Mortuza,S.M., Wang,Y., Shen,H.B. & Zhang,Y. (2017). NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. Bioinformatics 33, 2296-2306.
6. Zhang,Y. (2014). Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 82 Suppl 2, 175-87.
7. Zhang,Y. & Skolnick,J. (2004). SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25, 865-71.
8. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19, 1784-95.

9. Huang,X., Pearce,R. & Zhang,Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. Bioinformatics 36, 3758-3765.
10. Zhang,Y. & Skolnick,J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins 57, 702-10.
11. Zhang,J. & Zhang,Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One 5, e15386.
12. Park,J. & Saitou,K. (2014). ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. BMC Bioinformatics 15, 307.
13. Zheng,W., Zhou,X., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics 36, 3749-3757.
14. Wang,Y., Wang,J., Li,R., Shi,Q., Xue,Z. & Zhang,Y. (2017). ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. Nucleic Acids Res 45, W400-W407.
15. Zhou,X., Hu,J., Zhang,C., Zhang,G. & Zhang,Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. Proc Natl Acad Sci U S A 116, 15930-15938.
16. Zhang,Y. & Skolnick,J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302-9.

Zhang-TBM

Template-based protein folding guided by residue-residue distance and hydrogen-bond network prediction from deep-learning

Wei Zheng[1], Yang Li[1, 2], Xiaogen Zhou[1], Chengxin Zhang[1], Robin Pearce[1] and Yang Zhang[1]

1 - Department of Computational Medicine and Bioinformatics, Department of Biological Chemistry, University of Michigan, 100 Washtenaw Ave, Ann Arbor, MI 481091, 2 - School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094
yangzhanglab@umich.edu

Key: *Auto:Y; CASP_serv:N; Templ:Y; MSA:Y.MetaG; Fragm:N; Cont:Y; Dist:Y; Tors:Y; DeepL:Y; EMA:N; MD:Y*

The Zhang-TBM server in CASP14 is based on I-TASSER pipeline[1] but with structural templates identified by LOMETS3 (Zheng et al, in preparation) and deep-learning based spatial restraints (contact, distance, hydrogen-bond networks and torsion angles) predicted from a new deep-learning method TripletRes (Li et al, in preparation). The pipeline was designed for template-based modeling (TBM) with the procedure fully automated.

Methods

We utilized two methods (DeepMSA[2] and qMSA) to generate 7 multiple sequence alignments (MSAs). Here, DeepMSA[2] is our previous MSA construction program developed during CASP13, which uses HHblits2[3], Jackhmmer and HMMsearch[4] to search the query sequence against the Uniclust30[5], UniRef90[6] and Metaclust[7] databases in three stages, respectively. qMSA is an extended version of DeepMSA with a new search added between stage2 and stage3 of DeepMSA, where HHblits3 is used to search the BFD[8] metagenomics database. In addition, a new iteration stage (stage4) is added in qMSA to search the query through the Mgnify metagenomics database[9]. Thus, 5 different MSAs are generated by stages 1-3 of DeepMSA and stages 3-4 of qMSA. Furthermore, the MSA from qMSA stage3 (the MSA from the BFD database) is used as the starting point for HMMsearch to search through the IMG/M[10] database that contains more sequences than the Metaclust, BFD, and Mgnify databases. The resulting sequence hits are converted into a sequence database. This sequence database is then used as the target database for DeepMSA stage3 and qMSA stage4 to generate two additional MSAs. Finally, the 7 MSAs are input to TripletRes to obtain 7 contact-maps. The MSA with the highest sum of the top 10L (L is the length of the protein) predicted contact probabilities is selected as the final MSA for threading and residue-residue contact/distance prediction.

The MSA generated from the last step is used to produce sequence profiles or profile Hidden Markov Models (HMM) for the 10 profile-based threading methods[11] used by LOMETS3, and to predict contact-maps by TripletRes that are used by 4 contact-based threading methods in LOMETS3. The four contact-based threading methods are CEthreader[12], Map_align[13], EigenThreader[14] and DisCovER[15], where the last three methods are newly added components to LOMETS3 compared to the former version, LOMETS2[11]. To speed up the contact-based threading approaches, we select the top 1000 templates identified by HHsearch[16], and then re-rank the templates by the 4 contact-based threading methods individually. For proteins that are defined as "Hard" targets by the original LOMETS3 threading methods, the predicted contacts are used to re-rank the templates identified by the profile-based threading methods using the contact-map overlap score (CMO)[12]. The final 140 templates (10 templates from each individual threading method) are used as the initial conformations for the I-TASSER REMC folding simulations. In addition, the contacts, distances, torsion angles and hydrogen-bond networks calculated from the top 20 templates are used as additional restraints to guide the REMC simulations.

In the next step, the TripletRes pipeline, which utilizes deep residual network learning, is used to predict residue-residue contact-maps, distance distributions, inter-residue torsion angles and hydrogen-bond networks based on MSA collected above. Here, the Cα-Cα and Cβ-Cβ distance distributions are predicted in the form of 38 distance bins (1 bin for <2Å, 36 bins for 2 Å to 20Å with a bin width of 0.5Å, and 1 bin for ≥20Å), torsion angles are predicted with bin widths of 15°, and the hydrogen-bond networks defined in I-TASSER folding[1] are predicted with bin widths of 10°. These sequence-based spatial restraints from TripletRes are combined with the LOMETS-based restraints and used for guiding the I-TASSER REMC folding simulations.

The I-TASSER based REMC simulations generated 10,000 decoys for each target, which are then clustered by SPICKER[17] to obtain five clusters. REMO[18] is then used to generate full backbone models from the 5 largest cluster centroids. Following this, the side-chains for each model are packed by FASPR[19] in order to get full atomic models. The models are further refined by FG-MD[20] to remove steric clashes and refine the local structure packing.

For multi-domain proteins, the domain boundaries are detected by the consensus domain boundaries predicted by ThreaDom[21] and FUpred[22]. Structure models are predicted for both the full-length and individual domain sequences. The full-length model is used as the template for DEMO[23] to assemble the individual domain models into the final model.

1. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. & Zhang,Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nature Methods 12, 7-8.
2. Zhang,C., Zheng,W., Mortuza,S.M., Li,Y. & Zhang,Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 36, 2105-2112.
3. Remmert,M., Biegert,A., Hauser,A. & Söding,J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9, 173-175.
4. Potter,S.C., Luciani,A., Eddy,S.R., Park,Y., Lopez,R. & Finn,R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Research 46, W200-W204.
5. Mirdita,M., Von den driesch,L., Galiez,C., Martin,M.J., Söding,J. & Steinegger,M. (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research 45, D170-D176.
6. Suzek,B.E., Wang,Y., Huang,H., Mcgarvey,P.B., Wu,C.H. & Uniprot,C. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics (Oxford, England) 31, 926-932.
7. Steinegger,M. & Söding,J. (2018). Clustering huge protein sequence sets in linear time. Nature Communications 9, 2542.
8. Steinegger,M., Mirdita,M. & Söding,J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nature Methods 16, 603-606.
9. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale,V., Potter,S.C., Richardson,L.J., Sakharova,E., Scheremetjew,M., Korobeynikov,A., Shlemov,A., Kunyavskaya,O., Lapidus,A. & Finn,R.D. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research 48, D570-D578.

10. Chen,I.M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R., Seshadri,R., Smirnova,T., Kirton,E., Jungbluth,S.P., Woyke,T., Eloe-Fadrosh,E.A., Ivanova,N.N. & Kyrpides,N.C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Research 47, D666-D677.
11. Zheng,W., Zhang,C., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2019). LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. Nucleic Acids Research 47, W429-W436.
12. Zheng,W., Wuyun,Q., Li,Y., Mortuza,S.M., Zhang,C., Pearce,R., Ruan,J. & Zhang,Y. (2019). Detecting distant-homology protein structures by aligning deep neural-network based contact maps. PLOS Computational Biology 15, e1007411.
13. Ovchinnikov,S., Park,H., Varghese,N., Huang,P.-S., Pavlopoulos,G.A., Kim,D.E., Kamisetty,H., Kyrpides,N.C. & Baker,D. (2017). Protein structure determination using metagenome sequence data. Science 355, 294.
14. Buchan,D.W.A. & Jones,D.T. (2017). EigenTHREADER: analogous protein fold recognition by efficient contact map threading. Bioinformatics 33, 2684-2690.
15. Bhattacharya,S., Roche,R. & Bhattacharya,D. (2020). DisCovER: distance-based covariational threading for weakly homologous proteins. bioRxiv, 2020.01.31.923409.
16. Söding,J. (2005). Protein homology detection by HMM−HMM comparison. Bioinformatics 21, 951-960.
17. Zhang,Y. & Skolnick,J. (2004). SPICKER: a clustering approach to identify near - native protein folds. Journal of computational chemistry 25, 865-871.
18. Li,Y. & Zhang,Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins: Structure, Function, and Bioinformatics 76, 665-676.
19. Huang,X., Pearce,R. & Zhang,Y. (2020). FASPR: an open-source tool for fast and accurate protein side-chain packing. Bioinformatics 36, 3758-3765.
20. Zhang,J., Liang,Y. & Zhang,Y. (2011). Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. Structure 19, 1784-1795.
21. Xue,Z., Xu,D., Wang,Y. & Zhang,Y. (2013). ThreaDom: extracting protein domain boundary information from multiple threading alignments. Bioinformatics 29, i247-i256.
22. Zheng,W., Zhou,X., Wuyun,Q., Pearce,R., Li,Y. & Zhang,Y. (2020). FUpred: detecting protein domains through deep-learning-based contact map prediction. Bioinformatics 36, 3749-3757.
23. Zhou,X., Hu,J., Zhang,C., Zhang,G. & Zhang,Y. (2019). Assembling multidomain protein structures through analogous global structural alignments. Proceedings of the National Academy of Sciences 116, 15930.