Published online 14 May 2009

Nucleic Acids Research, 2009, Vol. 37, No. 11 ***e83***
doi:10.1093/nar/gkp318

MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming

Srayanta Mukherjee and Yang Zhang*

Center for Bioinformatics and Department of Molecular Bioscience, University of Kansas, 2030 Becker Dr, Lawrence, KS 66047, USA

Received December 24, 2008; Accepted April 17, 2009

ABSTRACT

Structural comparison of multiple-chain protein complexes is essential in many studies of protein–protein interactions. We develop a new algorithm, MM-align, for sequence-independent alignment of protein complex structures. The algorithm is built on a heuristic iteration of a modified Needleman–Wunsch dynamic programming (DP) algorithm, with the alignment score specified by the inter-complex residue distances. The multiple chains in each complex are first joined, in every possible order, and then simultaneously aligned with cross-chain alignments prevented. The alignments of interface residues are enhanced by an interface-specific weighting factor. MM-align is tested on a large-scale benchmark set of 205 × 3897 non-homologous multiple-chain complex pairs. Compared with a naïve extension of the monomer alignment program of TM-align, the alignment accuracy of MM-align is significantly higher as judged by the average TM-score of the physically-aligned residues. MM-align is about two times faster than TM-align because of omitting the cross-alignment zone of the DP matrix. It also shows that the enhanced alignment of the interfaces helps in identifying biologically relevant protein complex pairs.

INTRODUCTION

Protein–protein complex structures have rapidly accumulated in various protein quaternary structure libraries (1–3). As a consequence, large-scale automated structural comparisons of multiple-chain protein complexes have become routine in most contemporary structural biology studies, ranging from structure-based functional annotation (4–6) to protein quaternary structure modeling (7,8). While extensive efforts have been focused on the development of protein 'tertiary' structure comparisons (9–11), there is no efficient structural alignment algorithm for comparing protein 'quaternary' structures.

Tertiary-structure alignment algorithms, which were developed for structurally aligning two monomer structures, cannot be directly exploited for multimeric proteins. A simple treatment might be to join the multiple chains into an artificial monomer and then align the two 'monomers' using existing programs such as Dali (9), CE (10) or TM-align (11). However, non-physical cross-chain alignments, i.e. the alignment of one chain in the first complex to several chains in the second complex, will arise because the programs do not differentiate residues of different chains. Also, if the two protein complexes include more than two chains, then a combinatorial problem arises which the available methods are not designed to handle. An alternative approach is to align the monomer chains of the two complexes separately; however, this alignment cannot account for the differences in chain orientations within the complexes. Moreover, the structure of interface regions is usually of special importance in both biological function annotation and structural modeling. Neither one of these approaches take the special characteristics of the interface structures into account. Alternatively, some sequence independent approaches like Galinter (12), I2I SiteEngine (13), MAPPIS (14) can compare protein–protein interfaces but does not help in analysing the global structural similarity of complexes.

In this article, we develop a new algorithm, MultiMeralign (or MM-align), dedicated to multimeric protein structure alignment, as an extension of the monomeric alignment program TM-align (11). TM-align, developed by Zhang and Skolnick, uses a heuristic dynamic programming (DP) alignment procedure. Because the objective function and the rotation matrix in TM-align are consistent with each other, and are both based on TM-score (15), the DP iteration converges faster than that in many other heuristic algorithms. On average, TM-align is about 20 times faster than Dali and 4 times faster than CE; and yet the alignments of monomer

*To whom correspondence should be addressed. Tel: +1 785 864 1963; Fax: 001-785-864-5558; Email: yzhang@ku.edu

structures have higher TM-scores on average (TM-score is a combined measure of the accuracy and coverage of the structure superposition, see Equation 1 below) (11) or PSI-scores (Percentage of Structural Similarity) (16). Nevertheless, for monomer alignments, there are still some cases where we found TM-align could not identify the best alignment because of the limited number of initial alignments. The purpose of this work is first to improve the efficiency of TM-align by exploring more extensive search and then to extend the algorithm to deal with the problems of unphysical cross-chain alignments and the variance of chain orientations in protein complex structures. The alignment of interface residues is also reinforced in MM-align.

MATERIALS AND METHODS

For two given protein complex structures containing n and m chains ($n \geq m$), respectively, MM-align starts by generating all possible $P(n,m) = n!/(n-m)!$ permutations for selecting m chains in the first complex. MM-align then proceeds to join the C-terminus of one protein chain with the N-terminus of another chain, in the order generated by the permutation step, and treats the combined artificial chains as rigid-body alignment units (An example of dimeric complexes shown in Figure 1).

The structural alignment procedure is subdivided into three phases: (i) Selection of chains and chain order for chain-joining; (ii) constructing initial alignments; and (iii) performing the heuristic iteration of the superposition to optimize the TM-score. In general, several alignments are initially constructed, and the inter-complex distance matrix between the superimposed structures is used to guide a heuristic iteration to refine the alignment. The chains are joined in every possible order and the alignment obtained from the order with the highest TM-score is finally returned. For the purpose of saving time in comparing big complexes of more than three chains, we first sum the TM-scores obtained from a quick alignment of individual chain pairs and then process with those combinations that have a sum of individual TM-score higher than 90% of the maximum sum of the individual TM-scores.

TM-score

The TM-score was defined as a measure to assess the structural similarity of protein monomer chains (15). Here, we extend the definition to multiple-chain protein complexes, i.e.

$$\text{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L_{ali}}\frac{1}{1+d_{ij}^2/d_0^2(L)}\right] \quad \mathbf{1}$$

where L is the total length of all chains in the target complex and L_{ali} is the number of the aligned residue pairs in the complexes. d_{ij} is the distance between the Cα atoms of the aligned residues i and j after superposition of the complexes, and $d_0(L)$ is given by $d_0 = 1.24\sqrt[3]{L-15} - 1.8$.

One major advantage of TM-score over the often-used RMSD in assessing structural alignments is that TM-score accounts for both the similarity of the aligned regions and the alignment coverage in a single parameter. Second, even when alignments with the same coverage are evaluated, TM-score is more sensitive to the global topology of the structures because it down-weights the larger distances between aligned Cα pairs compared to the smaller ones. In RMSD, all distances are taken into account with equal weights, and therefore a local error (e.g. a mis-oriented tail) will result in a big RMSD value even though the global topology of the two structures may be similar. As in Zhang and Skolnick (15), a TM-score of 1 means that the two complex structures are identical, a TM-score >0.5 indicates that two complexes have a similar topology and chain orientation, and a TM-score <0.17 indicates that the structural similarity is close to random.

Chain selection and order of chain joining

For a pair of protein complexes with multiple chains, a combinatorial problem arises if the two proteins need

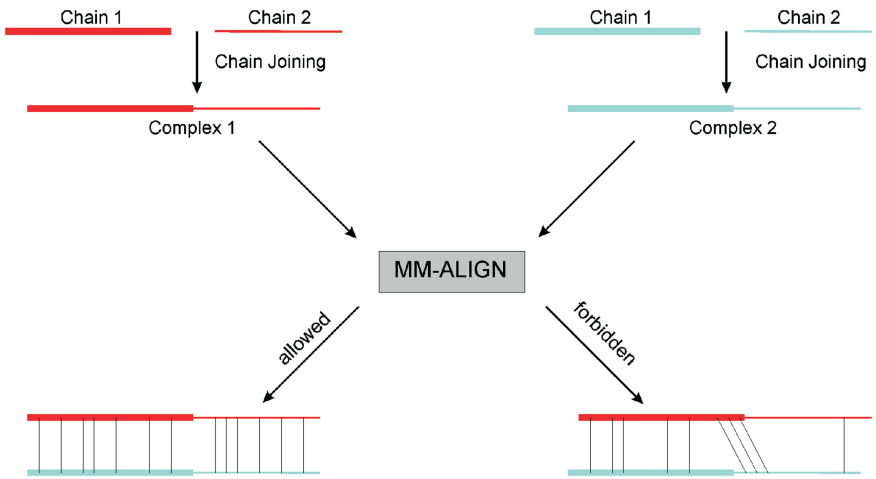

Figure 1. An illustration of the chain-joining procedure in MM-align. Both chains of the compared dimers are merged into single artificial chains and then aligned with cross-alignments forbidden. The chains corresponding to each other are presented by the same type of lines (thick and thin). Complex 1 is in red and Complex 2 is blue.

to be aligned without cross-chain alignments. Let's suppose two proteins contain n and m chains ($m < n$). Then m chains need to be selected from the n chains of the larger complex, which can be done in $C(n,m) = n!/(m!(n-m)!)$ ways. These m chains can be joined in $m!$ ways, giving rise to a total of $P(n,m) = n!/(n-m)!$ ways of comparisons. If the numbers of chains in both proteins are equal, the number of comparisons will become $n!$. When the number of chains is large, the number of possible chain orders becomes prohibitively large due to both memory and time constraints (e.g. 10 chains mean more than 3 million possible chain joinings). Therefore, to limit the number of total comparisons to a treatable range but without missing the meaningful matches, we quickly calculate the monomer TM-score for each chain in the first complex to match with the chains of the second complex based on a modified version of TM-align program, which exploits only the initial alignment from gapless threading (see below). For each chain order, we sum the TM-scores of the monomer chains that have been prescribed to be aligned. If the sum of the TM-scores of the monomer chains is >90% of the maximum sum of the monomer TM-scores obtained so far from previous steps, we then proceed further to align the complex as a whole. Otherwise, MM-align discards the particular chain order and moves on to the next order of chain joining. We find that the omission of these low-TM-score joinings does not decrease the average performance of MM-align in our testing results.

Initial alignments

MM-align uses five quickly constructed initial alignments, which are detailed as follows:

(i) An alignment of secondary structure (SS) elements using Needleman–Wunsch (NW) dynamic programming (17), using a score of 1 (0) for matching (non-matching) SS types (helix, strand or coil) of two aligned residues, and a gap penalty of –1.

(ii) Gapless alignment of the two structures (i.e. generating all possible gapless alignments by sliding one sequence along the other one with each step jumping five residues; the best alignment is selected on the basis of TM-score). Moreover, if the TM-score of any of the gapless alignment is greater than a cutoff (i.e. >95% of the maximum TM-score obtained so far), the alignment is further optimized by DP, and the alignment with the highest TM-score is selected. We find that the implementation of DP helps in generating much better starting alignments. But since only high-scoring gapless alignments are selected to do DP, this procedure does not increase the overall CPU time of the MM-align algorithm.

(iii) An alignment from DP where the score matrix is a half/half combination of the SS score matrix and the distance score matrix extracted from the second initial alignment. The gap-opening penalty is set to –1.

(iv) The fourth initial alignment is also gapless threading but the superposition of the structures is restricted to the longest continuous segments in each complex. This initial alignment is added because the second initial alignment could miss the best superposition when the joined chains have gaps (chain breaks) in the structure. This is especially the case when the algorithm is used to align interface structures that consist of chain fragments.

(v) A fragment of five continuous residues starting from the N-terminus of one protein is superimposed onto a similar fragment of five residues starting from the N-terminus of the second protein. The global TM-score is quickly calculated based on the rotation matrix of the five-residue fragments. If the TM-score is higher than 12% of the best TM-score obtained from the previous superimpositions, a DP alignment is performed to refine the initial alignment using the inter-residue distances from the initial superposition. The procedure is repeated for all five-residue fragments of either protein and the best alignment based on TM-score is finally selected. For saving CPU time, however, we skip those five-residue fragment pairs if they do not have similar secondary structure content.

Compared with TM-align, the last two initial alignments are new and the initial alignment (2) is improved by the additional DP iterations. These changes result in considerable improvement of the search engine of TM-align. In a benchmark test of aligning 4000 monomer pairs, we observed TM-score increase in 1337 cases, while the total CPU cost is kept essentially unchanged.

To prevent cross-chain alignment in the initial alignments, we have altered the conventional NW algorithm (17) so that regions in the DP matrix corresponding to cross-chain alignment are ignored. For example, if chains 1 and 2 of Complex 1 are to be aligned to chains 1 and 2 of Complex 2, respectively, the DP matrix regions corresponding to aligning chain 2 of Complex 1 with chain 1 of Complex 2 are omitted when filling up the alignment paths during DP (an example of aligning a three-chain complex pair is shown in Figure 2). The filling up of the DP matrix can be considered as a three-step process: (i) the region corresponding to the first chain (by the order prescribed by the chain joining step) of both complexes is filled up; (ii) a pseudo-layer uniformly assumes the value of the last cell of the preceding block; by doing this, the gap extension penalty will be ignored at the respective first residues of the second chains; and (iii) the region corresponding to the second chains (as per the order of chain joining) of both complexes is now filled up starting from the pseudo-layer values (instead of 0, which is used as the initial value for the first block). The process is repeated when aligning complexes with more than two chains.

While tracing back the pathway, we follow the reverse order and start the traceback in the region corresponding to the last chain of both complexes, crossing the junction of the diagonal blocks, and then continue the traceback in the area corresponding to the 'next to last' chains of both complexes. Traceback continues until we reach the first residue of the first chain of both complexes. We thus avoid the cross-alignment zones completely, and force the

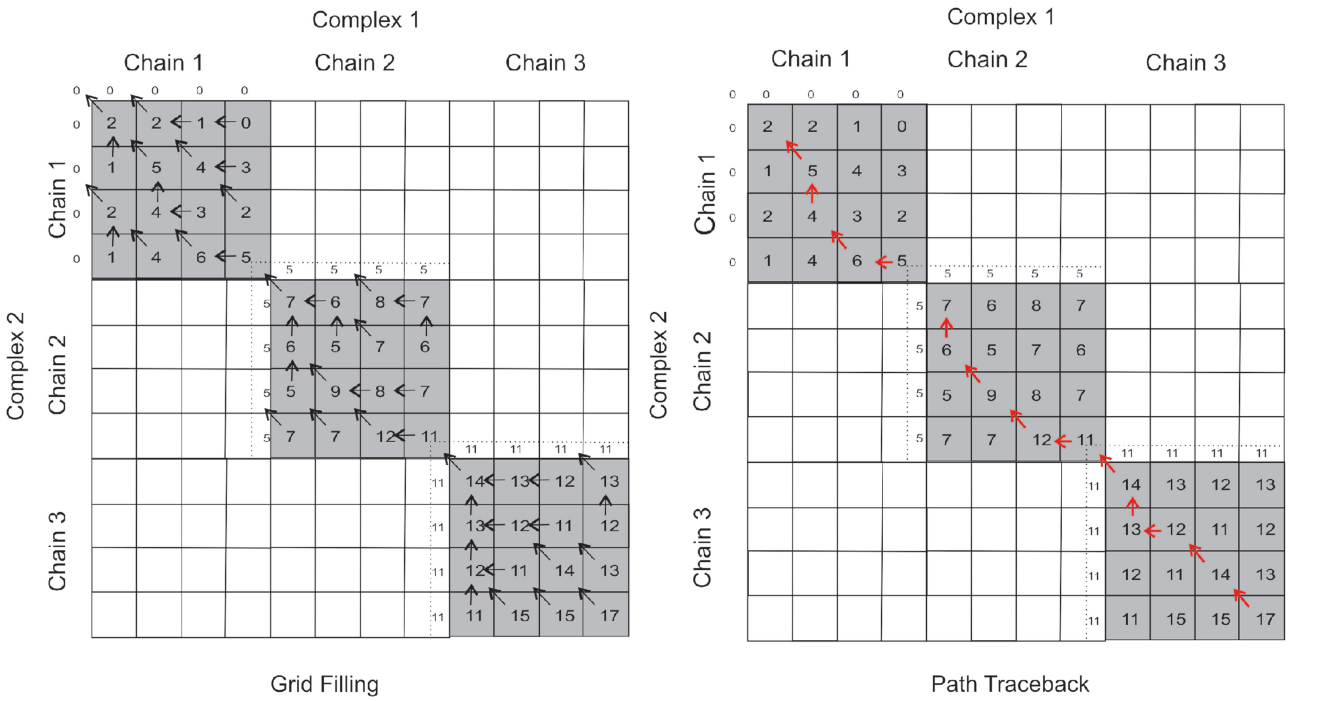

Figure 2. An illustration of the modified dynamic programming algorithm with cross-chain alignment prevented. The picture on the left panel illustrates the process of filling up the grid, with the cross-alignment zones (empty grids) ignored. The dashed lines represent a pseudo-layer which assumes the value in the last cell of the preceding block. The values of the pseudo-layer (5 and 11 in this example) are used as starting score of the next block corresponding to the next chain of both complexes. The picture on the right panel shows the traceback path (indicated by red arrows).

alignment to traverse a path, which does not lead to alignment of any residue of chains not prescribed to be aligned for that particular iteration. An illustration of the modified DP for a trimer is presented in Figure 2. An alternative treatment would be to employ a large penalty for cross-aligned regions, which is, however, more CPU-expensive because of the filling and backtracing procedures in the forbidden areas.

The five initial alignments thus derived are passed on to the heuristic iteration phase for further refinement.

Heuristic iterations

Once an alignment is obtained, the structures of the two complexes can be spatially superimposed by the TM-score subroutine (15). Based on the superimposed structures, a similarity scoring matrix is defined as

$$S_{ij} = \begin{cases} \frac{1}{1+d_{ij}^2/d_0^2(L_{\min})}, & \text{if } i \text{ and } j \text{ are aligned without cross} \\ \text{ignored}, & \text{if } i \text{ and } j \text{ are aligned with cross} \end{cases} \quad \mathbf{2}$$

where d_{ij} is the same as that defined in Equation (1). $d_0 = 1.24\sqrt[3]{L_{\min} - 15} - 1.8$, and $L_{\min}$ is the total length of the smaller complex. The purpose of using $L_{\min}$ instead of the target length (L) here is to avoid the asymmetry resulting when aligning Complex 1 to Complex 2 versus Complex 2 to Complex 1. Like in Figure 2, we omit the residue pair when i and j are from cross-aligned chains.

A new alignment can be generated based on the score matrix of Equation (2) by the modified NW dynamic programming as explained in Figure 2, with an optimal gap-opening penalty of –0.6. Based on the new alignment, we superimpose the complex structures by the TM-score subroutine again, which will give rise to a new similarity scoring matrix. This can again be used for the modified NW dynamic programming. The procedure is repeated for a number of times until the alignment between two protein complexes becomes stable. The alignment with the maximum TM-score encountered during the iterations starting from the five initial alignments is returned as the final alignment.

Because the score matrix of Equation (2) is consistent with the target function of TM-score of Equation (1), the iteration converges very fast, and usually two to three iterations are enough to find the best alignment. As we are mainly interested in the topological match between the compared complexes, no gap extension penalty is applied.

Preferential alignment on interfaces

The structures of protein–protein interfaces are usually more conserved than other regions, and generally have special importance in the inference of biological function (18). In the MM-align program, we have a special option for preferentially aligning the interface residues of dimers, which constitutes the largest subgroup of multimeric protein complexes.

For the given dimer structures, the interface residues are defined using a default Cα distance cutoff of 8 Å (a different value can optionally be specified by the user), i.e. any

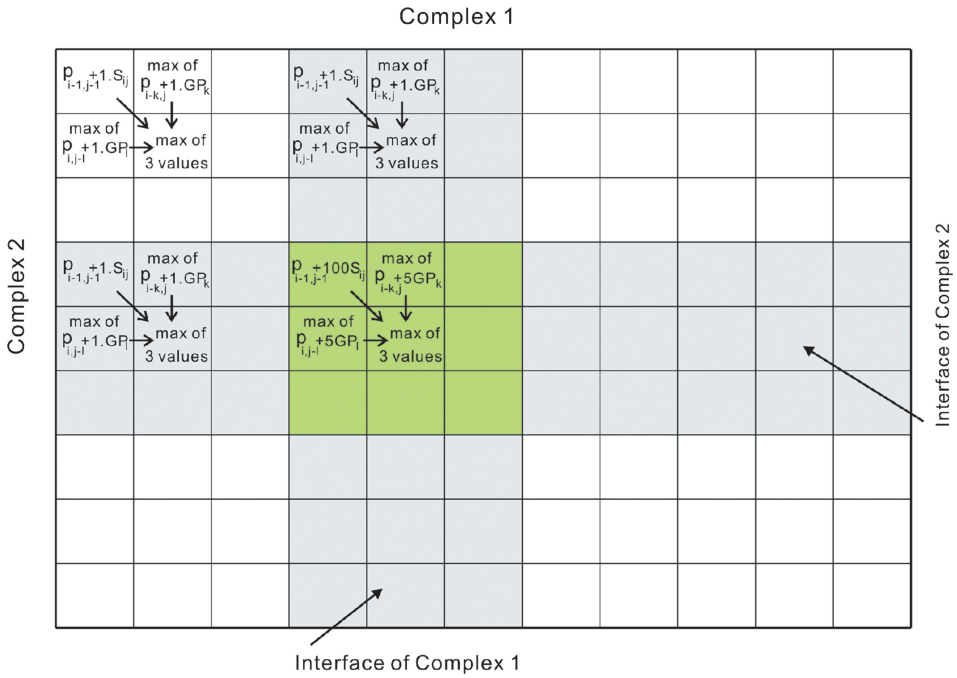

Figure 3. A modified dynamic programming scheme with the alignment of interface residue pairs reinforced. The interface areas are highlighted in color. If the residue pairs are both from an interface (the area in green), the score is increased by a factor w and the gap penalty is increased by a factor x.

residue whose Cα atom is at a distance <8 Å from any Cα atom in the other chain of the complex is considered to be an interface residue. The alignment of the interface residues can be enhanced by a modified dynamic programming scheme where the alignment path is defined by

$$P(i,j) = \max\{p(i-1,j-1) + wS_{ij}, \max_{k\geq 1}\{p(i-k,j) + xGP_k\}, \max_{l\geq 1}\{p(i,j-1) + xGP_l\}\} \quad (3)$$

where $P(i, j)$ is the maximum score of an alignment path ending at (i, j) and $GP_k < 0$ is the normal gap penalty. Because we have no gap extension penalty, GP_k actually does not depend on k. For non-interface residue pairs, $w = x = 1$. If both i and j are from interfaces, $w > 1$ is used to encourage the alignment of the interfaces and $x > 1$ to discourage gaps at the interfaces (see Figure 3). The gap penalty is always neglected at the boundary of two chains.

Since interface alignment is most important for complex pairs with weak structural similarity (see below), we optimized the parameters of w and x based on 2000 complex pairs with TM-scores <0.3 and interface alignment coverage below 10%. In general, higher values of w and x will increase the number of aligned interface residues but too large values will reduce the TM-score of the overall alignment. After a comprehensive grid search of the parameter space, we found that $w = 100$ and $x = 5$ work the best for generating the highest number of aligned interface residue pairs while still maintaining a reasonable TM-score of global alignments.

RESULTS

Benchmark sets

Dimers constitute by far the largest subgroup of multimeric protein complexes and, therefore, it is on dimers that MM-align has been mostly tested. However, MM-align also has the capability of accurately aligning larger multimers which is tested on a number of the higher-order multimer cases. For testing MM-align on dimeric complexes, we constructed two sets of protein complex structures. The first set consists of 205 non-redundant dimers with various sizes and a pair-wise sequence identity of <30%. The second set consists of 3897 dimers collected from Dockground [2], with a pair-wise sequence identity <70%. The pair-wise sequence identity between the first and the second complex sets is <98%. A complete list of the two benchmark sets is available at http://zhang.bioinformatics.ku.edu/MM-align/benchmark.

Prevention of cross-chain alignment

We first check the ability of MM-align to exclude the unphysical cross-chain alignments. Using MM-align, we align all dimer structures in the first benchmark set against all dimers in the second set. For each of the 205 complexes in the first set, we select the complex from the 3897 complexes in the second set that has the best match based on TM-score. A summary of the results on the 205 pairs is presented in Table 1. For most dimers in the first set, MM-align identified similar dimer structures in the second set. As shown in Figure 4, 83% of protein complex pairs have a TM-score >0.5, indicating similar topology of the

two complexes (15). The average sequence identity between these best complex pairs is 44%. For the protein pairs having a sequence identity <30%, the average TM-score is 0.59, indicating that MM-align can identify structures with similar topology even when the sequence identity is very low. A complete list of the alignments for the 205 best complex pairs is available in Supplementary Table S1.

As a comparison, we also ran TM-align on the complexes, directly aligning them with chains joined and treating them as 'artificial monomers' (the results are shown as TM-align-I in Table 1). As expected, because TM-align does not distinguish between the different chains, a substantial portion of residue pairs gets non-physically aligned across chains. In the MM-align alignment, however, due to the exclusion of the cross-chain alignment paths in the DP matrix, there is no cross-chain alignment in any of the 205 × 3897 alignments. One example is presented in Figure 5, where both chains of the C. AhdI protein complex (PDB ID: 1y7y) are aligned by TM-align on the A chain of the *Xenopus laevis* nudix hydrolase nuclear SnoRNA decapping protein (PDB ID: 1u20). But when MM-align is used on the same structure pair, there is no cross-chain alignment and the interfaces of the two complexes are correctly aligned.

Table 1. Summary of results from TM-align (32) and MM-align on complex structure alignments

Method	⟨TM-score⟩	⟨RMSD⟩ Å	⟨Cov⟩[a] (%)	⟨N_{cross}⟩[b]
MM-align	0.759	2.65	60.4	0
TM-align-I[c]	0.750	2.70	60.2	12.6
TM-align-II[d]	0.710	3.00	58.5	0

[a]Average fraction of the aligned residue pairs divided by the length of the target complex.
[b]Average number of the non-physical cross-chain alignments.
[c]Using TM-align to align joined-chain complex structures.
[d]Same as TM-align-I but removing the cross-chain alignments.

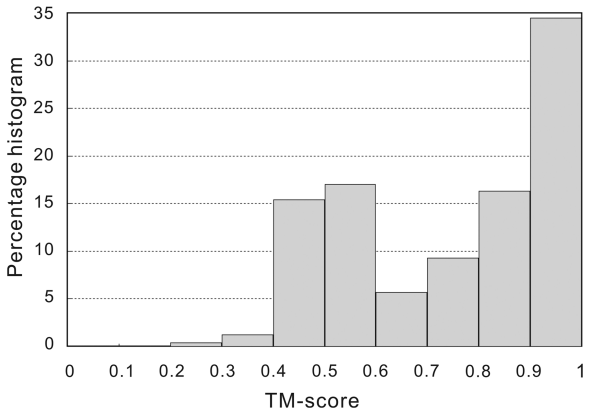

Figure 4. TM-score histogram of 205 protein complexes and their best-matching structures identified by MM-align in a non-redundant set of 3897 protein complexes.

Remarkably, despite the fact that MM-align searches far fewer possible alignment paths than TM-align (i.e. neglecting all the paths of cross alignments, see Figure 2), the average TM-score and RMSD of the best alignments by MM-align are better than those produced by TM-align-I. This improvement is mainly attributed to the newly added initial alignments in MM-align and the improved DP search in the existing paths of TM-align. On the other hand, the fact that much fewer paths catch up structure matches of similar or even better TM-scores reflects that the protein quaternary structures have inherent structural similarities of separate domain/chains. If we remove the cross-chain parts of the alignment from TM-align (shown as 'TM-align-II' in Table 1), the alignment score and coverage are much lower than that of MM-align, i.e. TM-score/RMSD/coverage by TM-align-II are 0.71/3.0Å/58.5% versus 0.759/2.65Å/60.4% for MM-align (Table 1).

Although no-chain-crossing rule is requested in most multimeric complex structure comparisons, there are also occasions where it may not be the case, e.g. aligning protein complexes, which involve domain swapping (19). For dealing with this issue, MM-align has a special option that allows cross-chain alignment between chains, when users suspect that domain swapping may be involved (or for any other reason where cross-chain alignment prevention is not required). There are also cases where no one-to-one correspondence is specified between subunits [e.g. gene-fusions (20) or aligning proteolytically cleaved chains to an uncleaved chain]. We therefore set up another special option of MM-align for aligning one chain to multiple chains. Similarly, the no-chain-crossing rule is taken off by using the normal DP for alignment instead of the modified DP illustrated in Figure 2.

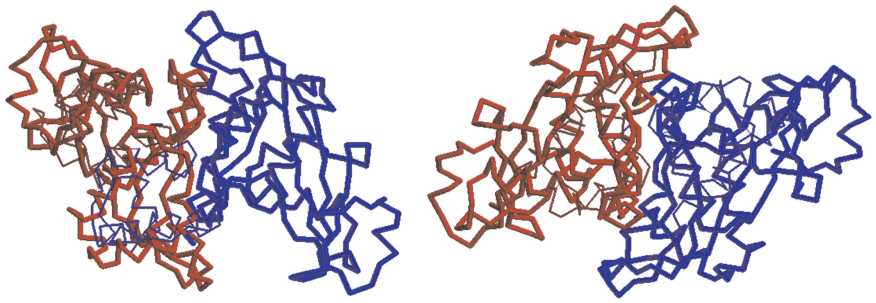

Figure 5. A typical example structures aligned by TM-align, containing cross-chain alignments (left panel), and the same structures aligned without cross-chain alignment by MM-align (right panel). The two complexes are from PDB files 1u20 (thick trace) and 1y7y (thin trace), with the two chains represented in blue and red, respectively.

Option for interface-enhanced alignments

Interface residues are usually related to biological activity, and evolutionarily more conserved than other regions of the protein (18,21–23). Matching subunit interfaces is of special importance when complex structures are compared. For protein complexes with obvious structural similarity and consistent interfaces, the normal version of MM-align can align both global structures and interfaces correctly. But when structural similarity is weak, the procedure may place the interfaces arbitrarily along the alignment path. For users interested only in aligning the interfaces of such complexes, MM-align provides an option to optimize the interface match in addition to optimizing the TM-score.

To reinforce the alignment of the interfaces, MM-align assigns a higher weight to the alignment scores and a higher gap penalty if the alignment involves the interface residues as described in Equation (3) and Figure 3. For testing this option, we randomly selected 2000 complex pairs from the 205×3897 pairs which have a TM-score <0.4 and an interface coverage $<10\%$ by the normal MM-align alignment. This set of protein complexes is different from the training protein pairs used to train the parameters as described in 'Materials and methods' section. The average fraction of aligned interface residues versus all interface residues is 3.3% in the normal MM-align alignments. After applying the interface-enhancement option, the average fraction of aligned interface residues increases to 14.3%, but the overall TM-score is similar to that without using the interface option (though the global structural match in this TM-score range is not very meaningful).

Functional relevance of structure alignments

The biological function of protein complexes depends on their 3D structures (6,24). An important goal of protein structural alignment algorithms is to assist in identifying the function-related structural similarities between complexes.

Out of the 205 non-redundant protein complex pairs identified by MM-align, 153 (75%) pairs have related functions as judged by the annotations in the original PDB files and Gene Ontology (GO) (25) annotations. The function of the complexes has been manually assessed by the following procedure: If the 'molecular function' GO term of the query and template complexes were the same, they were considered to have the same function. In a few cases, no 'molecular function' was associated with a complex; we then looked at the 'biological process' GO term. If the 'biological process' term was also missing, which occurs quite rarely, we further referred to the 'Classification' record in the PDB file. The function of all the 205 complexes could be obtained by this procedure.

Among the 135 protein complex pairs having a TM-score >0.7, 133 (98.5%) have the same function. Twenty-one percent of these protein pairs have a sequence identity below 30%. Out of the two complex pairs with different functions, one has a TM-score of 0.959 because both complexes are coiled-coils with very little deviation in structure. The sequence identity of this pair is very low (10%). In the other case, the TM-score is 0.709, but the compared structures are only fragments of their respective proteins rather than the complete complex structures. A complete list of TM-scores, RMSDs, and functional assignments of all 205 complex pairs is presented in Supplementary Table S1. These data demonstrate the ability of MM-align to identify structural similarities related to biological function.

In Figure 6, we present three illustrative examples of protein pairs from different protein classes (alpha-, beta- and alpha/beta-proteins); each having a high structural similarity but low sequence identity. The first target complex is from the protein allophycocyanin (PDB ID: 1all), a light harvesting protein (26) found in the cyanobacterium *Spirulina platensis*. Both its chains belong to the 'mainly alpha' class in CATH (27), and have an orthogonal bundle architecture and a globin-like topology. The complex selected by MM-align based on TM-score is alpha-phycoerythrocyanin (PDB ID: 2j96), which is also involved in photosynthesis (28) in the thermophilic alga *Mastigocladus laminosus*. According to CATH and SCOP, its both chains have the same architecture and topology as allophycocyanin (29). The sequence identity of the complex pair is 27% and the TM-score from MM-align is 0.895 (Figure 6a).

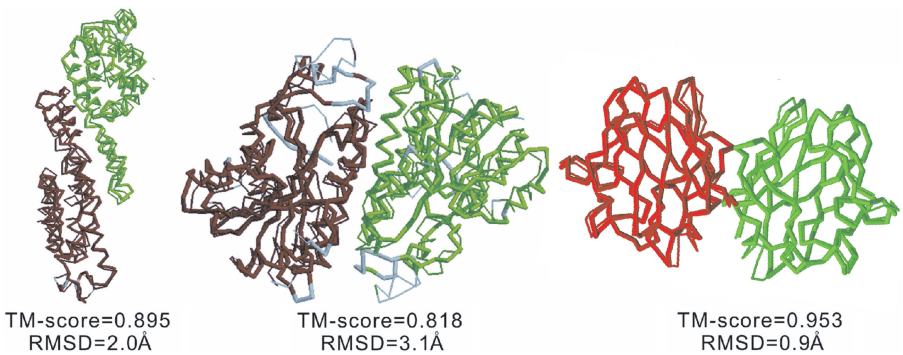

Figure 6. Three examples of protein dimeric complex alignments identified by MM-align, from three different protein classes (alpha-, alpha/beta- and beta-proteins). Thick and thin lines represent the Cα traces of different complexes, and red and green indicate different chains. The grey regions are those with a distance >5 Å in the superposition.

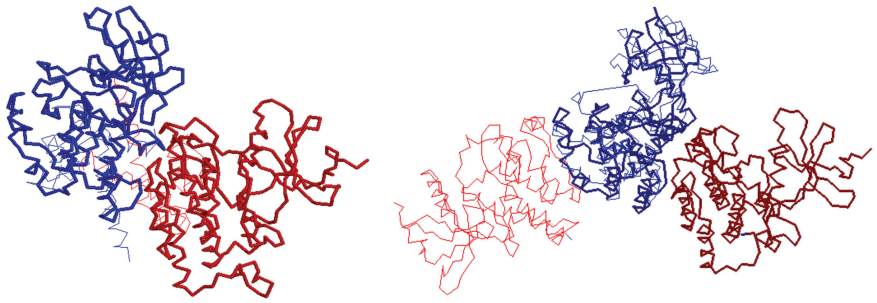

Figure 7. The structural alignment of casein kinase (1cki) with its best-matching structures in a non-redundant protein complex library. TM-align picks up human S100P (1j55) with 26 residues aligned across chains (left panel); MM-align picks up the tyrosine kinase domain of fibroblast growth factor (1fgk), without cross-aligned residues.

In the second example, the protein alcohol dehydrogenase from *Drosophila lebanonensis* (PDB ID: 1a4u) has an oxidoreductase activity (30) and its both chains are classified by CATH as alpha-beta proteins having a Rossman fold. The structurally closest complex chosen by MM-align is sorbitol dehydrogenase (PDB ID: 1k2w) from the bacterium *Rhodobacter sphaeroides*, which has the same activity (31) and belongs to the same class and fold according to CATH. The sequence identity of the complex pair is 24% and the TM-score from MM-align is 0.818 (Figure 6b).

The complexes in the third example are two 'mainly beta' proteins as classified by SCOP and CATH. The query protein is human copper superoxide dismutase (PDB ID: 1do5), and the best match found by MM-align is copper-zinc superoxide dismutase from *Xenopus laevis* (PDB ID: 1xso). The two proteins share a low sequence identity around 50% but have extremely similar structures with a TM-score of 0.953 (Figure 6c). Both have a similar topology and architecture of an immunoglobin-like sandwich according to CATH.

When the structural similarity is very high, functionally related protein pairs may also be identified by the naïve application of TM-align. However, cross-chain alignments may occur, and may lead to incorrect assignment of the protein family. One such example is casein kinase from *Rattus norvegicus* (PDB ID: 1cki). The closest complex identified by TM-align is the calcium-binding protein S100P (PDB ID: 1j55) (see Figure 7a). When we search Set 2 by MM-align, the closest protein complex found is a tyrosine kinase from human (PDB ID: 1fgk). In this example, the aligned complex structures derived from the naïve version of TM-align have a higher TM-score (0.409) than that from MM-align (0.396), but with 26 residue pairs aligned to the wrong chain, which results in an incorrect function assignment. By preventing the cross-chain alignment, MM-align aligns the complex structure correctly and assigns a similar function to it by the structure comparison. Only one chain is aligned by MM-align because of the different chain orientations.

Alignment of large oligomers

One of the important purposes of MM-align is to align proteins from large oligomers. Because the number of solved higher-order complexes in the PDB is much smaller than that of dimers, in Figure 8 we show four examples of MM-align with structures randomly selected from four families of big complexes, which include two of unequal number of chains and two of equal number of chains. The size of the complexes varies from 3 to 20 chains.

Figure 8a is an alignment of the photosynthetic reaction center of *Rhodobacter sphaeroides* (PDB ID: 2jiy) with that of *Rhodopseudomonas viridis* (PDB ID: 1dxr), which are randomly selected from the same family of bacteria. 2jiy has three subunits while 1dxr has four (the cytochrome *c* subunit is extra). Their alignment by MM-align yields a TM-score of 0.669 with the three chains of 2jiy being aligned to the second, third and fourth subunit of 1dxr, respectively. The first chain of 1dxr, which is cytochrome *c*, remains unaligned.

Figure 8b is another example of big complexes with unequal number of chains. The cytochrome *bc*1 complex from chicken (PDB id: 1bcc) has 10 chains while the bovine mitochondrial cytochrome *bc*1 complex (PDB id: 1qcr) has 11 chains. The automated MM-align procedure identified the correct chain combination and generated a structural match of TM-score = 0.907 and RMSD = 2.7 Å.

Figure 8b is an example of complexes with equal chain numbers, which come from phycocyanins in the *Gleobacter violaceus* (PDB id: 2vml) and the red algae *Gracilaria chilensis* (PDB id: 2bv8). Both complexes include 12 protein chains. MM-align correctly selects the chain combination and generates an alignment of TM-score = 0.657 and RMSD = 2.13 Å.

Figure 8d is an alignment of complexes of maximum size by MM-align. The structures come from the bacterial ribosome in *E. coli* (PDB id: 2qbd) and the ribosome of the bacterial species *Thermus thermophilus* (PDB id: 1fjg); both have 20 protein chains. MM-align generate a structure match of TM-score = 0.517 and RMSD = 4.16 Å. Owing to the large number of possible chain combinations, it takes MM-align nearly 1 h at a 2.6 GHz AMD processor to generate the best alignment in this example.

DISCUSSION

We have developed a new algorithm, MM-align, for quickly aligning and comparing the structures of multiple-chain protein complexes. Bearing in mind the

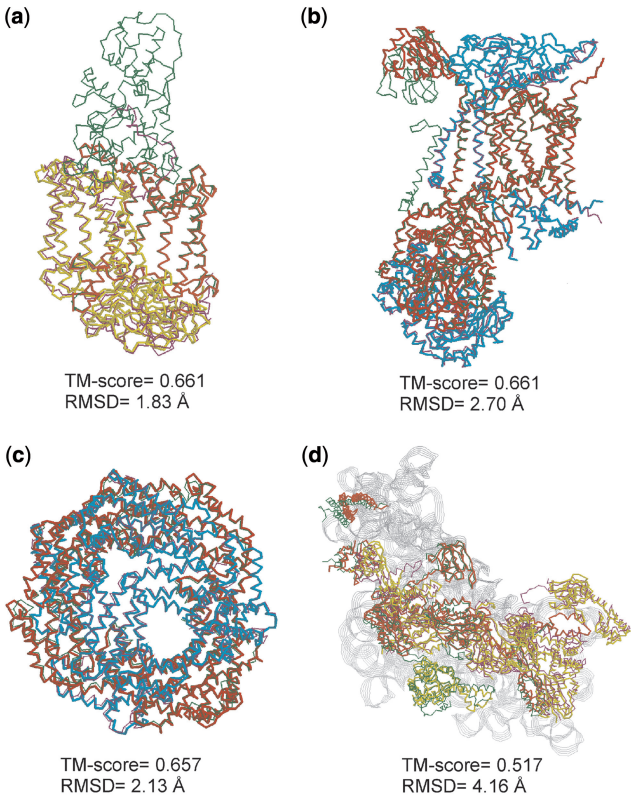

Figure 8. Examples of MM-align on big oligomers. (**a**) Alignment of the photosynthetic reaction center from *Rhodobacter sphaeroides* (PDB id: 2jiy, three chains, thick backbone) with that from *Rhodopseudomonas viridis* (PDB id: 1dxr, 4 chains, thin backbone). Yellow, cyan and yellow are for the first, second and third chains of 2jiy; dark green, magenta, dark green and magents are for the first, second, third and fourth chains of 1dxr. (**b**) Alignment of cytochrome *bc*1 complex from chicken (PDB id: 1bcc, 10 chains, thick backbone) with bovine mitochondrial cytochrome *bc*1 complex (PDB id: 1qcr, 11 chains, thin backbone). The chains are colored red and cyan alternatively for 1bcc and green and magenta for 1qcr. (**c**) Alignment of phycocyanin from the *Gleobacter violaceus* (PDB id: 2vml, 12 chains, thick backbone) with phycocyanin from the red algae *Gracilaria chilensis* (PDB id: 2bv8, 12 chains, thin backbone). The chains are colored in red and cyan alternatively for 2vml and green and magenta for 2bv8. (**d**) Alignment of bacterial ribosome from *E. coli* (PDB id: 2qbd, 20 chains, thick backbone) with ribosome of the bacterial species *Thermus thermophilus* (PDB id: 1fjg, 20 chains, thin backbone). The chains are colored red and yellow alternatively for 2qbd and green and magenta for 1fjg. The grey strands in background are RNA from 2qbd superimposed onto the aligned complexes.

importance of protein–protein interactions in structural biology studies, and the lack of computer algorithms dedicated to multimeric structure alignments, the MM-align method is expected to be of important use in many aspects of the field. The algorithm performs simultaneous alignment of all chains of protein complexes with both the monomer similarity and the relative chain-orientations accounted for by a single TM-score. The biologically irrelevant cross-chain alignments are efficiently prevented by the implementation of a modified DP algorithm which ignores the cross-alignment blocks of the DP matrix while filling up the cells and tracing back the pathway. This results in halving the necessary CPU time. Because of the consistency of the rotation matrix and the objective function, the convergence of the heuristic iteration stage is fast. For aligning a pair of protein dimers of 400 residues each, the average CPU cost is 0.35 s on a 2.6-GHz AMD processor.

The algorithm also includes a user-specified option to reinforce the structural alignment in the interface regions. The default weight for aligned interface residues has been carefully optimized using a benchmark set, balancing the overall topology match and the accuracy of interface alignment. Higher weights would result in aligning a higher number of interface residues but would, on average, deteriorate the overall structure match. This option is especially useful when the global structural match is inconsistent with the interface similarities but the user is interested in the interface match. In cases where there is reason to believe that prevention of cross-chain alignment is not desirable (e.g. complexes involving domain swapping or gene fusion), MM-align has a special option to utilize normal DP and hence does not prevent cross-chain alignment. It also allows alignment between one chain to multiple chains.

Noting the fact that proteins often function as complexes, a functional annotation study based on the conserved complex structures is relevant. In a test on 205 non-homologous proteins, MM-align was able to detect functionally similar proteins within a non-complete benchmark dataset of 3897 complexes. It often prevents false positives that may arise when dimer structures are aligned with tools dedicated to single-chain alignments only, like TM-align. MM-align also has the capability of aligning large multimeric complexes up to 20 chains and correctly identifying the corresponding subunits and the structure match. These data show that MM-align may serve as an effective function annotation tool if used for querying a complete library such as all complexes in the PDB. Because MM-align provides a single TM-score describing the global similarity of the complexes, it can also be conveniently used for automated and quantitative classification of protein complex structures. The related work is in progress.

An online MM-align server and the source code of the program are freely available at: http://zhang.bioinformatics.ku.edu/MM-align.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Alfred P. Sloan Foundation; NSF Career Award (DBI 0746198); and the National Institute of General Medical Sciences (R01GM083107, R01GM084222). Funding for open access charge: Alfred P. Sloan Research Fellowship.

Conflict of interest statement. None declared

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Douguet,D., Chen,H.C., Tovchigrechko,A. and Vakser,I.A. (2006) DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, **22**, 2612–2618.
3. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
4. Arakaki,A.K., Zhang,Y. and Skolnick,J. (2004) Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics*, **20**, 1087–1096.
5. Graille,M., Baltaze,J.P., Leulliot,N., Liger,D., Quevillon-Cheruel,S. and van Tilbeurgh,H. (2006) Structure-based functional annotation: yeast ymr099c codes for a D-hexose-6-phosphate mutarotase. *J. Biol. Chem.*, **281**, 30175–30185.
6. Zhang,Y. (2009) Protein structure prediction: When is it useful? *Curr. Opin. Struct. Biol.*, **19**, 145–155.
7. Janin,J., Henrick,K., Moult,J., Eyck,L.T., Sternberg,M.J., Vajda,S., Vakser,I. and Wodak,S.J. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, **52**, 2–9.
8. Vajda,S. and Camacho,C.J. (2004) Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol.*, **22**, 110–116.
9. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
10. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
11. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
12. Zhu,H., Sommer,I., Lengauer,T. and Domingues,F.S. (2008) Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, **3**, e1926.
13. Mintz,S., Shulman-Peleg,A., Wolfson,H.J. and Nussinov,R. (2005) Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins*, **61**, 6–20.
14. Shulman-Peleg,A., Shatsky,M., Nussinov,R. and Wolfson,H.J. (2008) MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic Acids Res.*, **36**, W260–264.
15. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
16. Teichert,F., Bastolla,U. and Porto,M. (2007) SABERTOOTH: protein structural alignment based on a vectorial structure representation. *B.M.C. Bioinformatics*, **8**, 425.
17. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
18. Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.
19. Bennett,M.J., Choe,S. and Eisenberg,D. (1994) Domain swapping: entangling alliances between proteins. *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.
20. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
21. Pawson,T. and Nash,P. (2000) Protein-protein interactions define specificity in signal transduction. *Genes Dev.*, **14**, 1027–1047.
22. Phizicky,E.M. and Fields,S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.*, **59**, 94–123.
23. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.
24. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Brejc,K., Ficner,R., Huber,R. and Steinbacher,S. (1995) Isolation, crystallization, crystal structure analysis and refinement of allophycocyanin from the cyanobacterium Spirulina platensis at 2.3 A resolution. *J. Mol. Biol.*, **249**, 424–440.
27. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
28. Schmidt,M., Patel,A., Zhao,Y. and Reuter,W. (2007) Structural basis for the photochemistry of alpha-phycoerythrocyanin. *Biochemistry*, **46**, 416–423.
29. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
30. Benach,J., Atrian,S., Gonzalez-Duarte,R. and Ladenstein,R. (1998) The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 A resolution. *J. Mol. Biol.*, **282**, 383–399.
31. Philippsen,A., Schirmer,T., Stein,M.A., Giffhorn,F. and Stetefeld,J. (2005) Structure of zinc-independent sorbitol dehydrogenase from Rhodobacter sphaeroides at 2.4 A resolution. *Acta Crystallogr. D. Biol. Crystallogr.*, **61**, 374–379.
32. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.