Worldwide Protein Data Bank Advisory Committee Meeting

October 12, 2012

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Nobelprize.org
The Official Web Site of the Nobel Prize

MEDICINE PRIZE
Shinya Yamanaka
Sir John B. Gurdon

PHYSICS PRIZE
Serge Haroche
David J. Wineland

CHEMISTRY PRIZE
Robert J. Lefkowitz
Brian K. Kobilka

11 oct 1 p.m. CET LITERATURE PRIZE

12 oct 1 p.m. CET PEACE PRIZE

15 oct 1 p.m. CET ECONOMIC SCIENCES PRIZE

Home | A-Z Index | FAQ | Press | Contact Us

Nobel Prizes | Alfred Nobel | Educational | Video Player | Nobel Organizations

Search

Nature, August 2011

FEATURE NEWS

It's all about the structure

For more than 20 years, Brian Kobilka worked to create a portrait of a key cell receptor. Sometimes, the slow, steady approach wins.

BY LIZZIE BUCHEN

Agenda

8:30	Welcome	
8:45	Overview	Haruki Nakamura
9:45	D&A Tool	Martha Quesada
10:30	*Break*	
11:00	Format, Raw Data, Validation	Gerard Kleywegt
12:00	*Lunch*	
1:30	NMR	John Markley
2:00	EM, SAS, Remediation	Helen Berman
3:00	wwPDB Organizational Update	Gerard Kleywegt
	Matters Arising, Discussion	
4:00	*Break*	
4:30	Executive Session	
5:30	Adjourn	

Overview

Haruki Nakamura

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

wwPDB

September 2011- October 2012

- Continued growth of archive
- Increased use of data
- Progress in Common Tool project
- Format working group
- Planning of next archive remediations
- Task Force activities
- Outreach: PDB40; 2012 Symposium
- PI editorials in *Structure* and *Biopolymers*
- Funding
- wwPDB Foundation

257 participants
35 travel awards
93 posters

PDB40 speakers

PDB past and present

PDB Depositions

By deposition and processing site

Year	Total Depositions	Deposited To			Processed By		
		RCSB PDB	PDBj	PDBe	RCSB PDB	PDBj	PDBe
2000	2983	2445	10	528	2297	158	528
2001	3287	2673	118	496	2408	383	496
2002	3565	2769	289	507	2401	657	507
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3082	1614	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	8130	5399	1603	1128	4703	2299	1128
2008	7073	5452	648	973	4106	1994	973
2009	8300	6715	527	1058	5069	2173	1058
2010	8878	6912	593	1373	5464	2041	1373
2011	9250	7172	582	1496	5938	1816	1496
2012	7514 *	5891	408	1215	5001	1297	1215
TOTAL	83278	62364	8569	12345	51419	19513	12345

Last Updated: 3 Oct 2012

***9768 depositions projected for 2012**

2011 FTP & Rsync Entry Downloads

RCSB PDB
2011: 282 million
2010: 159 million

PDBe
2011: 59 million
2010: 34 million

PDBj
2011: 38 million
2010: 16 million

Common Tool for Deposition and Annotation

- Sequence-annotation module v1.0 completed
- Ligand-annotation module v1.0 including new features for oligomers completed
- Workflow engine and management system running with annotation modules
- Validation module on track
- Deposition system in active development
- Cross-site data-sharing architecture in place

Format Discussions

- PDBx addresses limitations in molecular size and complexity and extensibility of existing PDB format
- Software developers committed to the production of PDBx/mmCIF for deposition by early 2013

Task Forces

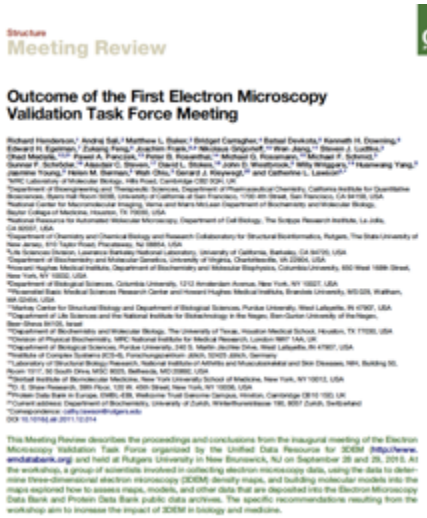
Collect recommendations and develop consensus on method-specific issues, including validation checks that should be performed and identification of validation software applications.

X-ray Validation

- 2008 Workshop
- 2011 *Structure* publication
- Chair: Randy J. Read (University of Cambridge)

3DEM Validation

- 2010 Meeting
- Chairs: Richard Henderson (maps, MRC-LMB), Andrej Sali (models, UCSF)
- 2012 *Structure* publication


Structure
Meeting Review

Outcome of the First Electron Microscopy Validation Task Force Meeting

NMR Validation

- 2009, 2011 Meetings
- Chairs: Gaetano Montelione (Rutgers), Michael Nilges (Institut Pasteur)
- Report in progress

Small-Angle Scattering

- 2012 Meeting
- Chair: Jill Trewhella (University of Sydney)
- Report in progress

Funding

- RCSB PDB competitive renewal funded by NSF
  - January 2009 - December 2013
  - Noncompetitive 5 years renewal due in 2013
- PDBe main funding from EMBL and Wellcome Trust
  - WT: competitive grant (2010-2014)
  - EMBL: core of ~15 posts
- PDBj competitive renewal funded by JST (Japan Science & Technology Agency)
  - April 2011 - March 2014
- BMRB competitive renewal funded from the National Library of Medicine
  - NLM will no longer fund BMRB after 2014

Worldwide
Protein Data Bank
Foundation

- Established to support specific wwPDB activities
  - Advisory committee meetings
  - Outreach and education activities, including seminars and workshops
- 501(c)3 organization
  - American, tax-exempt association dedicated to scientific, literary, charitable, and educational purposes
- Fundraising on-going


13:30－14:00
中村春木（大阪大学・蛋白質研究所）
「生命科学における蛋白質構造データバンク（PDB）の役割」
14:00－15:00
Stephen Kevin Burley（カリフォルニア大学サンディエゴ校）
「創薬へのPDBデータバンクのインパクト」（日本語解説付）
15:00－15:20 休憩
15:20－16:20
難波啓一（大阪大学・生命機能研究科）
「生命を支える生体分子機械 －ナノテクノロジーをはるかに超えた微細構造設計－」
16:20－16:30 質疑応答、討論
中村春木
Stephen K. Burley
難波啓一
一般社会人・学生（高校生以上）向け講演会
PDB データバンク：
タンパク質のかたちが支える生命科学と創薬への応用
日時
2012 年 10 月 13 日（土）
13:30－16:30（13:00 受付開始）
会場
ハートンホール毎日新聞ビル B 1
〒530－0001 大阪市北区梅田 3-4-5
主催
wwPDB Foundation
（国際蛋白質構造データバンク財団）
後援
（独）科学技術振興機構
バイオサイエンスデータベースセンター
大阪大学蛋白質研究所
大阪大学生命機能研究科
日本生物物理学会
日本蛋白質科学会
大阪医薬品協会
問合せ
Protein Data Bank Japan 事務局（堀氣）
Tel: 06-6879-4311
E-mail: nahokoh@protein.osaka-u.ac.jp
http://pdbj.org/pdbj_contact_j.html
■参加無料・事前申込不要
WORLDWIDE PDB PROTEIN DATA BANK
Worldwide Protein Data Bank Foundation
13:30－14:00
Haruki NAKAMURA, Osaka University
"wwPDB and its Impacts to Science and Society"
14:00－15:00
Stephen Kevin BURLEY, University of California at San Diego
"Impact of the Protein Data Bank on Drug Discovery"
15:00－15:20 Break
15:20－16:20
Keiichi NAMBA, Osaka University
"Molecular Nanomachines in Living Organisms
－Exquisite Structural Design far beyond State-of-the-Art Nanotechnology"
16:20－16:30 Q&A, Discussion
Haruki Nakamura
Stephen K. Burley
Keiichi Namba
wwPDB Foundation Outreach Seminar
Protein Data Bank:
Basis for Life Science and Drug Development
Date & Time
October 13, 2012 (Sat.)
13:30－16:30 (Opening of Reception 13:00)
Venue
Hearton Hall Mainichi Shimbun Bld. B1F
3-4-5 Umeda, Kita-ku, Osaka 530-0001, JAPAN
Organizer
wwPDB Foundation
Sponsors
National Bioscience Database Center
- Japan Science and Technology Agency
Institute for Protein Research, Osaka University
Graduate School of Frontier Biosciences, Osaka University
Osaka Pharmaceutical Manufacturers Association
Protein Science Society of Japan
The Biophysical Society of Japan
Contact
Protein Data Bank Japan Secretariat
E-mail: nahokoh@protein.osaka-u.ac.jp
http://pdbj.org/pdbj_contact.html
WORLDWIDE PDB PROTEIN DATA BANK
Worldwide Protein Data Bank Foundation
Free Admission -Advance Registration not necessary

wwPDB Interactions

- wwPDB leadership
  - Regular wwPDB Foundation phone meetings
  - Additional Skype and phone meetings
  - Yearly and ad hoc face-to-face meetings
- Common Tool for Deposition & Annotation Project
  - Weekly VTC meetings
  - Semi-annual in-person meetings
  - Daily phone, email and Skype meetings
- Regular annotator exchange visits
- NMR
  - Monthly phone/VTC meetings
- EMDB
  - Biweekly phone/VTC meetings

Activities for the Coming Year

- Roll-out of D&A system
- Phasing out of PDB format
- Production of PDBx format for deposition by refinement software
- Limited archive remediation
- Joint *pdb.org* website
- Joint publication about developments
- Continued Task Force activity
- International Year of Crystallography

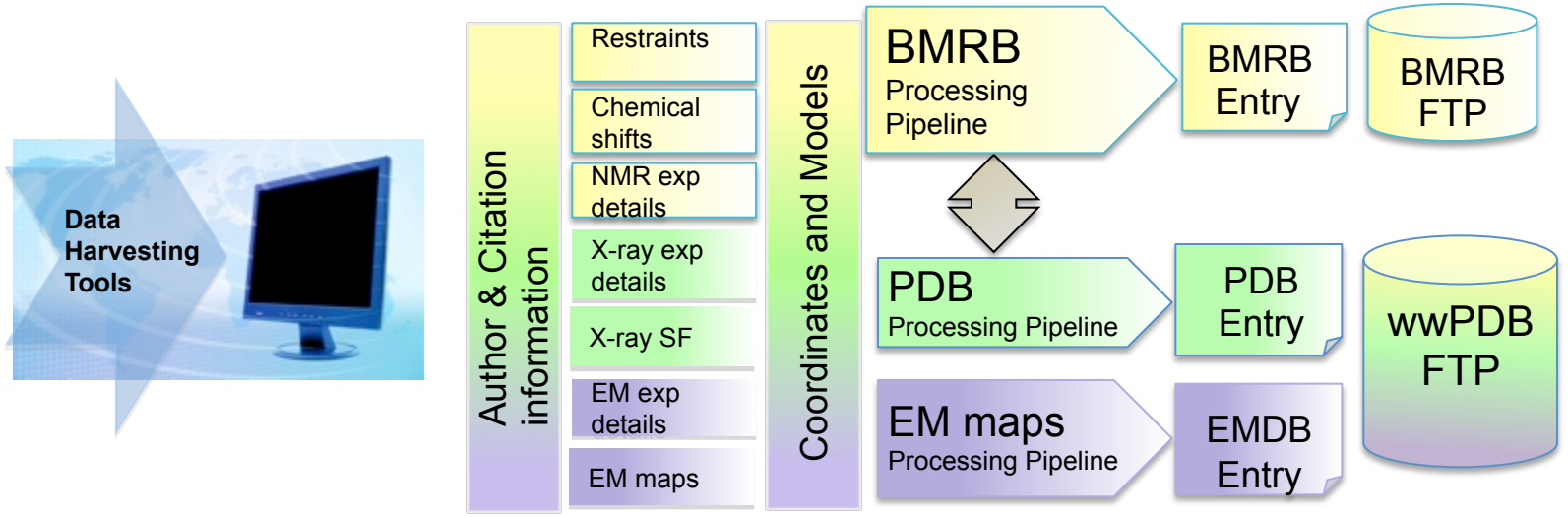
Common Deposition & Annotation (D&A) Tool

Martha Quesada

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

The Vision



Integrated Data Capture

Design Goals: Standardization, Quality and Efficiency

Supporting

- Larger and more complex biological molecules
- Expanded annotation
- Increased throughput: Automation and validation of routine submissions

wwPDB Common Deposition and Annotation Pipeline

Deposition Pipeline

Common Deposition Interface

X-ray-specific

EM-specific

NMR-specific

Data upload, Harvesting

Client-side Editor

Calculated annotations (PISA, SITE & LINK records, cross references, metal coordinates)

Validation

Submission

Progress Tracking/ Status

Common for coordinate processing
Method specific as appropriate

Communication System

Workflow-Automation System

Annotation Pipeline

Ligand Processing ID, Edit, Build

Sequence Processing

Validation

Calculated annotations (PISA, SITE & LINK records, cross references, metal coordinates)

Manual Annotation

Release Processing

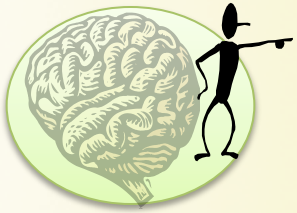
Green and yellow components are common for X-ray, NMR, & EM processing

D&A Deposition Pipeline Deliverables

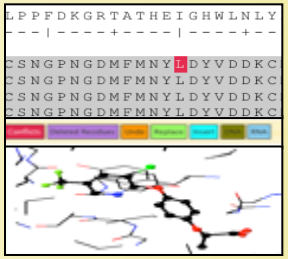
Deposition Pipeline

PDB EXTRACT

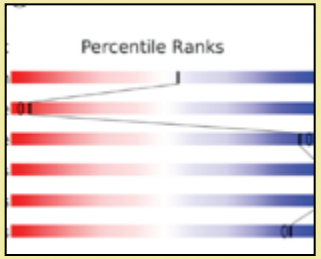
Data Harvesting



Deposition User Interface



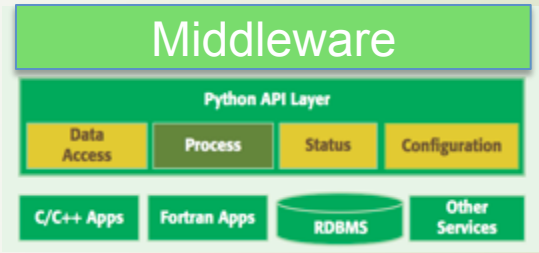
Functional Components

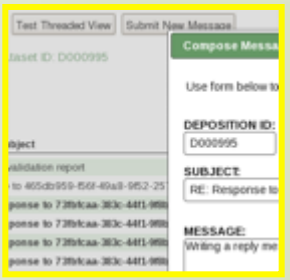


Infrastructure

PDBx

NetApp Snap Mirror Data Exchange





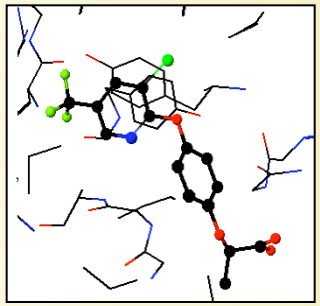
Communication

Release Processing

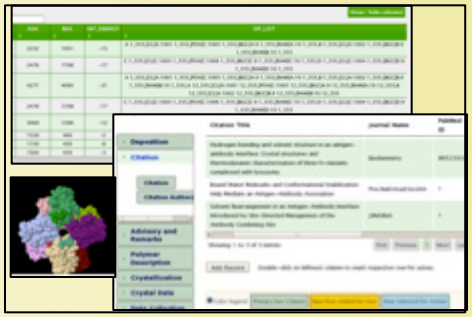
D&A Annotation Pipeline Deliverables

Annotation Pipeline

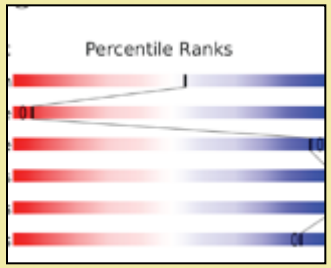
Sequence Processing



Ligand Processing



Manual and Automated Annotation

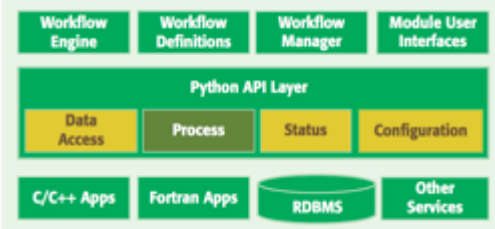


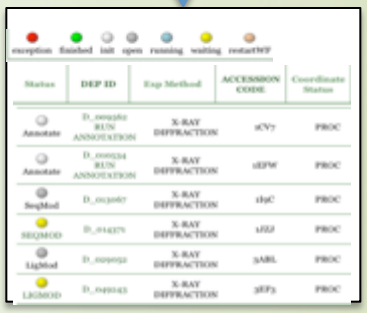
Validation

Infrastructure

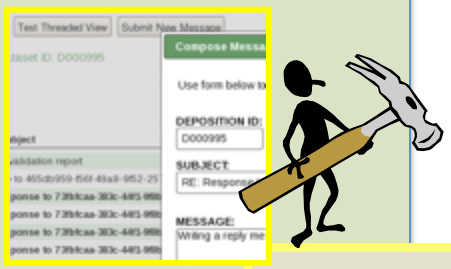
PDBx

NetApp Snap Mirror Data Exchange





Work Flow Manager

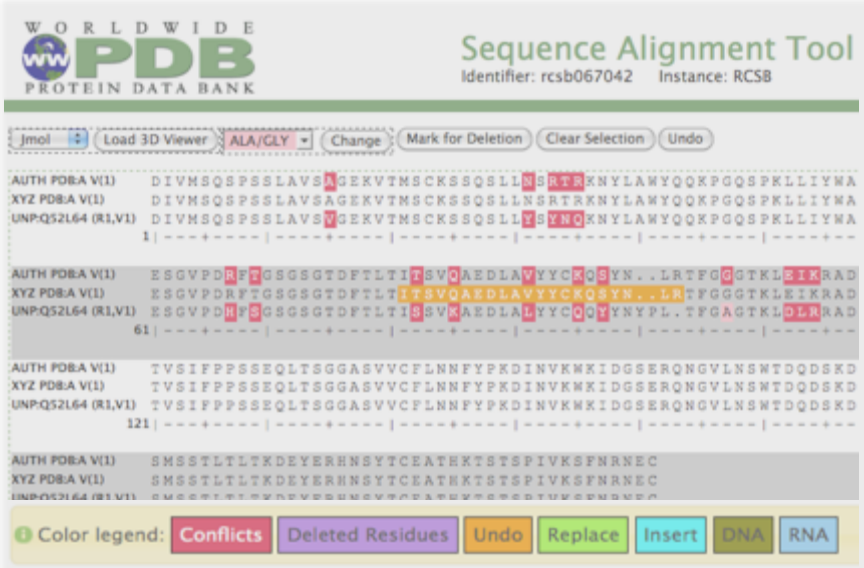


Communication

Release Processing

Processing Modules Updates

Sequence



2012 Enhancements

- Chimeric proteins supported
- Enhanced sequence match sorting
- Create new chemical definitions by splitting or merging existing definitions

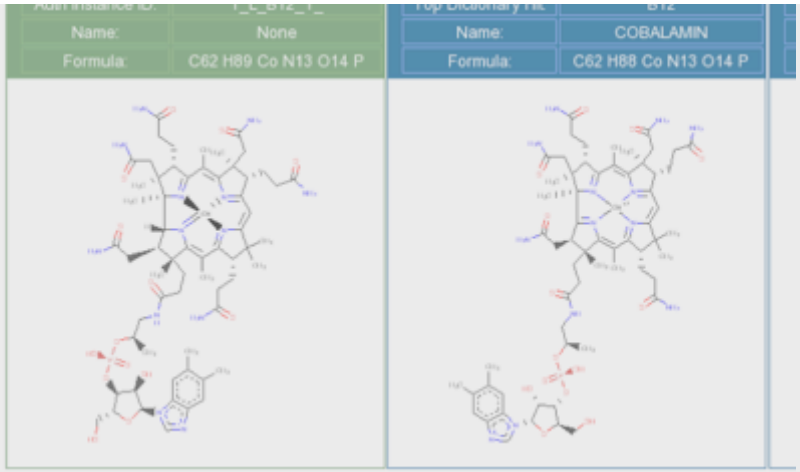
➢Unit and integration testing at all sites

Ligand

2012 Enhancements

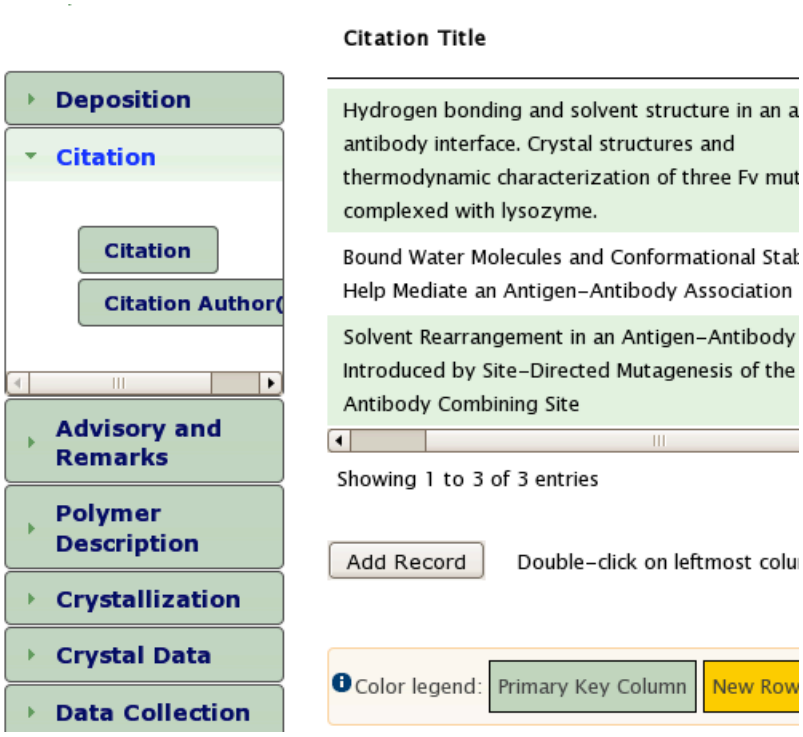
"Productionization"

- Create new chemical definitions by splitting or merging existing definitions
- Integration with Chemical Component Dictionary
- Component archiving User Interface
  - In production at RCSB PDB & PDBj
  - Unit and integration testing at PDBe



2012 Annotation Pipeline Deliverables

Web form data entry and editing



Annotation Tasks

- Dictionary check: Validation of PDBx, mmCIF, PDBML
- Biological assemblies (PISA)
- Automated annotation
  - Site environment
  - Solvent position
  - Linkage review module
  - Secondary structure

Code was refactored where needed, Middleware & User Interfaces for review and editing were developed

Validation Module

Requirements

X-ray VTF

Validation Functional Components

External Software
Protein geometry – MolProbity (Richardson lab)
Ligand geometry – Mogul (CCDC)
Structure factors – Xtriage (Phenix)
Map-model fit – EDS (Uppsala)

Internal Software
Nucleic-acid validation
Crystallographic symmetry clashes
Ligand stereochemistry and assignment
Sequence validation
R-factor comparisons from common refinement programs

Technical Components

D&A Workflow Manager

API (wrapper) for module integration

Validation Pipeline & Report Generation Software

mmCIF/PDBx Dictionary

Deposition Pipeline

Requirements → Functional Components + Technical Components

wwPDB Annotators

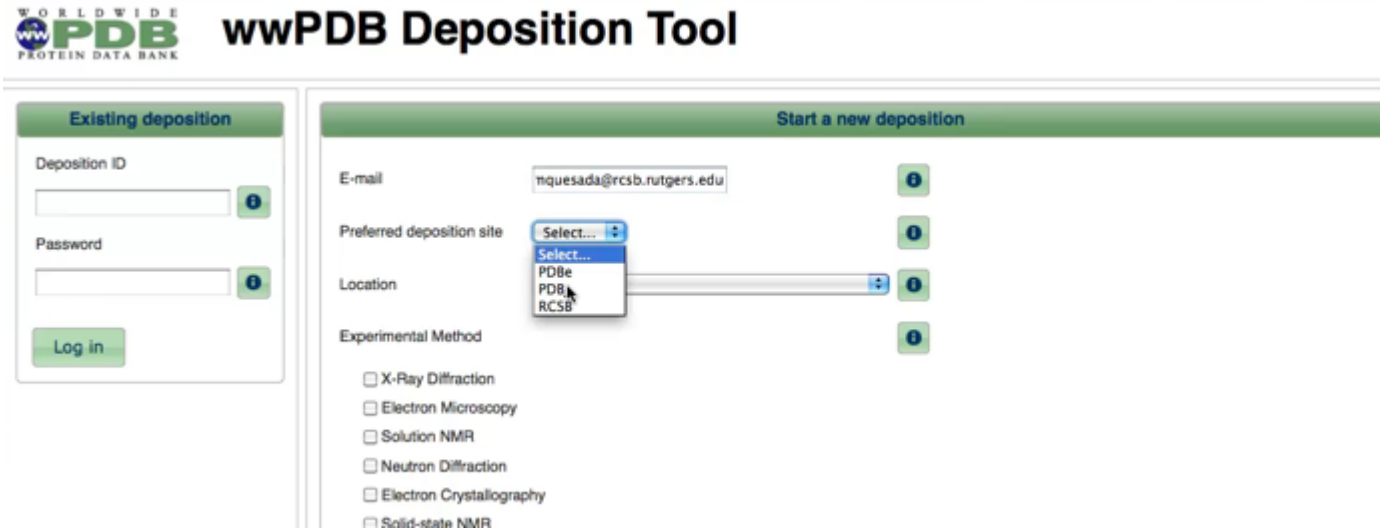
Internal Software
Load distribution
Harvesting (pdb_extract)
Webform pages
Validation module
Ligand module widget
Sequence module widget
Communication

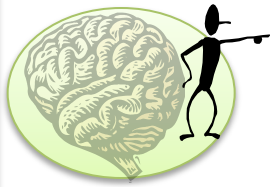
D&A Workflow System

Webform Middleware (XML)

API (wrapper) for functional components

Dictionary





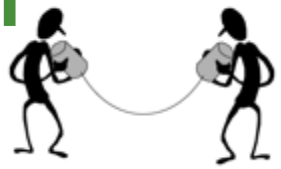
Workload Distribution

Distribution will take into account

- Restart of a deposition session
- Depositions based on previous entries
- Advisory and funding guidelines
- Time zone, to facilitate “help” and communication
- Load balance based on resource capacity
- User preferences

Single, wwPDB-branded, point of contact for all new depositions (e.g., http://wwpdb.org/deposit)

Communication Interface



- Look and feel of email
- Linked to web page content

EM V1.0

- Dictionary enhancements soon complete
- EM-specific interfaces being implemented
- Large data file requirements will be supported in the deposition module
- Ready for testing by end 2012

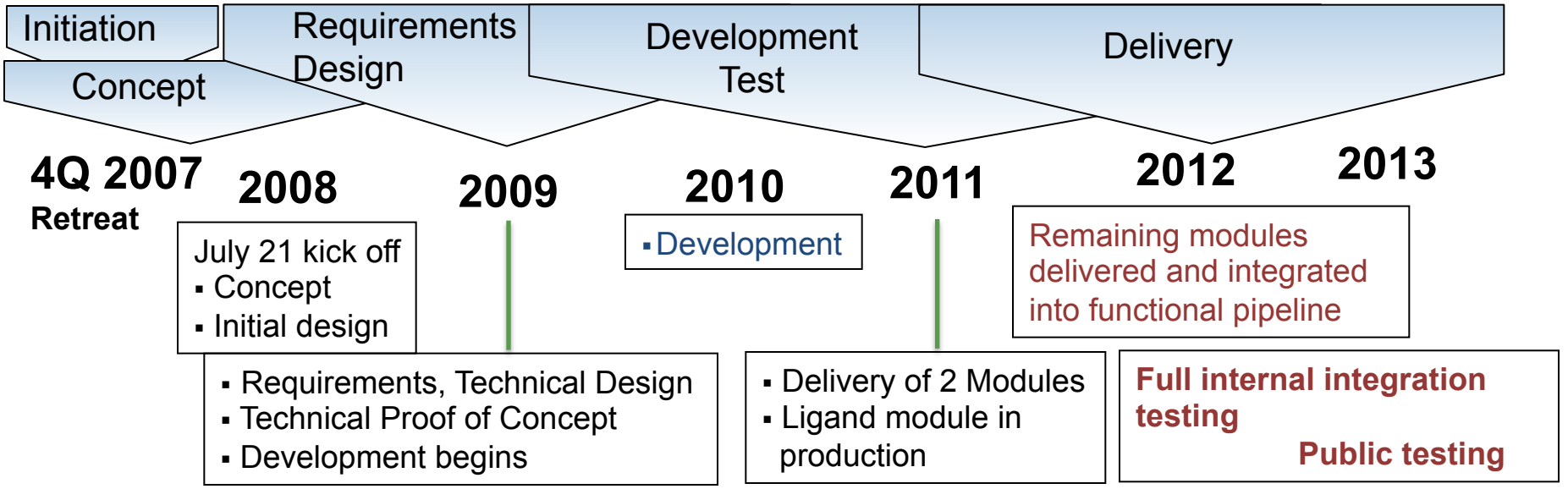
Future

- Additional visualization, data harvesting
- Validation requirements from EM VTF to be supported

NMR V1.0

- Dictionary data items in place
- Data requirements are defined and mapped for chemical shifts
- Integration of software for PDB atom nomenclature correspondence to NMR chemical shifts in place
- D&A NMR user interfaces being implemented
- Common D&A and ADIT-NMR data exchange to be implemented
- Ready for testing by end 2012

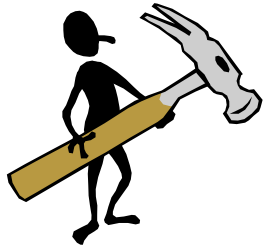
wwPDB Common D&A Tool Project



2012 Deliverable: wwPDB Common D&A System Version 1.0

- Able to process a file from deposition through annotation
- Supports all existing D&A processes and procedures
- Offers enhanced user interfaces for functional modules and deposition process (within reason)
- Provides a workflow infrastructure that enables task tracking and automation

Global Deployment Plan

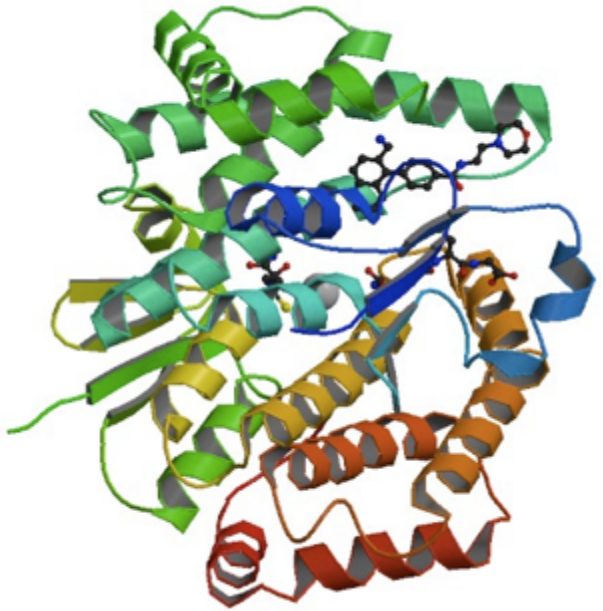


Depositor Perspective

1. Release for testing by select and representative group of depositors early 2013
2. Iterative development (refinement and debugging) based on feedback
3. Expand user base and continue dual deposition input stream
4. Before the end of 2013 - switch all NEW depositions to new system. Incomplete legacy depositions to be completed in “old” systems

The wwPDB Common Tool Demo

Deposition and Annotation of
4EC0, Human hematopoietic
Prostaglandin (PG)
D2 synthase (hH-PGDS)



Three Ligands: Glutathione,
4-[2-(aminomethyl)naphthalen-1-yl]-N-
[2-(morpholin-4-yl)ethyl]benzamide, Mg^{2+}

Pfizer, USA

Common Deposition Interface

Ligand Processing

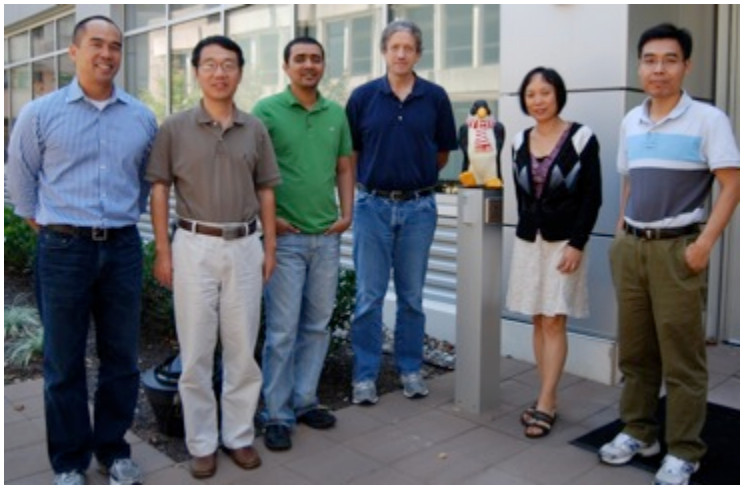
Sequence Processing

Annotations
(PISA, SITE & LINK records, cross references, metal coordinates)

Validation

Common D&A Project Team

Experience, *Expertise* and *Diverse Skills*
representing the broad interests of the wwPDB



Format, Raw Data and Validation

Gerard Kleywegt

Why the PDB Format Must Go

- Problem: PDB format is almost 40 years old and does not support today's science
  - Let alone tomorrow's science…
- Some of the limitations
  - Max 62 chains
    - and that's stretching it
  - Max 99,999 atoms
    - 5 ribosomes in ASU=10 PDB entries!
  - Very short chain, residue and atom names
    - 1, 3, 4 characters, respectively
  - No bond orders or chirality specified for ligands
  - No support for NMR, EM, hybrid methods, …
  - Meta-data specification cumbersome and inflexible

Towards a “new PDB format”

- 2010 – started process of defining new format, consulting many software developers
- 2011 – agreement to adopt PDBx (mmCIF) as the new format and to phase out the old PDB format
  - Commitments from CCP4, Phenix and Global Phasing (*i.e.*, ~85% of all PDB depositions)
  - Agreement on managing development between these software providers and wwPDB
  - Established working group
  - Projected completion – January 2013

Update on “New PDB Format”

- PDBx/mmCIF Deposition Working Group
  - Goal: support deposition of X-ray structures in PDBx format
  - Participants: developers of major X-ray software packages and wwPDB staff
  - Continued virtual meetings to discuss content and representation issues
  - Paul Adams has replaced Oliver Smart as chair
  - Delivery target early 2013

Format Compatibility Plan

- Adopt a *PDB-friendly* mmCIF/PDBx style
  - All records on a single text line
  - Columns presented in standard column order
  - Tabular presentation with leading record names
    (*e.g.* ATOM, CELL, REFINE)
  - Method-independent features in left-most columns
    (*e.g.* identifiers & coordinates)
  - Method-specific features in the right-most columns
    (*e.g.* ADPs, NMR order/disorder parameters)
  - Continue to support PDB nomenclature semantics
    (*e.g*. PDB-style chains, residue numbering and insertion codes)

```
ATOM      1  N   GLN A  39      24.690 -27.754  24.275  1.00 60.76           N
ATOM      2  CA  GLN A  39      23.581 -26.768  24.416  1.00 60.98           C
ATOM      3  C   GLN A  39      23.990 -25.379  23.905  1.00 59.98           C
ATOM      4  O   GLN A  39      25.070 -25.209  23.330  1.00 60.25           O
ATOM      5  CB  GLN A  39      23.136 -26.685  25.878  1.00 60.69           C
ATOM      6  N   VAL A  40      23.115 -24.395  24.122  1.00 59.58           N
ATOM      7  CA  VAL A  40      23.342 -23.010  23.690  1.00 57.26           C
ATOM      8  C   VAL A  40      24.000 -22.152  24.778  1.00 56.00           C
```

PDB

PDBx/mmCIF

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.type_symbol
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.pdbx_PDB_model_num
_atom_site.occupancy
_atom_site.pdbx_auth_alt_id
_atom_site.B_iso_or_equiv
ATOM   1  N   N  GLN  A  39  24.690  -27.754  24.275  1  1.00  .  60.76
ATOM   2  CA  C  GLN  A  39  23.581  -26.768  24.416  1  1.00  .  60.98
ATOM   3  C   C  GLN  A  39  23.990  -25.379  23.905  1  1.00  .  59.98
ATOM   4  O   O  GLN  A  39  25.070  -25.209  23.330  1  1.00  .  60.25
ATOM   5  CB  C  GLN  A  39  23.136  -26.685  25.878  1  1.00  .  60.69
ATOM   6  N   N  VAL  A  40  23.115  -24.395  24.122  1  1.00  .  59.58
ATOM   7  CA  C  VAL  A  40  23.342  -23.010  23.690  1  1.00  .  57.26
ATOM   8  C   C  VAL  A  40  24.000  -22.152  24.778  1  1.00  .  56.00
```

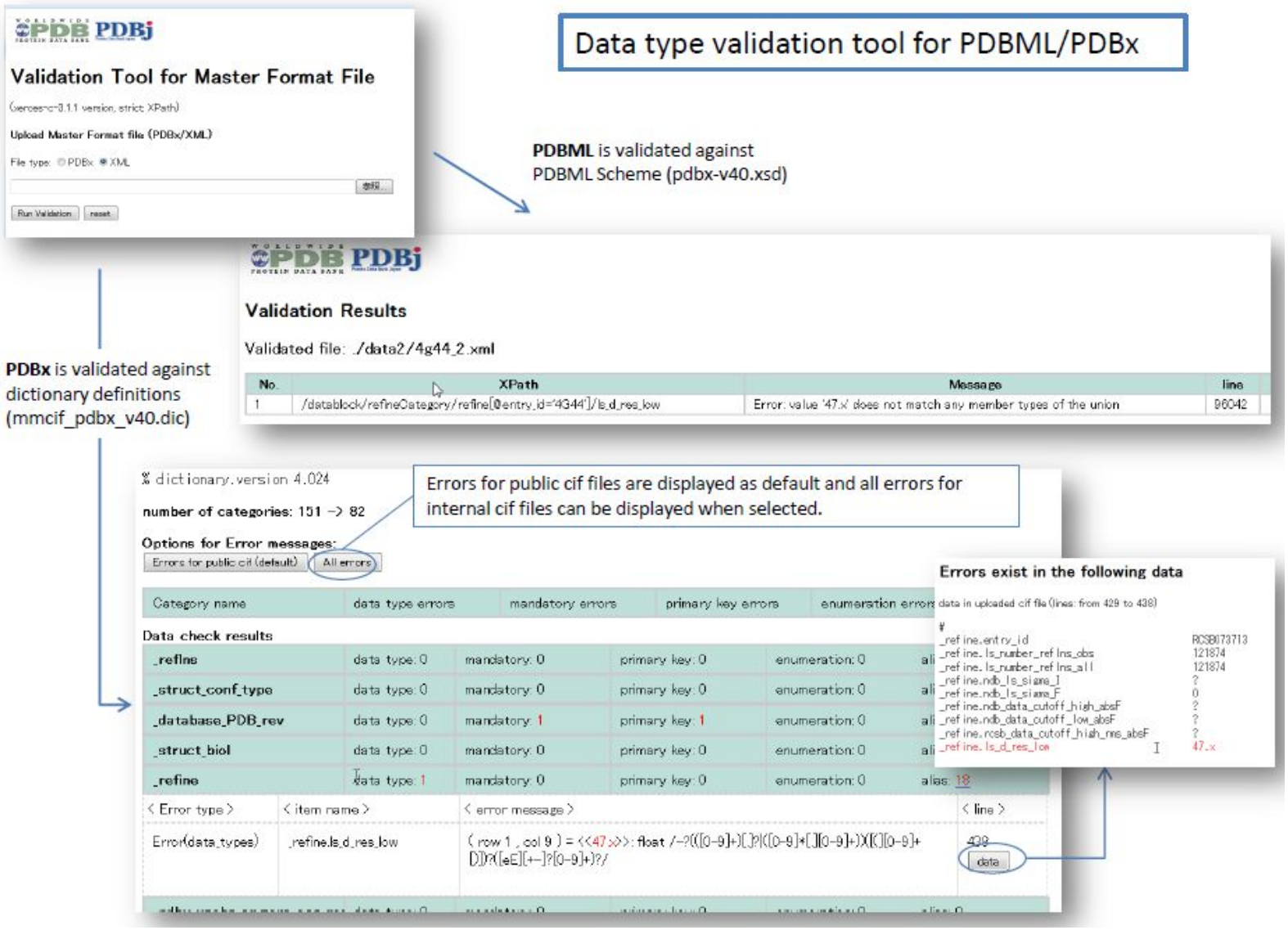
Guiding Principles for Deployment

- Preserve backward compatibility where possible
- Changes that do not fit the within the current PDB format will be implemented only if needed (e.g. to represent a large molecule)
  - Atom serial numbers
  - Chain identifiers
  - Residue names and numbers
- Continue to assign residue-level 3-letter codes even if more descriptive identifiers are adopted (e.g. for monosaccharides)

Future Support for the Current Format

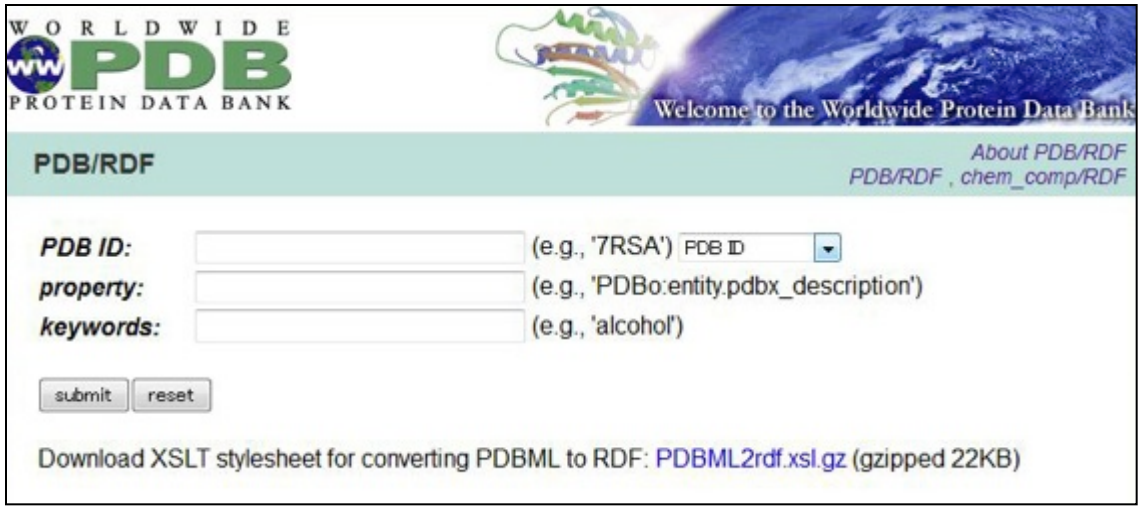
- Web service to create current PDB format data files
- PDB-like report format

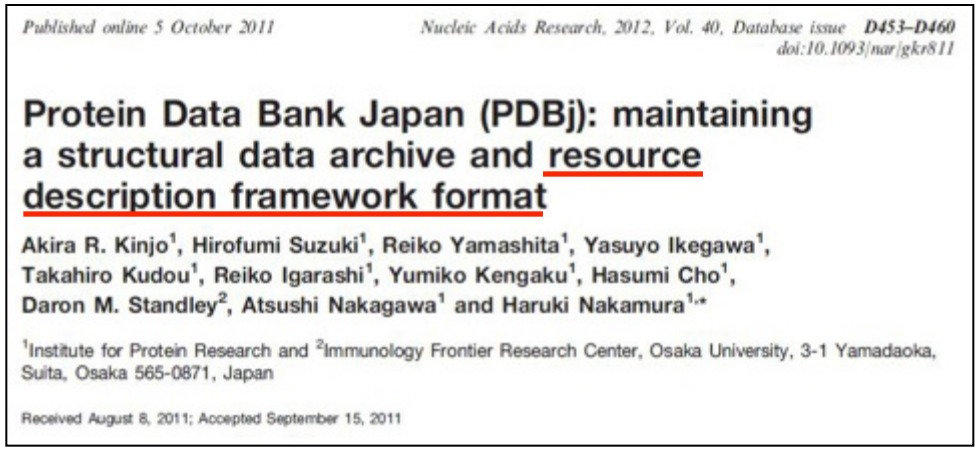
Validation of PDBx/mmCIF and PDBML Files



PDB/RDF Format for Semantic Web

http://rdf.wwpdb.org/





Published online 5 October 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* **D453–D460**
doi:10.1093/nar/gkr811

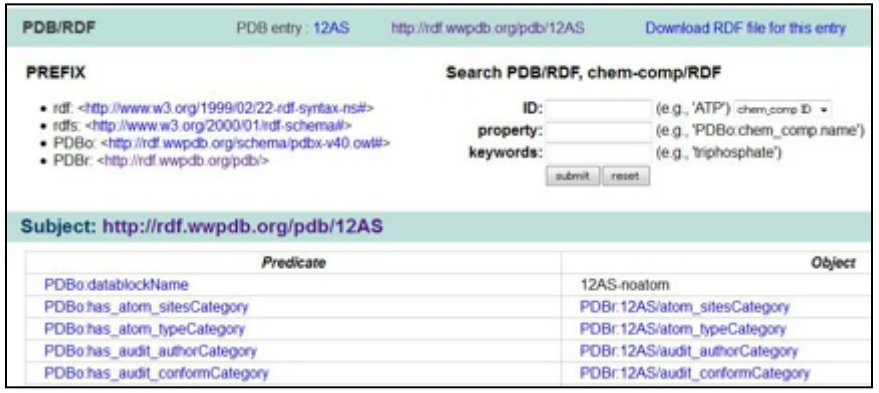
Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format

Akira R. Kinjo[1], Hirofumi Suzuki[1], Reiko Yamashita[1], Yasuyo Ikegawa[1], Takahiro Kudou[1], Reiko Igarashi[1], Yumiko Kengaku[1], Hasumi Cho[1], Daron M. Standley[2], Atsushi Nakagawa[1] and Haruki Nakamura[1,*]

[1]Institute for Protein Research and [2]Immunology Frontier Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan

Received August 8, 2011; Accepted September 15, 2011

Example: http://rdf.wwpdb.org/pdb/12as

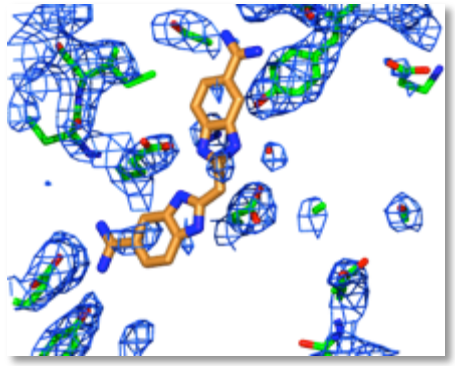


Raw Data Archiving

- Participated in workshops at IUCr Madrid, ACA Boston and ECM Bergen
- Surveyed existing repositories of image data and found usage to be limited
- Converging *communis opinio* suggests image storage at synchrotron sites and institutional repositories where supported
- Assign DOIs to publicly accessible data sets and link DOIs to PDB entries
- Future: unmerged intensities and NOE peak lists

Validation by wwPDB – where are we (heading)?

Questions



- Entry-specific validation (quality control)
  - Is this model ready for archiving and publication?
  - Is this model a faithful, reliable and complete interpretation of the experimental data?
  - Are there any obvious errors/problems?
  - Are the conclusions drawn in the paper justified by the data?
  - Is this model suitable for my application?

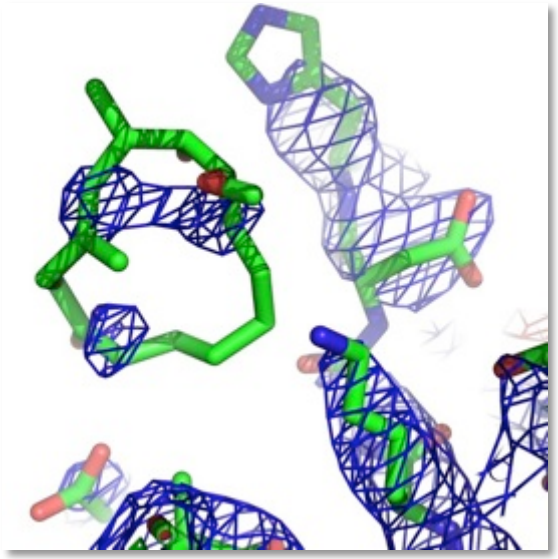
- Archive-wide validation
  - What is the best model for this molecule/complex to answer my research question?
  - Which models should I select/omit when mining the PDB?

Validation in a Nutshell

- Compare a model to the experimental data and to the prior knowledge. It should:
  - Reproduce knowledge/information/data used in the construction of the model
    - R, RMSD bond lengths, chirality, …
  - Predict knowledge/information/data not used in the construction of the model
    - R_{free}, Ramachandran plot, packing quality, …
  - Global and local
  - Model alone, data alone, fit of model and data
  - … and if a model fails to do this, there had better be a plausible explanation!

Validation - Outline

- wwPDB strategy for validation
- Implementing the wwPDB X-ray validation pipeline
- Update on wwPDB validation plans for NMR and EM



Validation

wwPDB strategy for validation

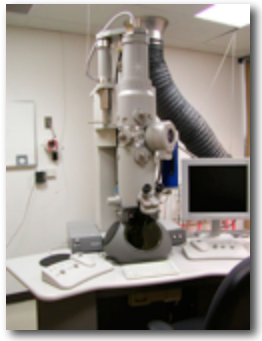
Two Major Recent wwPDB Projects

- Development of a new joint wwPDB Deposition and Annotation (D&A) system
  - Will handle X-ray, NMR, EM, …
  - Will be used at all wwPDB sites
  - Replaces ADIT, AutoDep, EMdep, parts of ADIT-NMR
  - Public release 2013

- Validation using community-recommended methods will be integral part of new D&A
  - 2008: X-ray Validation Task Force (VTF)
  - 2009: NMR VTF
  - 2010: EM VTF
  - Implementation of recommendations in validation-software pipelines

Validation by wwPDB - Advantages

- Applies community-agreed methods uniformly
- Improves the quality and consistency of the PDB archive
- Supports editors and referees
- Helps users assess if an entry is suitable
- Helps users compare related entries
- Enables identification of outliers when mining the PDB
- Stimulates adoption of better protocols by the community



Experimental Data for Structure Papers

We are writing to address the retraction of five papers on structural studies of ATP-binding cassette (ABC) transporters—three in *Science* (G. Chang *et al.*, "Retraction," Letters, 22 Dec. 2006, p. 1875), one in the *Proceedings of the National Academy of Sciences* (*1*), and one in the *Journal of Molecular Biology* (*2*). We have much sympathy for your readers but very little for the magazine. This is not the first time incorrect structures have been published in *Science* (*3*), and it will not be the last time. We

Since February 2008, data deposition mandatory for PDB!

Storing diffraction data

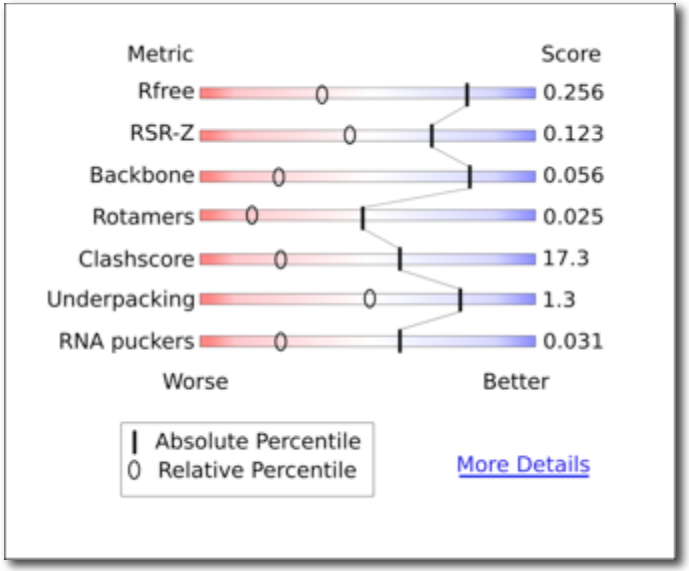
SIR — *Nature* must consider structural biology to be of some interest to its readership, as almost every issue contains a new macromolecular structure. We there-

active-site nucleophil
called 'disallowed' co
we would expect up to
glycine residues to

Mandatory Data Deposition

- This is great news!
  - Data available for all eternity
    - Even after a student/post-doc has left the lab…
  - Calculate maps
    - How credible is the density for a ligand or active site?
    - Are there any unmodelled features in the density?
  - Re-refine models
  - Re-interpret data
  - Validation using data
    - R, R_{free}, real-space fit, k_{sol}, B_{Wilson}, …
  - Assess outliers identified by coordinate-only validation methods
    - (1) genuine, but unusual feature of the structure?
    - (2) probable error in the model?

wwPDB X-ray VTF



- "Relative" = compared to structures of similar resolution
- Reference values and distributions will be recomputed annually

Report: Read *et al*., *Structure* **19**, 1395 (2011)

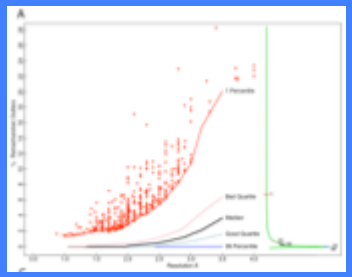
Validation

Implementing the wwPDB X-ray validation pipeline

wwPDB X-ray Validation Pipeline

External reference files (*e.g.*, Engh & Huber)

Distributions

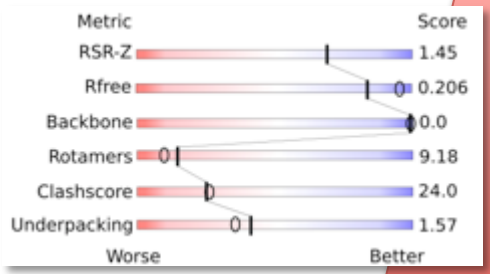


Validation XML file

Validation pipeline 1.0 *

MolProbity | Xtriage | EDS | Mogul | Percentiles | PDF maker

Deposited data (coordinates & reflections)

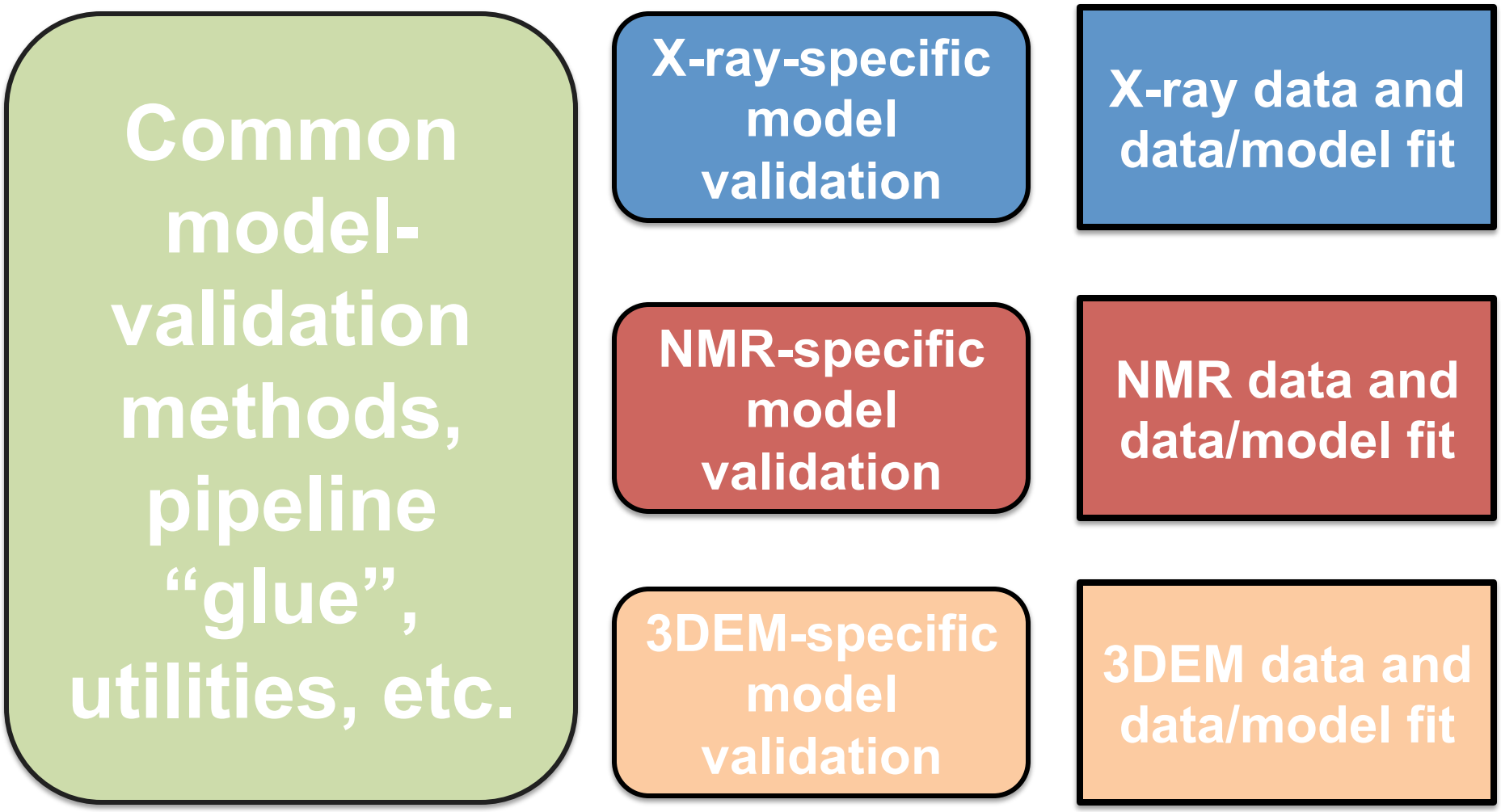


PDF report for depositor & referees - Statistics and plots for the entry, per chain, per residue, and list of unusual features

* WhatCheck coming soon

Gore *et al.*, *Acta Cryst.* **D68**, 478 (2012)

Validation-Software Components

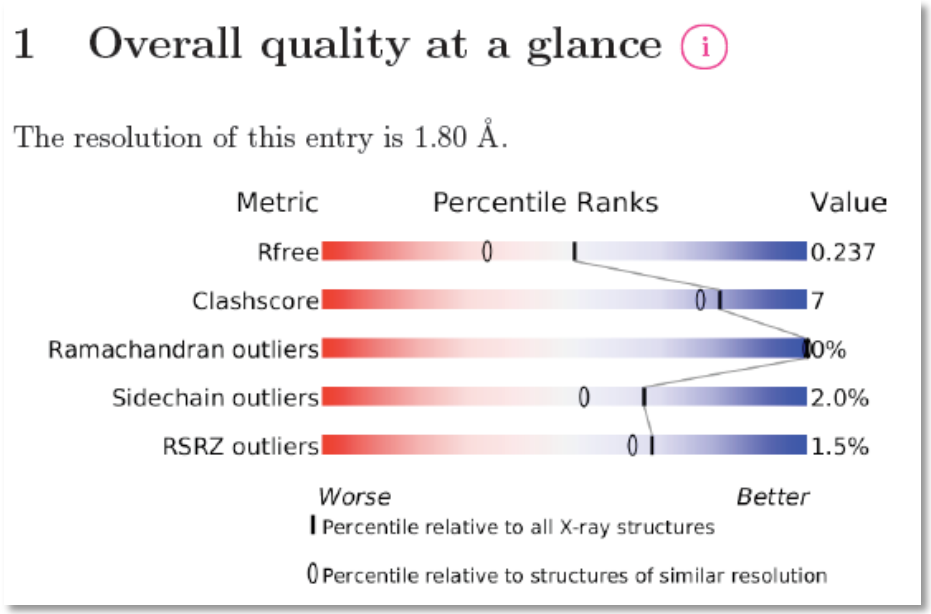


What Does it Mean for a Crystallographer?

- There will be three uses of the validation pipeline
  - At deposition time
    - Not all checks can be run, e.g. some sequence and ligand checks
    - Report for depositor
  - At annotation time
    - Complete validation report, also suitable for editors/referees
  - Independently of deposition
    - Anonymous web-based server to use on models not (yet) in the PDB
    - Will be developed once the production pipeline is up and running
    - Will not be available as a stand-alone software package

What Will a Validation Report Include?

- Report = summary
  - Gory details in XML file
  - Explanations on web site
- Title page
  - Authors, title, PDB code (if assigned), time-stamp
- Overall quality at-a-glance
  - Slider plots of key statistics
- "Table 1"
  - Key data and refinement stats
- Entry composition
  - Macromolecules (including sequence diagnostics, if available)
  - Ligands (including diagnostics, if available)

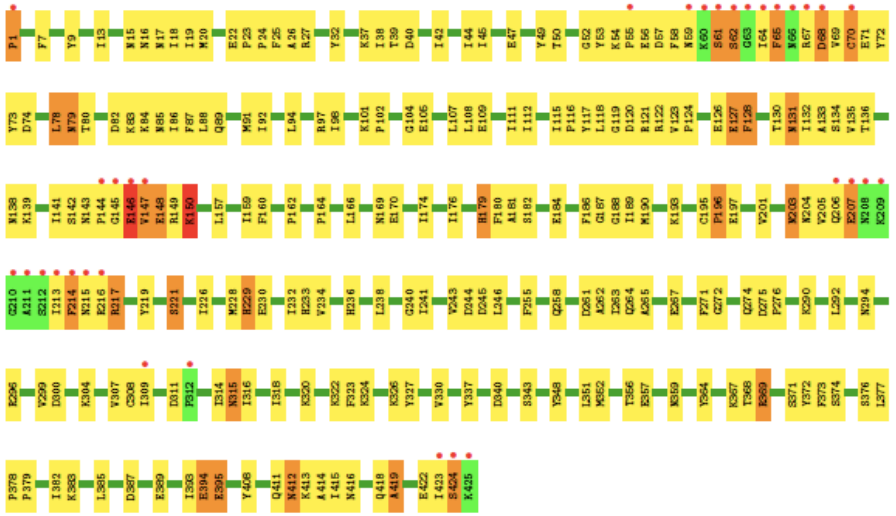


What Will a Validation Report Include?

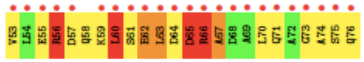
- Model quality
  - Bond lengths and angles (outlier info, RMS-Z)
  - Chirality, planarity
  - Close contacts (incl. clashscore, worst clashes)
  - Torsion angles (Ramachandran, rotamers for proteins)
  - Ligand geometry (Mogul analysis)
- Model/data fit
  - Macromolecules: RSR, RSR-Z, B-factors, partial occupancies
  - Ligands: same, but RSR-Z undefined
- Residue plots
  - Residues with model-quality outliers (0, 1, 2, >2)
  - Residues with RSR-Z > 2 get a •
  - Unmodeled residues

Residue Plots

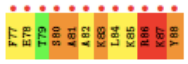
- Molecule 1: BOTULINUM NEUROTOXIN TYPE B
  Chain A:



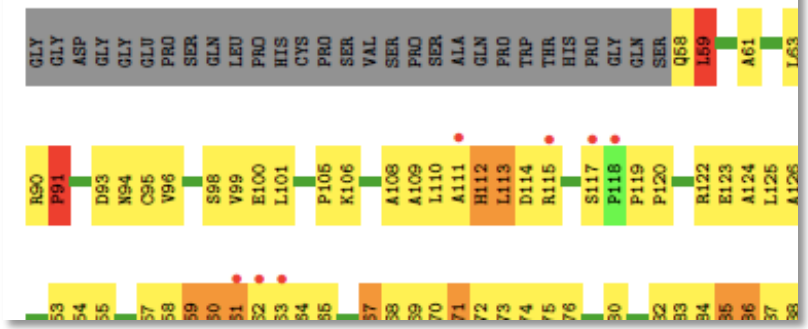
- Molecule 2: SYNAPTOBREVIN-II
  Chain B:



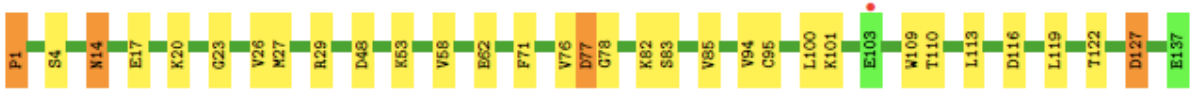
- Molecule 3: SYNAPTOBREVIN-II
  Chain C:



- Molecule 1: MEMBRANE COPPER AMI
  Chain A:



- Molecule 1: CELLULAR RETINOIC ACID BINDING PROTEIN TYPE II
  Chain A:



Status and Timeline

- MolProbity, EDS, Mogul, Phenix modules, nucleic-acid validation, symmetry clashes, ligand and sequence validation implemented
  - Archive coverage at present >98% (EDS ~95%)
- Production of XML file and PDF report done
- Contents and presentation/wording details of report fine-tuned with wwPDB partners and wwPDB X-ray VTF
- Integrated into new wwPDB D&A system
- Debugging of the pipeline on-going
- Internal and external testing
- Version 1.0 for public release Q1 2013
- Version 2.0 will include WhatCheck plus changes based on feedback and experience

Validation

Update on wwPDB validation plans for NMR and EM

What About Other Methods?

- Model validation using same criteria as X-ray
  - MolProbity, WhatCheck, Mogul
- Some special model-related issues per technique
  - X-ray: alternative conformations
  - NMR: ensemble of models; ill-defined regions
  - 3DEM: clashes of rigid-body fitted models; wrong species
- Data quality and model/data-fit assessment will be different for each technique

wwPDB NMR VTF Update

- Report with recommendations in preparation
- Mock-ups and specific questions sent to VTF
- Geometry validation as for X-ray
- Well-defined *vs* not-well-defined regions
- Treatment of NMR ensembles

- Chemical-shift validation to report completeness, outliers, referencing corrections
- Constraints validation to report number for each category (intra-residue, sequential, intermediate, long-range, inter-chain) and worst violations

NMR VTF: Montelione, Nilges *et al*., to be published

NMR Validation Work and Plans

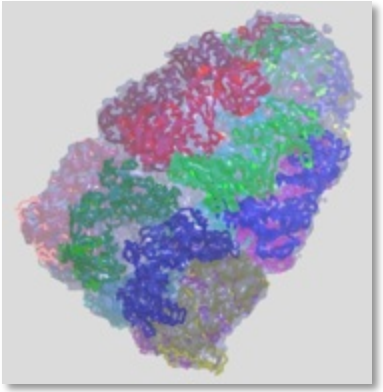
- FindCore (PSVS) and NMRCore (Olderado) algorithms in place for definition of well-defined regions
- Starting adaptation of the X-ray pipeline for NMR entries (MolProbity and Mogul)
- Once the recommendations are finalised
  - Sanity checks on nomenclature for constraints and coordinates (already in place for chemical shifts)
  - Chemical shift report (referencing corrections, outliers, completeness)
  - Calculation of global and per-residue scores for geometry validation (need advice on ensemble representations)
  - Identification of constraint violations
  - Validation of RDCs and other types of experimental restraints

EMDataBank EM VTF Update

- Main recommendations for EM maps
  - Standards for assessing resolution and accuracy of a map need to be developed
  - Structural features in a map should be in accordance with the claimed resolution
- Main recommendations for models fitted into EM maps
  - Criteria for assessing models need to be developed
  - Capability to archive coarse-grained representations of models is needed
- More research and development needed!

3DEM VTF: Henderson, Sali *et al*., *Structure* **20**, 205 (2012)

EM Validation Work and Plans



- D&A 1.0
  - Map visual analysis (Chimera): visual sanity check of the map and map/model overlay
  - Minimal model validation à la X-ray
- Later
  - Harvest more validation-related data (e.g., results of tilt-pair analysis)
  - As new methods are developed and become community-accepted they can be incorporated into the validation pipeline

NMR - BMRB

John Markley

WORLDWIDE PDB
PROTEIN DATA BANK

wwpdb.org

Depositions (09/15/11 – 09/14/12)

- 828 new BMRB depositions
  - 62% associated with coordinates
- 556 new combined PDB and BMRB entries
  - 40 new coordinate sets associated with earlier BMRB depositions
  - 498 new depositions through BMRB
  - 28 new depositions through PDBj-BMRB
  - 30 new depositions through PDBe
- BMRB has handled restraint validation with help from European colleagues

Additional Activities

- Participation in the wwPDB common deposition and annotation (D&A) project
  - Interface development
    - NMR mock-up design support
    - Incorporation of NMR-STAR data items into the PDBx dictionary
  - Infrastructure development
    - NetApp data exchange
    - BMRB D&A test platform
  - Software development
    - NMR-STAR/PDBx data conversion
    - Coordinate/chemical shift atom nomenclature checker

Funding Status

- Grant (~70% of previous award) through August 2014
- Funding cuts have reduced staffing by one annotator and two programmers
- We prepared a “white paper” describing the activities of BMRB and its funding requirements, which was approved by the BMRB Advisory Board and sent to representatives of US granting agencies (NIH, NSF, and DOE)
- The white paper elicited little positive support from the granting agencies

Community Support for BMRB

- Editorial in *Nature Structural and Molecular Biology*
  - The Editor (Ines Chen) learned of the funding situation at BMRB and offered to look into writing an editorial
  - The result was an editorial plus 7 pages of testimonials from scientists in the field volunteering their strong support for BMRB
- Editorials in *Nature* and *The Scientist*
- Coverage by science bloggers

FEATURE

In support of the BMRB

The Biological Magnetic Resonance Bank (BMRB) is facing the threat of having its funding discontinued. Concerned about this situation, the editors of *Nature Structural & Molecular Biology* have asked the community why it is important to continue to support the BMRB. We have also asked John Markley, head of the BMRB, to present his case.

A word from the BMRB

Stable funding is essential for a data bank such as the BMRB. Ideally, the funding should support maintenance activities (data deposition, storage and dissemination), growth of the data bank to track the emergence of new types of data, and development of improved technology to reduce costs and improve the impact of the data bank. The BMRB has played a key part in developing standards for the representation of biomolecular NMR data, and continued efforts in this area are needed as new kinds of data, such as those for small-angle X-ray (or neutron) scattering and cryo-EM, are reported and need to be archived. The BMRB, through its association with the Worldwide Protein Data Bank (wwPDB), is participating in the development of new standards and software for the validation of structures determined by NMR spectroscopy. Opportunities exist for expediting the creation of (more extensive) BMRB depositions through collaboration with instrument manufacturers and software developers. Such developments can facilitate the deposition of peak lists associated with assignments and structure determination, as recommended by our advisory board members. The challenge of the future will be in linking information across different data banks. The wwPDB is leading the way in demonstrating how this can be done.

Most grant regulations now require the timely deposition of experimental results, and an increasing number of journals have data deposition as a requirement for publication. Several growing areas of research are making extensive use of the BMRB. These include investigations of intrinsically disordered proteins, development of automated analysis of NMR data, solid-state NMR and NMR-based metabolomics.

With the budget cuts that the BMRB has suffered (reduced by 40%, compared to the previous operating budget), we currently are at the minimal level of keeping up with depositions, data validation and data-dictionary development. In addition, the BMRB is barely managing to meet its obligations as a partner in the wwPDB. We have had to lay off people who were developing new software and functionality. The wwPDB advisory-committee meeting, held at Rutgers University on 1 October 2010, had a session on funding, which enabled us to inform members of the US granting agencies about the impending expiration of remaining funding from the National Library of Medicine in September 2014. To date, no plan has been advanced to keep the BMRB functioning. None of the three agencies has expressed an interest in funding more than a part of the needed budget, so a multiagency approach appears to be needed. To stimulate this, BMRB staff members prepared a 'white paper' (see **Supplementary Note**), which was approved by its advisory board and then sent to representatives of the US grant agencies (National Institutes of Health (NIH), Department of Energy and National Science Foundation). Given the lead time for applications and review, it appears critical that a funding plan be developed within the coming year.

John L. Markley, University of Wisconsin–Madison, Madison, Wisconsin, USA

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nsmb.2371

Voices from the community

The BMRB is playing a very important part in determining structures and elucidating functions and interactions between biological molecules by NMR. The BMRB unit of Protein Data Bank Japan (BMRB-PDBj) has collected more than 600 chemical-shift data sets produced by RIKEN in the Protein 3000 project. Even if some of the structures themselves might not be that important, the chemical shift data can be used for drug discovery and to understand how they relate to secondary and tertiary structures. That relationship is used in many software tools for NMR data analysis, such as TALOS, SHIFTX and SPARTA. Thus, the chemical shift data are an important outcome of the structural proteomics effort, and this is the reason why the BMRB is a member of the wwPDB. The software programs mentioned above are commonly used by biological NMR researchers around the world. Personally, I am also using the database and software tools in my own research. I think they are indispensable in biological NMR and related fields.

The BMRB was established by John Markley at the University of Wisconsin–Madison. Now, a network with BMRB-PDBj and the European Bioinformatics Institute (EBI) has been formed, and the BMRB has an essential role in the development and management of the database. To provide a high-quality NMR database to researchers in the life sciences and related fields, the activity at the BMRB should be kept in full swing.

Hideo Akutsu, Institute for Protein Research, Osaka University, Osaka, Japan

Current Plan

- We learned subsequently and indirectly that the NIGMS would accept an “R01” application for BMRB support
  - We intend to submit an application for the February 5, 2013 deadline

3D Electron Microscopy, SAS, Hybrid Methods

Helen Berman

EMDataBank: Unified Data Resource for 3DEM

- Collaborative project between PDBe, RCSB PDB and Baylor-NCMI
- "*One-Stop Shop*" for collection of EM maps and coordinate models
- Standardized map redistribution format
- Cross referencing between maps and models (EMDB←→PDB)

EMDataBank Services

- **Joint map + coordinate deposition**
- News, software list, information about dictionaries, conventions, FAQ, community links
- Search by ID, author, sample type, keyword, deposition date
- Recently released entries
- Map+model 3D java viewer

PDB-EMDB merger:
EM to be part of Common D&A Tool

6-March-2012

EM Data Bank joins the PDB archive

The EM Data Bank (EMDB), the primary archive for experimentally-determined maps obtained using three-dimensional electron microscopy methods, has joined the PDB archive (**ftp://ftp.wwpdb.org**), as **announced previously**.

The merger makes 3DEM results available in a single archive, including over 1300 electron microscopy derived maps from EMDB and 400 coordinates for EM map-derived models in PDB. It is also an essential step in the wwPDB's development of a Common Deposition & Annotation Tool that will cover all experimental methods, including hybrid methods.

With the addition of EMDB data, the physical size of the complete wwPDB archive jumps to roughly 180 GB (from its previous 130 GB). Sites that mirror the full wwPDB archive will need to increase storage capacity accordingly.

Summary information regarding the merger and detailed specifications for data access are posted at **wwpdb.org/em/**.

EMDataBank Project Funding

- Funded by NIH, BBSRC, and EMBL
- Collaborative Grant for 3DEM Validation under review by NIH

Late breaking news!

Grant is in the 5th percentile!!

EMData Bank

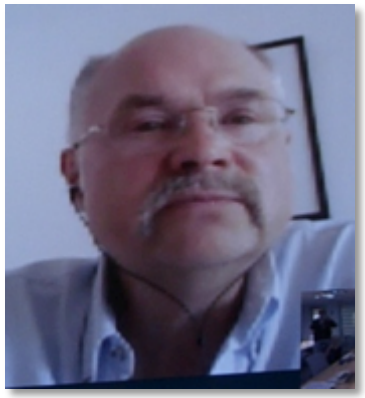
Wah Chiu PI, Helen Berman and Gerard Kleywegt coPIs

Specific Aims

1. **Establish map-validation methods**
   - Use representative raw image datasets from both our laboratory and broad group of collaborators
2. **Establish model-validation methods**
   - Use map and model data from EMDB and PDB and community-contributed data
3. **Define standards for 3DEM data exchange and archiving**
   - Continue development of 3DEM terms in the EMDB data model and PDBx by adding metadata relevant to the validation procedures established above
   - Establish an agreed upon data exchange file format for maps, and develop or modify software converters to support the new and current data formats
4. **Facilitate the dissemination of 3DEM validation standards**
5. **Integrate 3DEM data standards and map and model validation into the wwPDB pipeline**
   - The map validation metadata and map-derived model-validation procedures developed through this project will be integrated into the wwPDB D&A system

SAXS/SANS Task Force

- Members
  - Jill Trewhella (Chair, University of Sydney)
  - Dmitri Svergun (European Molecular Biology Laboratory-Hamburg)
  - John Tainer (The Scripps Research Institute)
  - Wayne Hendrickson (Columbia University)
  - Mamoru Sato (Yokohama City University)
  - Torsten Schwede (University of Basel)



SAS Committee Charge

- Should the PDB accept (some types of) models based on SAS studies ?
- If so, which types of models should be included (and which should not)?
- What are the minimum requirements for these models?
- What are the requirements regarding the supporting experimental data that need to be deposited?
- What validation procedures should be applied in the deposition and annotation process?

Preliminary Recommendations

- Develop an international repository for SAS data
- Standard dictionary required for definition of terms involved in data collection
- Shape and atomistic models based on SAS data should be archived
- Criteria for assessment of the uniqueness and quality of models needs to be defined
- Models derived from diverse hybrid data should be archived
- There is a need for key people involved in the different wwPDB VTFs to come together to discuss what the PDB should be archiving

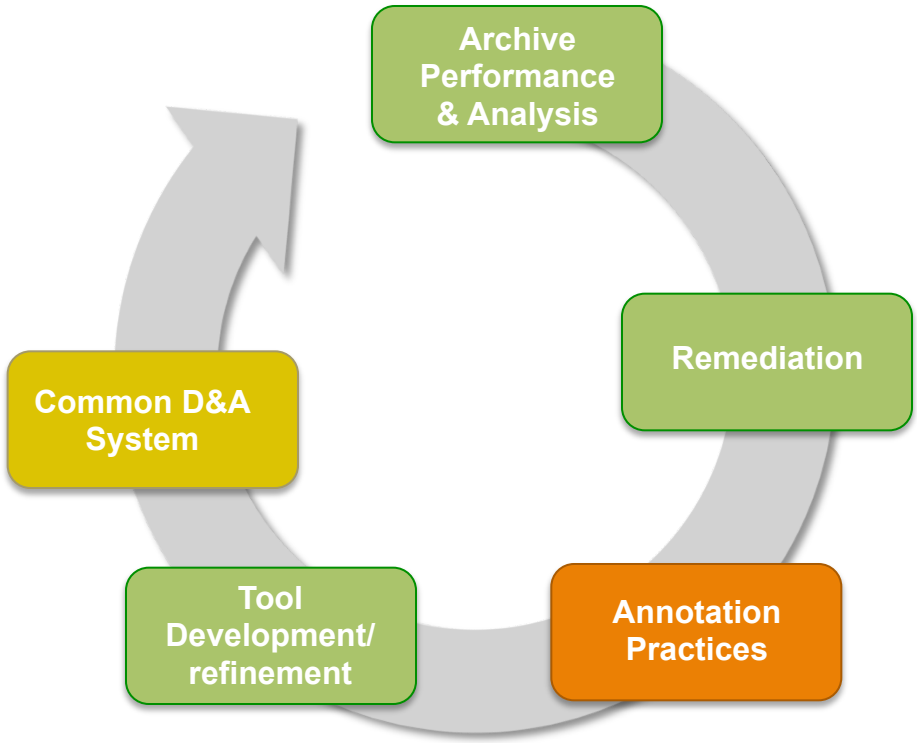
Hybrid Methods

- Meeting planned for 2013
- Include representatives from the different task forces

Remediation

- Informs all processes
- Improves consistency in entry and archive annotation
- Enhances chemistry representation

Better query capability



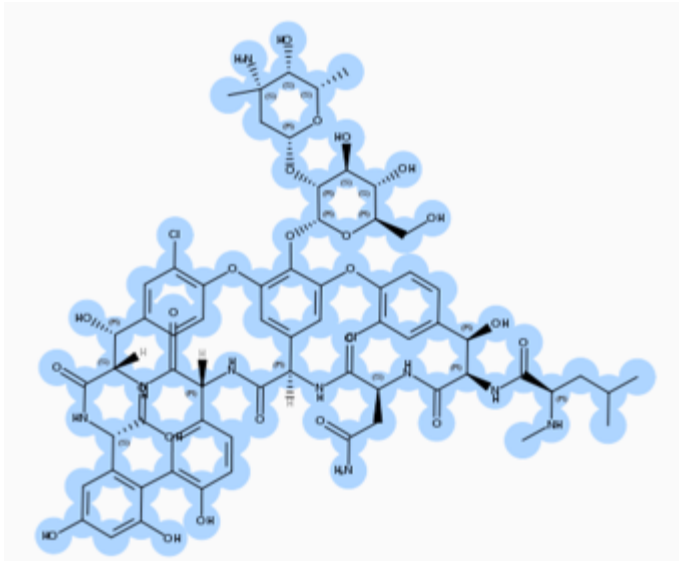
Latest remediation release: July 2011

Better Annotation of Biologically Interesting Molecules

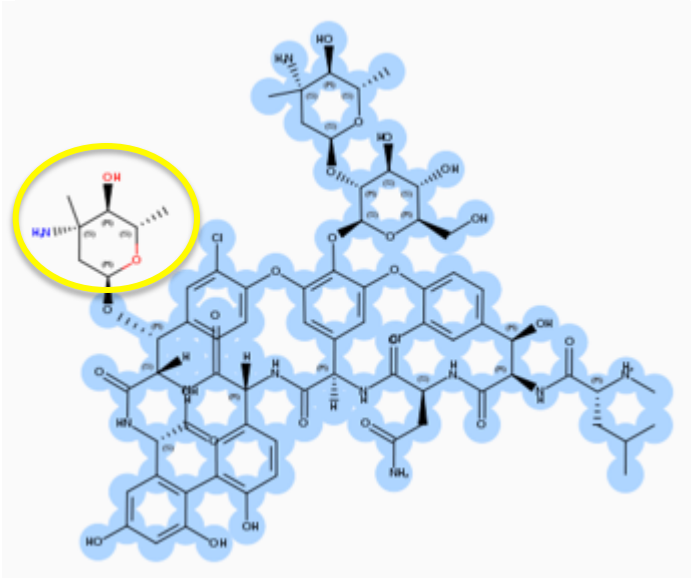
2011 remediation of inhibitors and antibiotics informed the development of an annotation system that supports

- Searches of small molecules and peptides against the new Biologically Interesting molecule Reference Dictionary (BIRD)
- 2D and 3D views
- Comparative analysis of structures
- Building new BIRD definitions
- Use of existing templates to maintain consistency in the data presentation

Target



Hit



Remediation 2013

- Transformation of non-standard crystal frames
- Recalculation of full B factors
- Transformation of dissociated assemblies
- X-ray multiple models

Remediation 2014

- Carbohydrates
  - Data analysis completed
- Protein modifications
  - Data analysis completed
- New amino acids – selenocysteine and pyrrolysine

Carbohydrate Remediation

Issues

- Multiple representations in naming and linking
- Non-standard nomenclature and incomplete linkages
- Representation of branched polymers

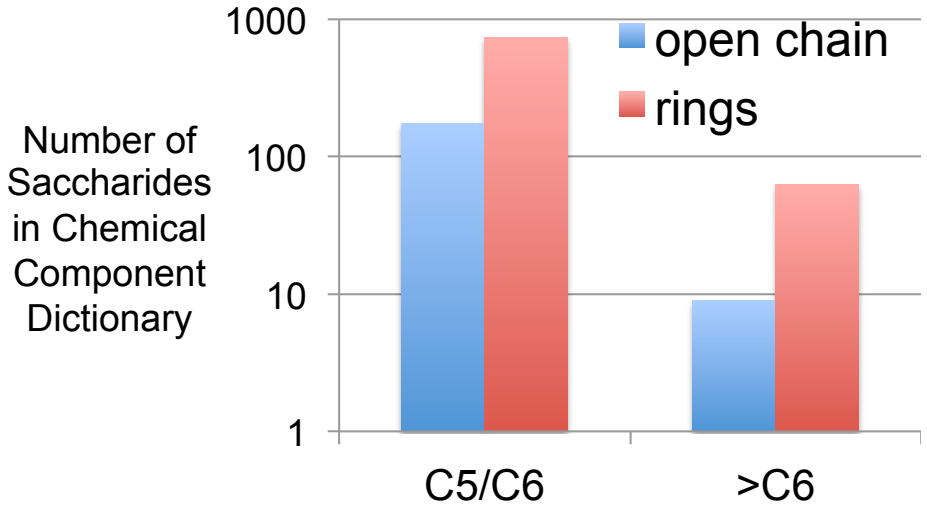
Goal

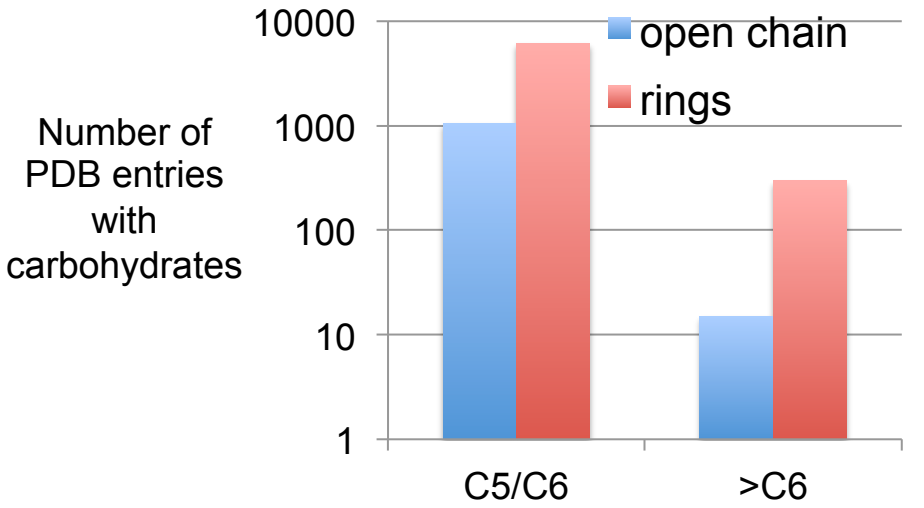
- Represent data consistently within the archive, in agreement with glycobiology community standards
- Enable searches for carbohydrates in the PDB archive

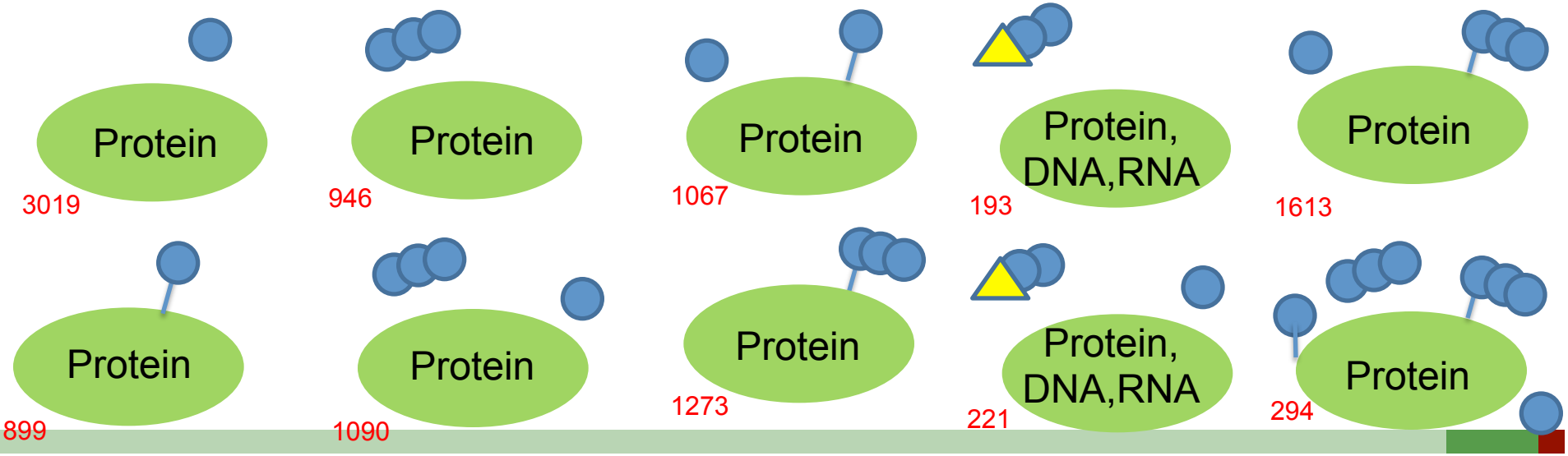
Plan

- Identify and analyze carbohydrate-containing entries
- Create standard representation for branched polymers
- Incorporate standard nomenclature
- Create a strategy for remediation

Carbohydrate Remediation Scope







Carbohydrate Remediation Plan

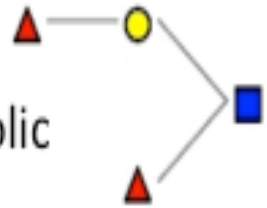
- Represent carbohydrate molecules as polymers of monosaccharides as appropriate
- Adopt glycobiology community standard nomenclature (LINUCS and IUPAC)

LINUCS: [][b-D-GlcpNAc]{[(3+1)][a-L-Fucp]{}[(4+1)][b-D-Galp]{[(2+1)][a-L-Fucp]{}}}

IUPAC

```
a-L-Fucp-(1-2)-b-D-Galp-(1-4)+
                             |
                       b-D-GlcpNAc
                             |
                 a-L-Fucp-(1-3)+
```

Symbolic



PDB ID: 2wmg

Protein Modifications

Issue

- Inconsistent annotation of PMs in the archival files results in the inability to search for these important structures

Goal

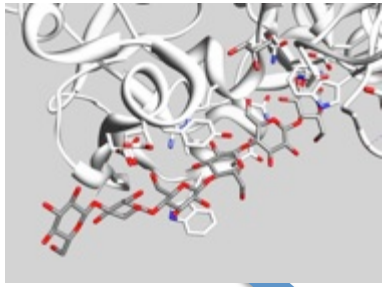
- To identify, classify, and represent all natural protein modifications consistently within the PDB archive and mutually mapped to UniProtKB

Scope

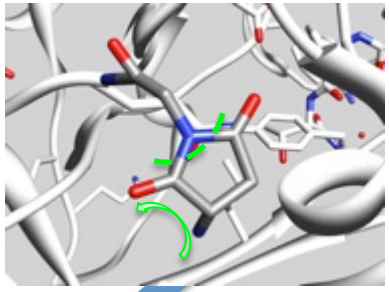
- Chemically modified ribosomal protein including post-translation

Remediated Data

Carbohydrates

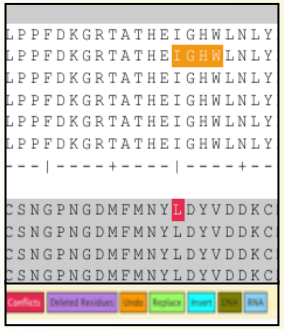


Protein modifications

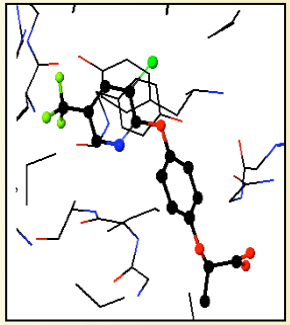


wwPDB Common D&A System

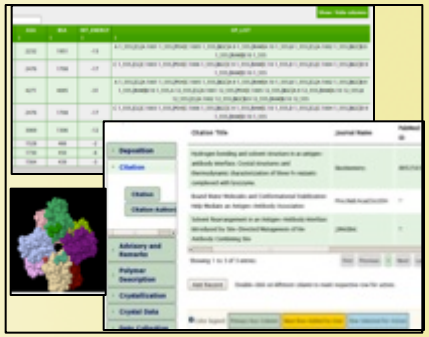
Annotation Pipeline



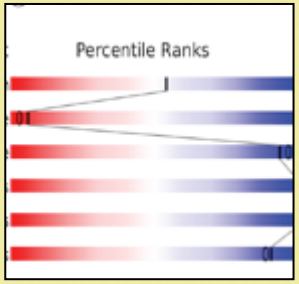
Sequence Processing



Ligand Processing



Manual and Automated Annotation



Validation

Putting it all together

2013 is the start of the big transition

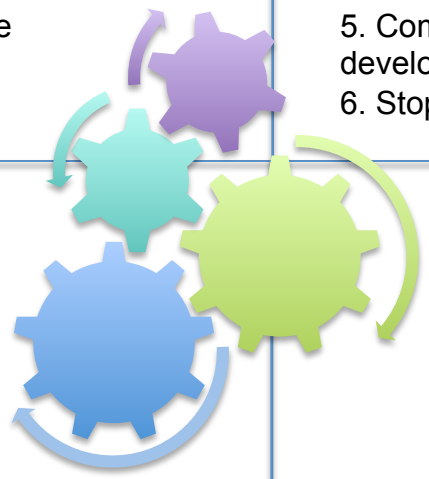
Coordinated Transition

Deposition & Annotation

1. Annotation running at all sites
2. Deposition alpha-2 stage at all sites
3. Deposition functional at all sites
4. Old Deposition system operation as normal
5. Old Deposition system – no new depositions allowed
6. Old annotation system still active
7. Old weekly release
8. New weekly release

Format

1. PDBx Working Group to finalize work
2. Release example new style mmCIF files for ribosomes
3. Start accepting new style format depositions
4. Release server to produce best-effort PDB-format files
5. Community outreach about format developments
6. Stop supplying PDB files in ftp archive



wwPDB Website(s)

1. Announce upcoming changes to the world
2. Unveil new pdb.org website and ftp site
3. Unveil expanded wwpdb.org

Archive

1. Reformat PDBx/mmCIF data files in archive to conform to new style guidelines
2. Remediate “low-hanging fruit”
3. Remediation of carbohydrates and PTMs
4. Introduce versioned entries

wwPDB Organizational Update

Gerard Kleywegt

wwPDB Organisation

- Original wwPDB charter expires in 2013

Effective Date July 1, 2003

By signing this agreement, all members agree to accept the terms contained within.

RCSB	MSD	PDBj
Helen M. Berman	Kim Henrick	Haruki Nakamura
Name	Name	Name
Signature	Signature	Signature
June 30, 2003	June 19. 2003	June 9, 2003
Date	Date	Date

Institutional Official	Institutional Official	Institutional Official
Michael E. Breton	Janet Thornton	Katsuya Nagai
Name	Name	Name
Signature	Signature	Signature
6-30-03	June 19. 2003	June 9, 2003
Date	Date	Date

wwPDB Organisation

- New wwPDB charter 2013
  - Separate principles and implementation details
    - Principles in charter – about wwPDB
      - Updated from 2003 text and circumstances
      - To be signed by PIs and heads of parent institutes
      - Covers 2013-2023, with review possible in 2018
    - Details in appendix – about PDB
      - Updated from 2003 text and circumstances
      - To be signed by PIs
      - Can be modified at any time by mutual agreement

wwPDB Organisation

- wwPDB Advisory Committee
  - "Re-form and reform" in 2013
  - New Terms of Reference to specify:
    - Remit/scope of advice
    - Meetings/reports
    - Membership – who/how long/chair
      - Representatives of wwPDB partners
      - Community representatives
      - Geographical representatives
    - Observers
      - Funding agencies, publishers, *etc*.; wwPDB staff

Discussion Points

- New Terms of Reference
- New wwPDB charter
- Transition strategy