RCSB Protein Data Bank: Overview

RCSB PDB AC

October 2, 2010

Helen M. Berman

Vision

To provide a global resource for the advancement of research and education in biology and medicine by curating, integrating, and disseminating biological macromolecular structural information in the context of function, biological processes, evolution, pathways and disease states.

We will implement standards, and anticipate and develop appropriate technologies to support evolving science.

Structural Views of Biology and Medicine

Mission

Support a resource that is by, for, and of the community by providing

- Leadership in the representation of biological structures derived via experimental methods
- Data in an accurate and timely manner
- Comprehensive, integrated view and unique views of the data

so as to enable scientific innovation and education

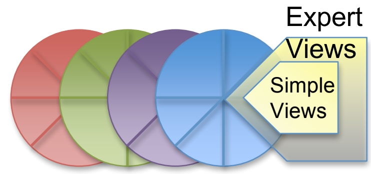
Response to 2009 Major Recommendations

- Develop a coordinated 5-year plan ... balancing costs with benefits, maximizes impact, and establishes productive ties with PDB educator champions
  - Drafted
- Work with scientific journal editors to establish a uniform requirement for author submission of the PDB validation report together with the manuscript describing the structure(s)
  - Reports created, communicating with journals
- Source of biological assembly annotation be identified, and how the biological assembly annotations are decided be documented
  - Source identified on Structure Summary page
  - Process defined in online processing manual

Strategic Plan

Vision: To provide a structural view of biology that frames the access to and understanding of the PDB archive, serving both the scientific and educational communities

Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



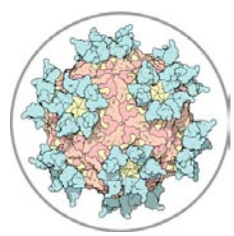
Strategy: Enable new scientific views of the archive, through the RCSB PDB website, that reflect structural biology and support both expert and novice access pathways through categorization of the PDB archive. This strategy will drive all activities including web development, enhanced annotation and outreach design.

The result will be more effective access to the archive content and search functionality.

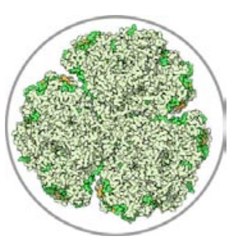
RCSB PDB PROTEIN DATA BANK
Overview
www.pdb.org

Protein Synthesis

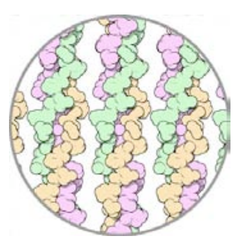
Enzymes



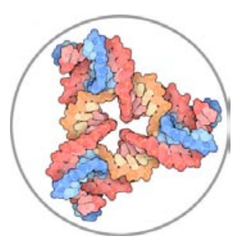
Health & Disease



Biological Energy



Infrastructure & Communication



Biotechnology & Nanotechnology

Data In

- Improved display of large structures
- New validation reports
- Updates on restraint files and EM maps
- ADIT 2.0
- Remediation
- Peptide reference dictionary (PRD)
- wwPDB Validation Task Forces
- NMR: Implementation of chemical shifts
- New format

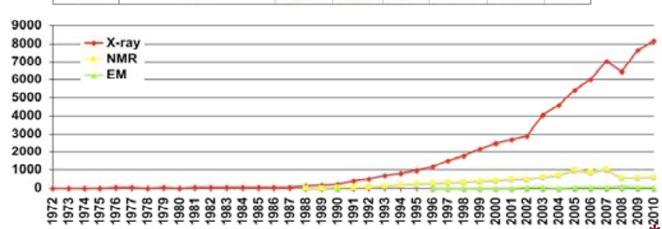
PDB Depositions

By deposition and processing site
*(2010 projected)

Last Updated: 1 Sep 2010

Year	Total Depositions	Deposited To			Processed By		
		RCSB	PDBj	EBI	RCSB	PDBj	EBI
2000	2983	2445	10	528	2297	158	528
2001	3286	2673	118	495	2408	383	495
2002	3563	2769	289	505	2401	657	505
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3083	1613	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	5928 (*8754)	4701	368	859	3766	1303	859
TOTAL	63561	47090	7354	9117	38783	15661	9117

By experimental type
*(2010 projected)



PDB Depositors (1999-2009)

RCSB PDB PROTEIN DATA BANK
Overview
www.pdb.org

wwPDB Projects

- Common Deposition and Annotation Tool
- Task Forces
- Remediation
- wwPDB Foundation

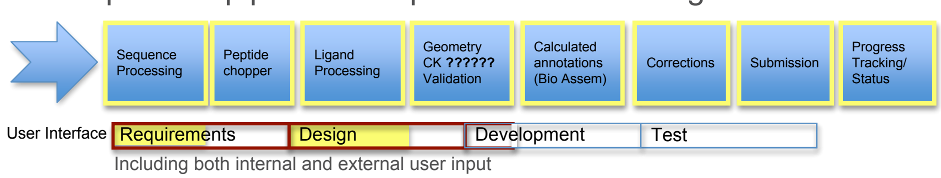
Common Deposition and Annotation Tool

The goal is to implement a set of common deposition and annotation processes and tools that will enable the wwPDB to deliver a resource of increasingly high quality and dependability over the next 10 years.

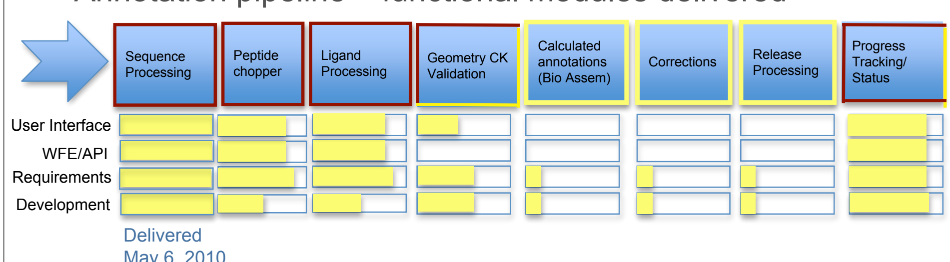
- addresses the increase in complexity and experimental variety of submissions and the increase in deposition throughput
- maximizes the efficiency and effectiveness of data handling and support for the scientific community

2010 Goals

Deposition pipeline – requirements and design



Annotation pipeline – functional modules delivered



wwPDB Validation Task Forces

Method-specific Validation Task Forces have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks.

X-ray

- Workshop on Next Generation Validation Tools for the wwPDB (April 2008)
- White paper nearly complete
- Members

  Paul Adams (Lawrence Berkeley Laboratory), Axel Brünger (Stanford University), Paul Emsley (University of Oxford), Robbie Joosten (University Nijmegen Medical Centre), Gerard Kleywegt (Uppsala University), Thomas Luetteke (Utrecht University), Garib Murshudov (University of York), Zbyszek Otwinowski (UT Southwestern Medical Center at Dallas), Tassos Perrakis (Netherlands Cancer Institute), Randy J. Read (University of Cambridge), Jane Richardson (Duke University), Will Sheffler (University of Washington), Janet Smith (University of Michigan), Ian J. Tickle (Astex Therapeutics Ltd.), Gert Vriend (Radboud Univ Nijmegen Medical Centre)

NMR

- Meeting held September 2009
- Members

  Gaetano Montelione (Co-Chair, Rutgers), Michael Nilges (Co-Chair, Institut Pasteur), Ad Bax (NIH), Wim Vranken (Free University Brussels), Peter Guentert (University Frankfurt), Torsten Herrmann (CNRS/ENS Lyon), Jane Richardson (Duke University), Charles Schwieters (NIH), Geerten Vuister (Radboud University), David Wishart (University of Alberta).

wwPDB Validation Task Forces

CryoEM

- Meeting September 2010
- Members

  Richard Henderson (Map Chair, Cambridge University), Andrej Sali (Models Chair, UCSF), Kenneth Downing (LBL), Edward Egelman (U Virginia), Joachim Frank (Columbia), Niko Grigorieff (Brandeis), Wen Jiang (Purdue), Steven Ludtke (Baylor), Ron Milligan (Scripps), Pawel A. Penczek (UT Houston Medical School), Peter Rosenthal (National Institute for Medical Research), Michael G. Rossmann (Purdue), Michael Schmid (Baylor), Gunnar Schroeder (Forschungszentrum Juelich), Alasdair Steven (NIAMSD), Florence Tama (University of Arizona), Maya Topf (Birbeck, University of London), Willy Wriggers (DE Shaw Research)

Small Angle Scattering

- Members

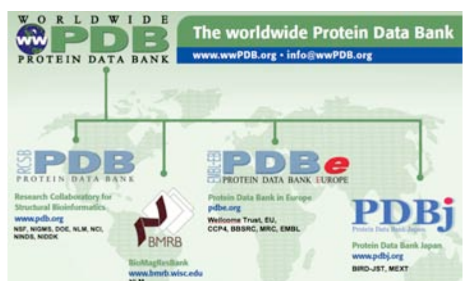
  Jill Trewhella (University of Sydney), Dmitri Svergun (EMBL Hamburg), Andrej Sali (UCSF), Mamoru Sato (Yokohama City University), John Tainer (Scripps)

Common D&A Tool and Remediation Are Collaborative wwPDB Projects

Funding for wwPDB curation and distribution of the archive comes from grants to the individual wwPDB member groups

Foundation was established to fundraise for wwPDB education and outreach activities

- PDB 40

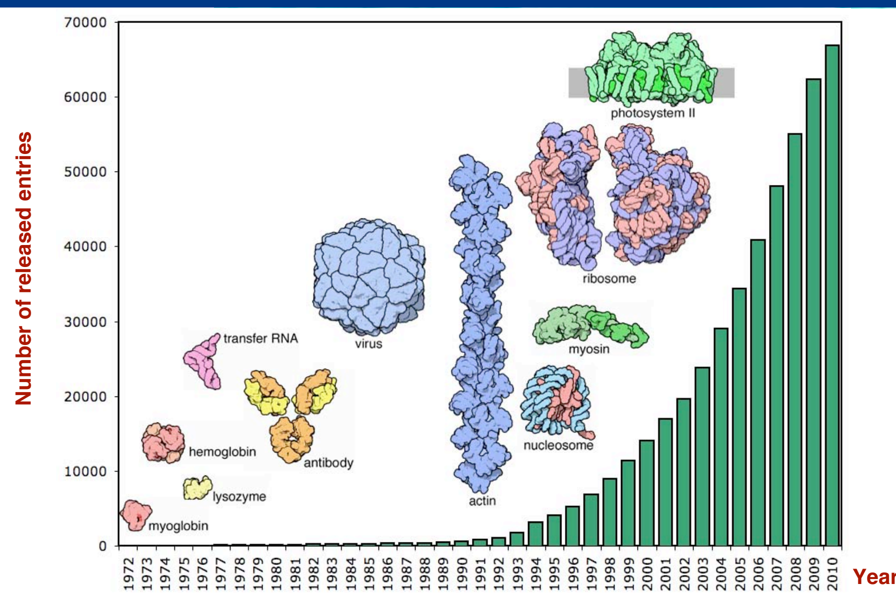


Worldwide Protein Data Bank Foundation

Data Out

- New home page layout and view
- Web widgets
- Customizable home page
- Educational view for molecule of the month
- Ligand summary page
- Chemical components search
- Customizable query results
- Query refinement through drill-down
- Improved tabular reports
- Pair-wise sequence and structure comparison
- Improved structure visualization
- Improved performance

RCSB PDB PROTEIN DATA BANK
Overview
www.pdb.org



PDB FTP & Rsync Traffic
(July 2009 – June 2010)

RCSB PDB
173,416,704
data downloads

PDBe
32,344,547
data downloads

PDBj
14,053,071
data downloads

Outreach and Impact

Goals

- RCSB PDB resource should meet its mission in the interest of science, medicine and education
- RCSB PDB is defined by, designed for, and owned by the communities it serves

Communities

- Biologists
- Other scientists
- Students and educators (all levels)
- Media writers, illustrators, textbook authors
- General public

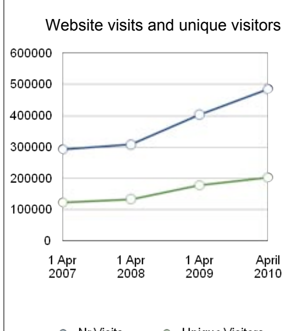
Current and Expanding Initiatives

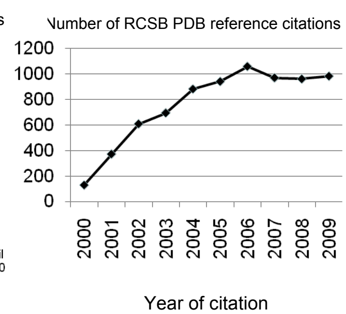
- Electronic help desks, discussion groups
  - New tracking system
- Demonstrations and presentations at professional meetings
  - New meetings, improved materials and assessment systems
- Personal interactions
- Workshops and posters
- Surveys
- PDB 40

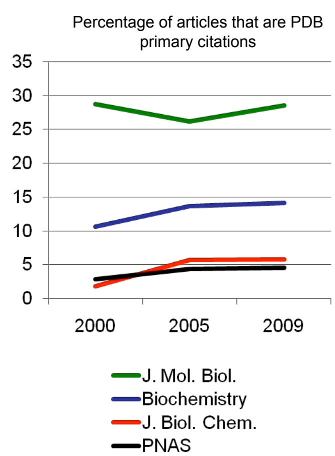
Biophysical Society Meeting, 2010

PDB Depositors' Lunch, ACA 2010

Increasing Impact







Management and Oversight

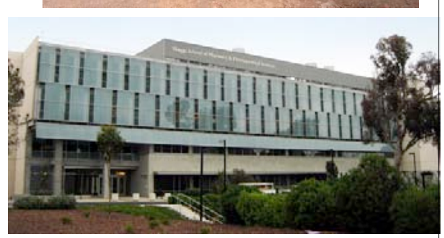
- Director, Helen M. Berman
  - Overall direction of RCSB PDB
  - Direction of Rutgers site
- Deputy Director, Martha Quesada
  - Coordination of all projects across the RCSB PDB
  - Facilitation of wwPDB initiatives
- Associate Director, Philip E. Bourne
  - Direction of UCSD site
- PDBAC and wwPDBAC
  - Stephen K. Burley, Chair

Institutional Commitments

Rutgers

- Center for Integrated Proteomics Research (CIPR)
  - Intellectual home
  - New building
  - New hires

UCSD



- Skaggs School of Pharmacy and Pharmaceutical Sciences
  - New collaborators (*e.g.*, Ruben Abagyan, Michael Gilson)

PDB-Related Funding

Project	Agency	Period	Award
RCSB PDB PROTEIN DATA BANK	NSF	03/01/09-2/28/14	$28 million
PSI nature StructuralBiologyKnowledgebase	NIH	07/01/10-06/30/15	$12.5 million
EMDataBank.org Unified Data Resource for CryoEM	NIH	08/15/07-05/31/12 PI Wah Chiu	$2 million

RCSB PDB Mid-cycle Review

Rutgers, November 1-2, 2010

Selected Topics

- Sustainability
- How do we measure our impact on
  - education
  - non-structural biology
- International relations
- D&A tool development

RCSB PDB & Friends, 2009

RCSB PDB PROTEIN DATA BANK
Overview
www.pdb.org

Agenda

- Introduction & Overview — Helen Berman
- Data In — Jasmine Young
- Common D&A Tool — Martha Quesada, John Westbrook
- Data Out — Phil Bourne
- Outreach and Impact — Christine Zardecki, Andreas Prlic
- Executive Session
- General Discussion

Data In: Deposition, Annotation and Remediation

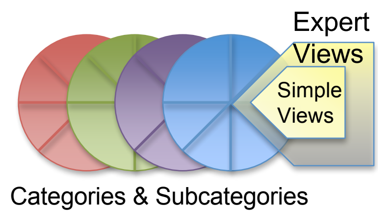
RCSB PDB AC

October 2, 2010

Jasmine Young

RCSB PDB PROTEIN DATA BANK
Data In
www.pdb.org

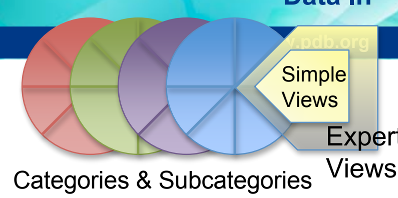
Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



Strategy: Enable new scientific views of the archive, through the RCSB PDB website, that reflect structural biology and support both expert and novice access pathways through categorization of the PDB archive. This strategy will drive all activities including web development, enhanced annotation and outreach design.

The result will be more effective access to the archive content and search functionality.

Annotation Goal



- *Routine* annotation and validation tasks fully automated
- Principal annotation activities shifting from routine data handling to expanded expert annotation
  - Integration with other biological data
  - Expanding and maintaining data uniformity
  - Support larger and more complex biological molecules, and new methods
  - Extending and representing new content, *e.g.* functional annotation (categories)

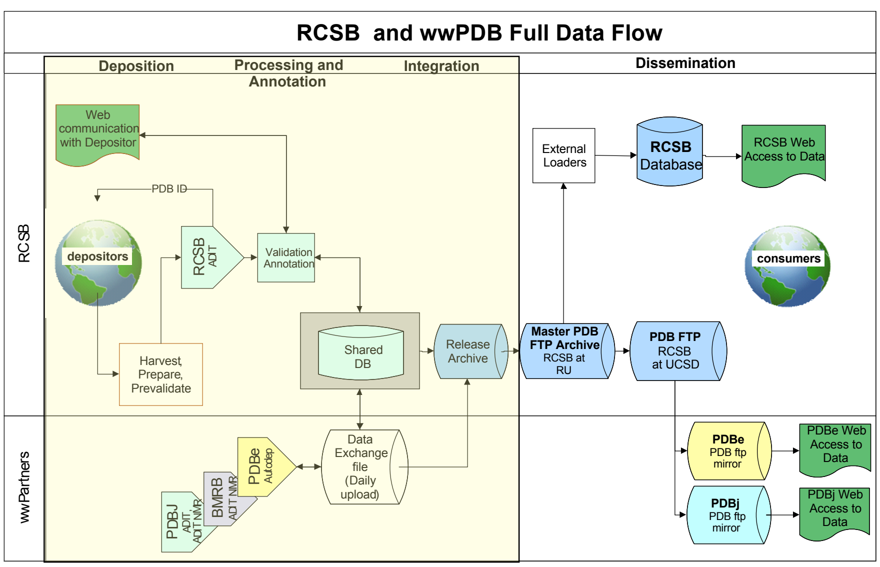
Depositor locations

Download locations

- RCSB PDB
- PDBe
- PDBj

RCSB PDB PROTEIN DATA BANK

WORLDWIDE PDB PROTEIN DATA BANK



Deposition Statistics

Month	Deposited to RCSB	Deposited to PDBj	Deposited to PDBe	Processed by RCSB	Processed by PDBj	Processed by PDBe	Total deposition
Jul 2009	568	37	88	429	176	88	693
Aug 2009	507	48	79	393	162	79	634
Sep 2009	613	41	105	474	180	105	759
Oct 2009	596	71	103	456	211	103	770
Nov 2009	528	52	92	399	181	92	672
Dec 2009	501	43	68	348	196	68	612
Jan 2010	538	55	106	424	169	106	699
Feb 2010	488	51	109	347	192	109	648
Mar 2010	613	39	121	485	167	121	773
Apr 2010	578	49	90	454	173	90	717
May 2010	625	49	99	512	162	99	773
Jun 2010	705	27	90	541	191	90	822
Total	6860	562	1150	5262	2160	1150	8572
	80%	6%	13%	61%	25%	13%	

Data In

www.pdb.org

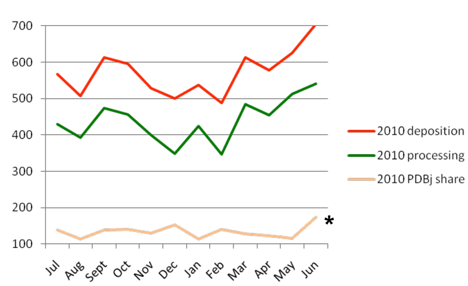
Deposition and Annotation

RCSB PDB and PDBj

Entries deposited at RCSB PDB

Number of depositions increased in 2010

Entries processed at RCSB PDB



*PDBj processes some entries deposited at the RCSB PDB

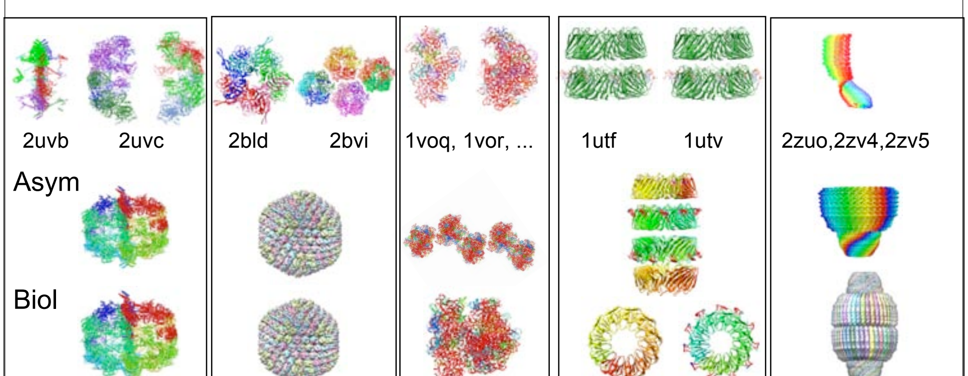
Data In: Recently Completed Projects

- Improved display of large structures
- New validation reports
- Update on restraint files and EM maps
- ADIT 2.0
- Supported new methods

Improved Display of Large Structures

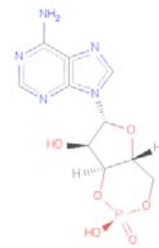
(December 2009)

Complete biological assembly views available for structures that are split across multiple PDB coordinate files



New Validation Reports (May 2010)

- High level summary
- PDF format for authors to easily send to journal reviewers
- Geometry validation
  - Atom clashes, peptide linkage, covalent geometry
- Sequence validation
- Biological assembly
- Ligand chemistry
- Structure factor validation



R-factors	
R-factor (Author reported)	0.150
R-factor (Calculated by SFCHECK, V7.02.4)	0.212
R-factor (Calculated by REFMAC, V5.5.0109)	0.1960
Free R-factor (Author reported)	0.188
Free R-factor (Calculated by SFCHECK, V7.02.4)	0.236
Free R-factor (Calculated by REFMAC, V5.5.0109)	0.2200

Structure quality	
Average Real space R-factor (Deviation) (Calculated by SFCHECK, V7.02.4)	0.0767
Average Real space R-factor (Deviation) (Calculated by MAPMAN, V7.8.5)	0.1007
Average Real-space correlation coefficient (Deviation) (Calculated by SFCHECK, V7.02.4)	0.9858
Average Real-space correlation coefficient (Deviation) (Calculated by MAPMAN, V7.8.5)	0.963
Average Occupancy-weighted avg temperature factor (Deviation)	35.25

Wilson statistics (PHENIX, V1.6-289)	
Wilson B-factor	31.65
Wilson Scale	0.12

NMR Restraint Files and EM Maps

NMR Restraint Files (version 2) (June 2010)

- BMRB in collaboration with PDBe and CMBI/IMM
- NMR-STAR 3.1 format
- Contain current PDB atom nomenclature
- Provide accurate atom-level correspondences to the NMR model coordinate files in the current archive
- Original restraint files (Version 1) remain on the site and will continue to be updated regularly

EM Maps (+730 maps) (September 2010)

- Map headers: corrected voxel size, density statistical values (min, max, avg, rmsd)
- Maps have been repositioned to superimpose over corresponding fitted PDB models (~50)

ADIT 2.0 (July 2010)

- Deployed in July
  - Designed to improve data quality and processing efficiency
- Validation mandatory
  - Checks file format with suggestions for solutions
  - Checks for consistency between sequence and coordinates
- Allows easier organization of sequence information
- Simplifies entering author, title, and citation information

Support New Methods

Expanded dictionary to support

- EM
- SAX (preliminary dictionary)
- Joint refinement (e.g., Neutron/Xray diffraction)

Data In: Ongoing Projects

- Remediation
- Peptide Reference Dictionary (PRD)
- wwPDB Validation Task Force
- NMR: Implementation of chemical shifts
- New format

Remediation

- Biological assemblies
  - PISA vs PQS
  - Missing PISA
- Residual B factors
- Peptide inhibitors and antibiotics

Expected Rollout Q1 2011

Biological Assemblies

Problem

- Inconsistent and missing computational annotation of biological assemblies

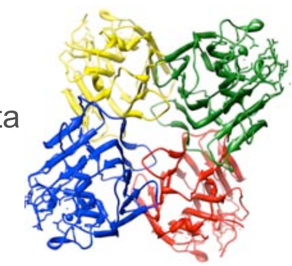
Approach

- Compared curated PQS generated assemblies with PISA generated assemblies and preferentially included the PQS data in entries with missing data

Result

- 5800 entries updated with PISA and/or PQS

Author-deposited



PISA-generated

Crystal strucure of a lectin from *Canavalia gladiata* (CGL) in complex with man1-2man-Ome. Bezerra, Oliveira, Moreno, de Souza, da Rocha, Benevides, Delatorre, de Azevedo, Cavada (2007) *J.Struct.Biol*. 160: 168-176

2OVU

Residual B Factors

Problem

- Inconsistent deposition of temperature factor data in PDB ATOM records for 7629 entries refined using TLS with REFMAC

Approach

- Analyzed these entries by back calculation of new isotropic B-values, and compared refinement statistics before and after correction
- Closer reproduction of reported statistics used to assign full or residual B-value

Result

- 6296 entries labeled as LIKELY containing residual B-values.154 entries determined to contain full B-values based on other information in the deposited entry
- 1179 entries require further analysis

Residual B Factors – Format Details

Remediated data files for the 6296 entries identified as likely containing residual B-values will include the following new records

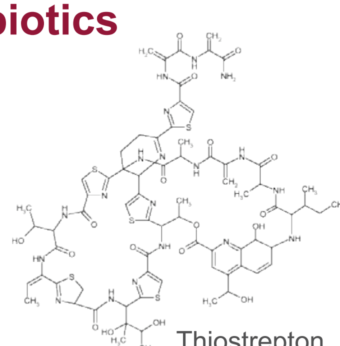
- PDB FORMAT

```
REMARK   3  B VALUES
REMARK   3  B VALUE TYPE : LIKELY RESIDUAL
```

- PDBx/mmCIF and PDBML

  In the REFINE category, a new item PDBX_ADP_TYPE will be added and assigned the value 'LIKELY RESIDUAL'

Peptide Inhibitors and Antibiotics



Thiostrepton

Contain

- Complicated chemistry
- Important functions
- 300 polymeric antibiotics
- peptide inhibitors: 420 single component, 450 polymeric

Challenges

- Non-standard amino acid, nucleotides or other chemical groups in sequence
- Non-linear (cyclic or branched) sequences
- Microheterogeneity
- Non-uniform annotation of the same molecule in different PDB entries
- Lack of annotation regarding the source and function of these molecules

Peptide Inhibitors and Antibiotics: Solutions

Analysis and classification

- Identify antibiotics and inhibitors and group them into polymeric molecules or single molecules

Dictionary updates

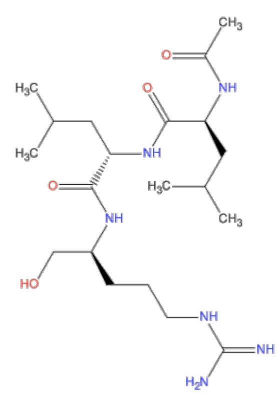
- Build single chemical components for appropriate cases
- Update dictionary with source, function and other details

Remediation and future processing

- Revise coordinate files to present chemistry in either sequence or single molecule form
- Create a Peptide Reference Dictionary (PRD)
- Establish rules and procedures to make new annotations consistent

Status

- Chemistry corrected
- Inhibitor annotation completed
- Load testing to be done
- Annotation guideline documentation completed
- Annotation training ongoing
- To be released January 2011



ACE LEU LEU ARG

Leupeptin

Peptide Reference Dictionary (PRD)

An supplementary information resource about peptide inhibitors and antibiotics:

- Provides help in consistent PDB data processing
- General resource for community
- Integrate with other biological data: source, physical, chemical, functional, and other commercial information
- Dual presentation: sequence and SMILES strings
- Links to CAS, KEGG, ChEBI, Norine, UniProt, *etc*.
- Functions extracted from these resources as well as from primary citations
- mmCIF files have been created and checked for PRD
- Search interface to be done

wwPDB Validation Task Forces

Method-specific Validation Task Forces have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks.

X-ray

- Workshop held April 2008
- White paper nearly complete

NMR

- Meeting held September 2009

CryoEM

- Meeting September 2010

Small Angle Scattering

NMR: Implementation of Chemical Shifts (CS)

- Installation, testing and training on CS deposition
  - ADIT-NMR: check format and sanity check at deposition
  - Substitute explicit atoms for pseudo-atoms
- Testing and training on CS data processing
  - Maintain nomenclature correspondence during annotation
  - Data files to be transferred to BMRB for further annotation
- To be deployed early December 2010
- PDB will release CS files in NMR-STAR format along with coordinate data files

New Format

- PDB format defined in 1970s
  - FORTRAN (column-oriented)
  - "Small" molecules
- Limitations
  - Max 62 chains (and that's stretching it)
  - Max 99,999 atoms (5 ribosomes in ASU- 10 PDB entries!)
  - No bond orders specified for ligands
  - Meta-data specification cumbersome and inflexible

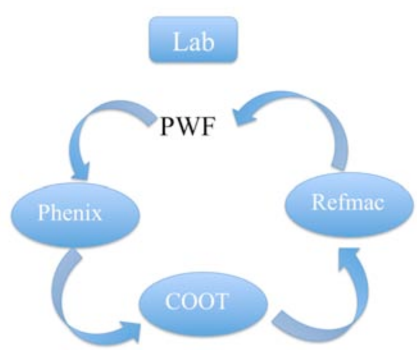
New Format

- wwPDB archival/exchange format is PDBx
  - No uptake in community despite libraries
  - Good for machines, not so good for humans
- Pragmatic solution needed
  - Specify new working format for data exchange between software used in labs
    - Refinement, model-building, graphics, validation, …
  - Define new "human-readable report" content and format for meta-data

New Format: PDB Working Format (PWF)

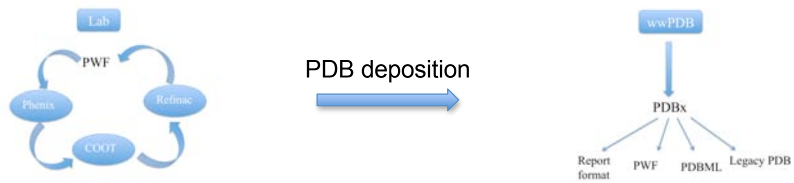
- Support large and complex structures
- Support for new and hybrid experiments
- Addresses PDB format issues

  e.g. Fixed character width and text REMARKs



PDB Working Format (PWF)

- Preserve simple style and readability of PDB format
- Provide extensible framework for capturing larger systems and information from multiple experimental methods
- Best combine of both worlds
- One master archival format (PDBx)
- FTP will contain PDBx, PWF, report format and PDBML files




```
#!BEGIN_TABLE_DECLARATION atom_site
#!BEGIN_COLUMN_LIST 17
_atom_site.group_PDB
_atom_site.id
_atom_site.auth_atom_id
_atom_site.label_alt_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.type_symbol
_atom_site.pdbx_formal_charge
_atom_site.pdbx_tls_group_id
_atom_site.pdbx_PDB_model_num
#!END_COLUMN_LIST
#!FORMAT_STRING_C VER1 ROW (%-6s %8d %-10s %-2s %-10s %-10s %6d %-2s %10.3f %10.3f %10.3f %5.3f %8.3f %-2s %3d %3d %4d\n)
#!FORMAT_STRING_F77 VER1 ROW (A6,1X,I8,1X,A10,1X,A2,1X,A10,1X,A10,1X,I6,1X,A2,1X,F10.3,1X,F10.3,1X,F10.3,1X,F5.3,1X,F8.3,1X,A2,1X,I3,1X,I3,1X,I4)
#!END_TABLE_DECLARATION

#!BEGIN_TABLE_DATA atom_site
ATOM          1 N          _  MET        0              1 _     -38.945    118.157    160.952 1.000  156.580 N    0   1    1
ATOM          2 CA         _  MET        0              1 _     -40.032    119.180    160.981 1.000  156.580 C    0   1    1
ATOM          3 C          _  MET        0              1 _     -41.382    118.537    161.236 1.000  156.580 C    0   1    1
ATOM          4 O          _  MET        0              1 _     -42.016    118.788    162.262 1.000  199.790 O    0   1    1
ATOM          5 CB         _  MET        0              1 _     -40.089    119.956    159.655 1.000   98.680 C    0   1    1
ATOM          6 N          _  ALA        0              2 _     -41.813    117.704    160.294 1.000  146.870 N    0   1    1
ATOM          7 CA         _  ALA        0              2 _     -43.109    117.046    160.389 1.000  146.870 C    0   1    1
ATOM          8 C          _  ALA        0              2 _     -44.136    118.150    160.154 1.000  146.870 C    0   1    1
ATOM          9 O          _  ALA        0              2 _     -45.109    118.286    160.897 1.000  199.790 O    0   1    1
ATOM         10 CB         _  ALA        0              2 _     -43.290    116.422    161.778 1.000   37.370 C    0   1    1
ATOM         11 N          _  HIS        0              3 _     -43.898    118.937    159.107 1.000  124.100 N    0   1    1
```

PDB-like stylized PDBx

New Format

Timeline

- First written draft of well-defined PWF was written June 2010
- Bring in key software developers in 2010
  - Coot, Phenix, CNS, Refmac, Buster, Shelx, CCP4
  - ARIA, CYANA, UNIO, XPLOR-NIH
  - Visualization, computational biology, bioinformatics, commercial
- Finalize written format Q2 2011
- Implementation Q1 2012

D&A Project

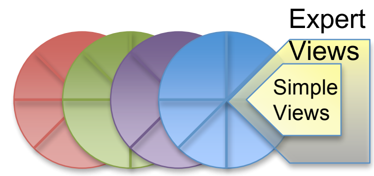
wwPDB Common Deposition & Annotation (D&A) Tool

RCSB PDB AC

October 2, 2010

Martha Quesada, John Westbrook
for the wwPDB D&A Project Team

Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



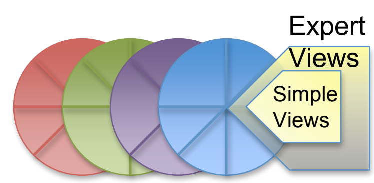
Categories & Subcategories

Strategy: Enable new scientific views of the archive, through the RCSB PDB website, that reflect structural biology and support both expert and novice access pathways through categorization of the PDB archive. This strategy will drive all activities including web development, enhanced annotation and outreach design.

The result will be more effective access to the archive content and search functionality.

RCSB PDB PROTEIN DATA BANK
D&A Project
www.pdb.org

Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



Categories & Subcategories

Annotation Goals

- *Routine* annotation and validation tasks fully automated
- Principal annotation activities shifting from routine data handling to expanded expert annotation
  - Integration with other biological data
  - Expanding and maintaining data uniformity
  - Support larger and more complex biological molecules, and new methods
  - Extending and representing new content, *e.g*. functional annotation (categories)

Multi-Disciplinary Project Team Representing All Four wwPDB Sites

Experts in:

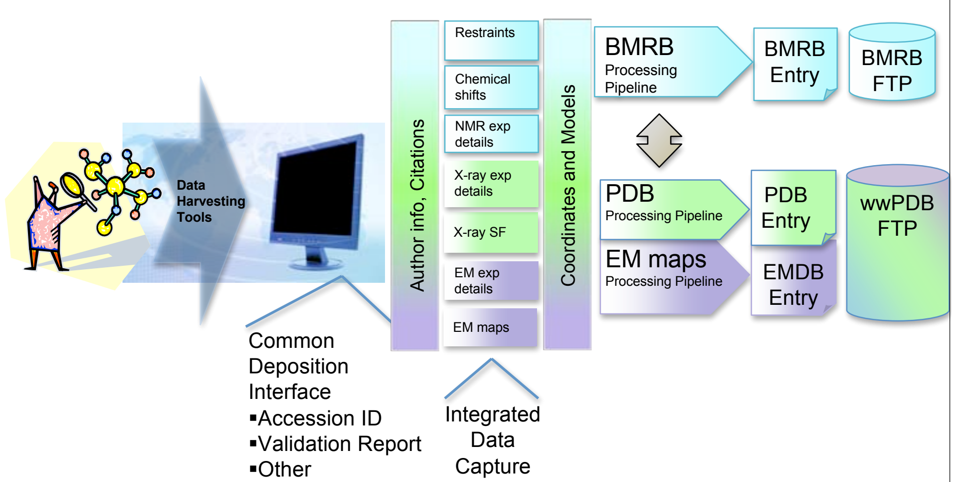
- Content - annotators
- Functional applications - scientific programmers
- Graphical user interfaces
- Databases
- Application programming interfaces
- Workflow engine design
- Data sharing architecture

wwPDB Common D&A Project

Project Drivers: Scope Growth, Quality and Efficiency

- Meeting the evolving data needs of our user community
  - Larger and more complex biological molecules
  - New methods
  - Expanded annotation
  - Improved quality – new validation strategies
  - Larger throughput – automation and validation of routine submissions
- Recognition of the need to “pool” our resources to meet the challenges before us

The Operational Vision



Project Goal

The goal is to implement a set of common deposition and annotation processes and tools that will enable the wwPDB to deliver a resource of increasingly high quality and dependability over the next 10 years.

The tools and processes will:

- Address the increase in complexity and experimental variety of submissions and the increase in deposition throughput
- Maximize the efficiency and effectiveness of data handling
- Provide for higher quality and completeness of submissions and annotation through improved use of graphical interfaces

What's in it for...

Depositors

- Interactive and informative deposition interface
- Value-added validation input and annotation during deposition
- Faster processing

Annotators

- Improve efficiency, freeing time for more advanced annotation
  - Improved quality early in the process
  - Automation of appropriate processing steps
  - Best-of-breed tools
  - Expanded functionality
- Enable system evolution through modularity

Data users

- Higher quality archive

2010 Goals

Deposition pipeline – requirements and design

Sequence Processing | Peptide chopper | Ligand Processing | Validation | Calculated annotations (Bio Assembly) | Corrections | Submission | Progress Tracking/ Status

User Interface: Requirements | Design | Development | Test

Including both internal and external user input

Annotation pipeline – functional modules delivered

Sequence Processing | Peptide chopper | Ligand Processing | Validation | Calculated annotations (Bio Assembly) | Corrections | Release Processing | Progress Tracking/ Status

User Interface

WFE/API

Requirements

Development

D&A Project

Sequence Processing Overview

Author-provided

```
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: MYOGLOBIN;
SOURCE MOL_ID: 1;
SOURCE 2 ORGANISM_SCIENTIFIC: PHYSETER CATODON
SOURCE 4 ORGANISM_TAXID: 9755
DBREF 1MBN A 1 153 UNP P02185 MYG_PHYCA 1 153

SEQRES   1 A  153  VAL LEU SER GLU GLY GLU TRP GLN LEU VAL LEU HIS VAL
SEQRES   2 A  153  TRP ALA LYS VAL GLU ALA ASP VAL ALA GLY HIS GLY GLN
SEQRES   3 A  153  ASP ILE LEU ILE ARG LEU PHE LYS SER HIS PRO GLU THR
SEQRES   4 A  153  LEU GLU LYS PHE ASP ARG PHE LYS HIS LEU LYS THR GLU
SEQRES   5 A  153  ALA GLU MET LYS ALA SER GLU ASP LEU LYS LYS HIS GLY
SEQRES   6 A  153  VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU LYS LYS
SEQRES   7 A  153  LYS GLY HIS HIS GLU ALA GLU LEU             GLN
SEQRES   8 A  153  SER HIS ALA THR LYS HIS LYS
SEQRES   9 A  153  GLU PHE ILE SER GLU ALA ILE
SEQRES  10 A  153  ARG HIS PRO GLY ASP PHE GLY ALA ASP ALA GLN GLY ALA
SEQRES  11 A  153  MET ASN LYS ALA LEU GLU LEU PHE ARG     ASP ILE ALA
SEQRES  12 A  153  ALA LYS TYR LYS GLU LEU GLY TYR GLN GL

ATOM      1  N   VAL A   1      -2.900  17.600  15.500        0.00           N
ATOM      2  CA  VAL A   1      -3.600  16.400  15.300        0.00           C
ATOM      3  C   VAL A   1      -3.000  15.300  16.200        0.00           C
ATOM      4  O   VAL A   1      -3.700  14.700  17.000        0.00           O
ATOM      5  CB  VAL A   1      -3.500  16.000  13.800        0.00           C
ATOM      6  CG1 VAL A   1      -2.100  15.700                              C
ATOM      7  CG2 VAL A   1      -4.600  14.900                              C
ATOM      8  N   LEU A   2      -1.700                                      N
ATOM      9  CA  LEU A   2      -0.900  14.100                              C
ATOM     10  C   LEU A   2      -1.000  13.900                              C
ATOM     11  O   LEU A   2      -0.900  14.900  19.000  1.00  0.00           O
```

Taxonomy

Sequence

Atom-site records

Cross-check with

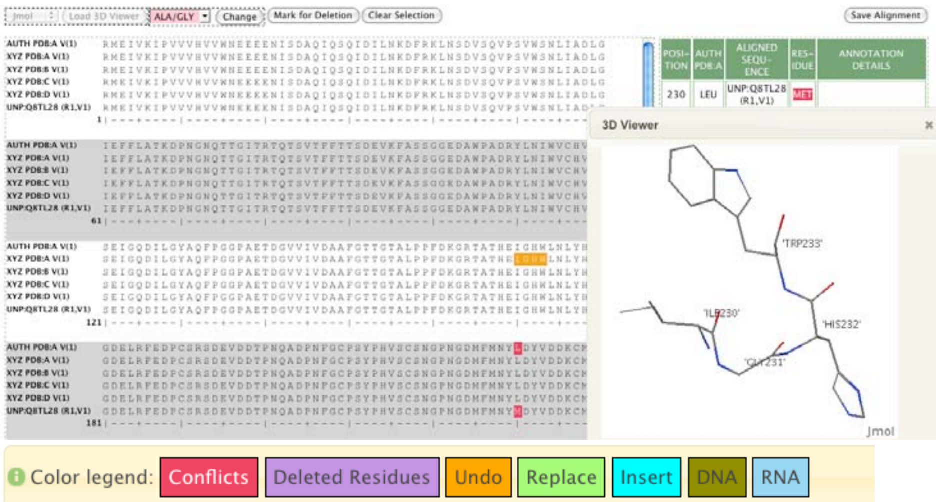
Taxonomy

UniProt

Sequence

D&A Project

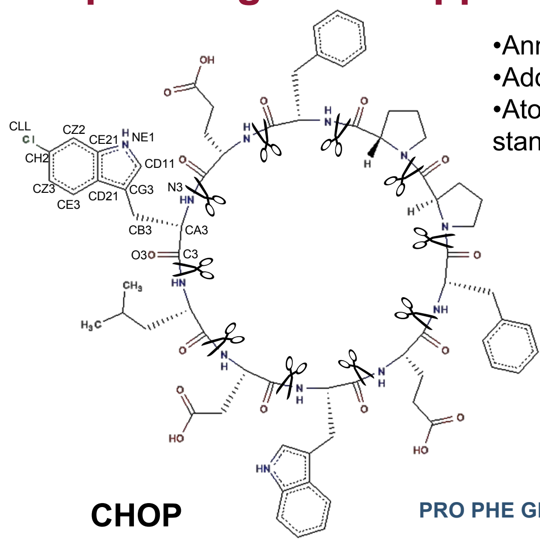
Annotator Integrated View



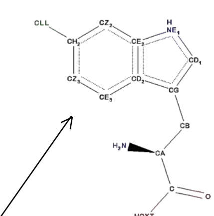
D&A Project

Peptide Ligand Chopper

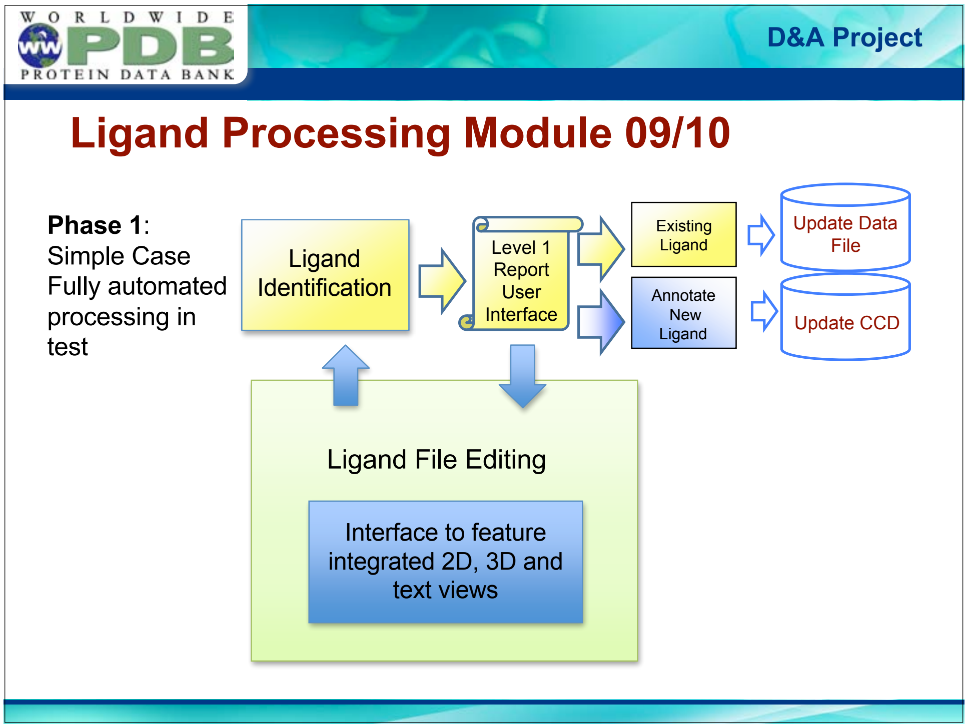
- Annotator directed bond breaks
- Add leaving groups (ie. -OH, -H, -Cl)
- Atom naming and numbering standardized

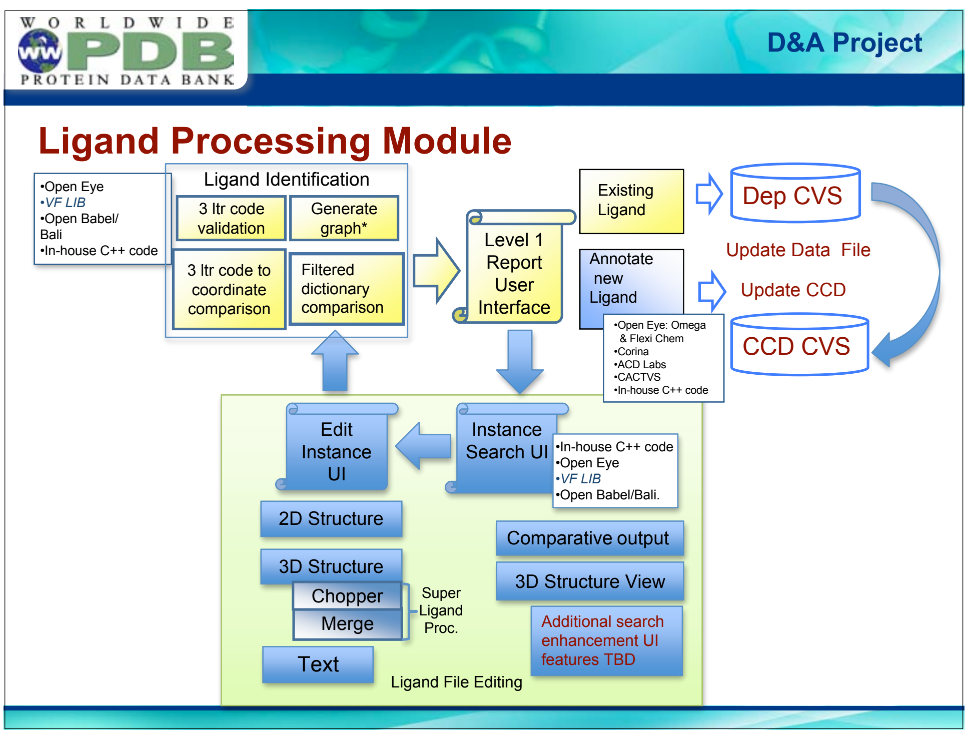


CHOP



PRO PHE GLU 6CW LEU ASP TRP GLU PHE DPR


WORLDWIDE PDB PROTEIN DATA BANK
D&A Project
Ligand Processing Module 09/10
Phase 1:
Simple Case
Fully automated processing in test
Ligand Identification
Level 1 Report User Interface
Existing Ligand
Annotate New Ligand
Update Data File
Update CCD
Ligand File Editing
Interface to feature integrated 2D, 3D and text views


WORLDWIDE PDB PROTEIN DATA BANK
D&A Project
Ligand Processing Module
•Open Eye
•VF LIB
•Open Babel/ Bali
•In-house C++ code
Ligand Identification
3 ltr code validation
Generate graph*
3 ltr code to coordinate comparison
Filtered dictionary comparison
Level 1 Report User Interface
Existing Ligand
Annotate new Ligand
•Open Eye: Omega & Flexi Chem
•Corina
•ACD Labs
•CACTVS
•In-house C++ code
Dep CVS
Update Data File
Update CCD
CCD CVS
Edit Instance UI
Instance Search UI
•In-house C++ code
•Open Eye
•VF LIB
•Open Babel/Bali.
2D Structure
3D Structure
Chopper
Merge
Super Ligand Proc.
Text
Comparative output
3D Structure View
Additional search enhancement UI features TBD
Ligand File Editing

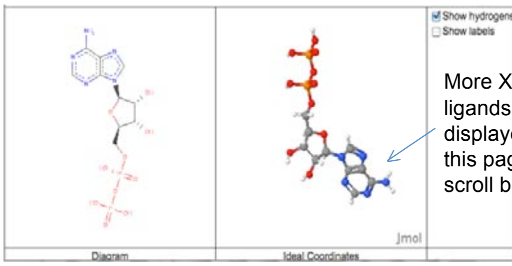
Ligand Editor Mock Up

Deposition id: D_012345

Ligand id: XYP_B_287
Name: [(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-3,4-dihydroxy-oxolan-2-yl]methyl phosphono hydrogen phosphate
Formula: C10 H15 N5 O10 P2
Formal Charge: 0

Undo | Save

ID	Instance	Status	Select
XYP	A503	CLOSE MATCH	
XYP	A504	NO MATCH	



More XYP ligands to be displayed on this page with scroll bar

Search results for Ligand instances

XYP_B_287	ID	Score (%)	Select for comparison
	0AI	98	
	1NA	97	
	5AX	96	
	A2G	96	

Create Ligand | Split/Merge | Run Search

Input new parameters here

Input your notes here

WORLDWIDE PDB PROTEIN DATA BANK
D&A Project

Ligand Validation

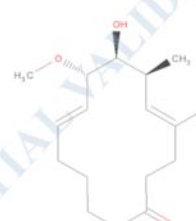
Ligand Chemistry

Ligand chemistry has been checked against the Chemical Component Dictionary. The following is a summary.

There are outstanding issues with following ligand(s) in the coordinates.

```
The real space R value indicates that the model for ligand 48D does not correlate to the structure factors.
```

Identifier: 48D
Name: (4E,6S,7S,8S,9E)-7-hydroxy-8-methoxy-4,6-dimethylcyclotetradeca-4,9- dien-1-one
Formula: C17 H28 O3



Type	Program	Version	Descriptor
InChI	InChI	1.02	InChI=1S/C17H28O3/c1-13-10-11-15(18)8-6-4-5-7-9-16(20-3)17(19)14(2)12-13/h7,9,12,14,16-17,19H,4-6,8,10-11H2,1-3H3/b9-7+,13-12+/t14-,16-,17-/m0/s1
InChIKey	InChI	1.02	GNVGUUJMZICZND-MGPSKUMVSA-N
SMILES CANONICAL	CACTVS	3.352	CO[C@H]1C=CCCCCC(=O)CCC(=C/[C@H](C)[C@@H]1O)/C
SMILES	CACTVS	3.352	CO[CH]1C=CCCCCC(=O)CCC(=C[CH](C)[CH]1O)C
SMILES CANONICAL	OpenEye OEToolkits	1.7.0	C[C@H]1/C=C(/CCC(=O)CCCC/C=C/[C@@H]([C@H]1O)OC)C
SMILES	OpenEye OEToolkits	1.7.0	CC1C=C(CCC(=O)CCCCC=CC(C1O)OC)C

Common Tool Enhancements to Ligand Processing

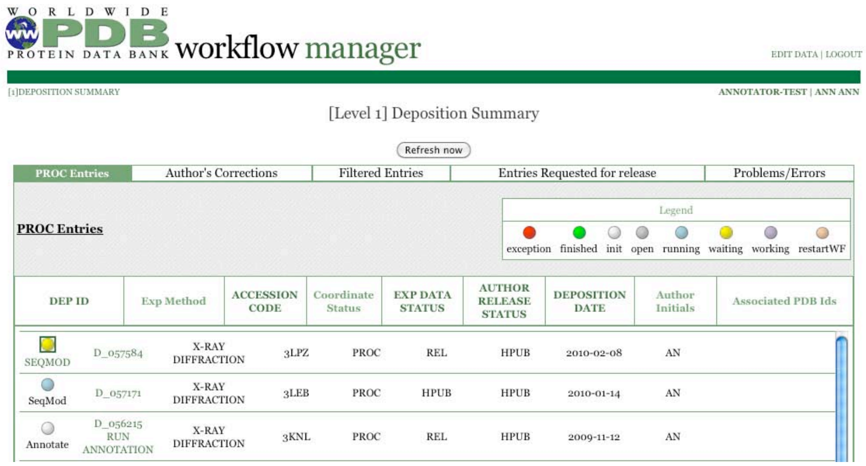
- Automated processing of "correct" existing ligands
- Better integration of process steps during annotation
- User interface to provide 2D, 3D and text views concurrently for ease of analysis
- Use of author provided SMILES descriptor to facilitate ID
- Provide ideal geometry reference through validation against CCD

The Workflow Manager Interface

wwPDB annotators will access the new D&A workflow using the Workflow Manager interface

- Interface provides
  - Summary display of the active workflows
  - Processing status of each entry throughout the annotation process
- Action buttons
  - Launch tasks
  - Provide navigation to view details and browse output files produced by each task

Workflow Manager Example: Level 1



	DEP ID	Exp Method	ACCESSION CODE	Coordinate Status	EXP DATA STATUS	AUTHOR RELEASE STATUS	DEPOSITION DATE	Author Initials	Associated PDB Ids
SEQMOD	D_057584	X-RAY DIFFRACTION	3LPZ	PROC	REL	HPUB	2010-02-08	AN	
SeqMod	D_057171	X-RAY DIFFRACTION	3LEB	PROC	HPUB	HPUB	2010-01-14	AN	
Annotate	D_056215 RUN ANNOTATION	X-RAY DIFFRACTION	3KNL	PROC	REL	HPUB	2009-11-12	AN	

Deposition Interface

Goal To provide a depositor interface that supports data quality, processing efficiency and communication between the annotators and depositors.

Process

- Requirements – annotator and community driven
- Community input and feedback
  - Questionnaire distributed at ACA workshop
  - Mock-ups in preparation and community review planned

ACA 2010 - PDB Depositor Lunch

- 100 attendees
- Introduction of the D&A Project goals
- Review of depositor interface questionnaire
- Answers to questionnaire itself

System Architecture–Drivers & Goals

Scope Growth

- Enable integration of new applications, now and in the future through modularity
- Support for new and hybrid experimental methodologies at the forefront of structural biology

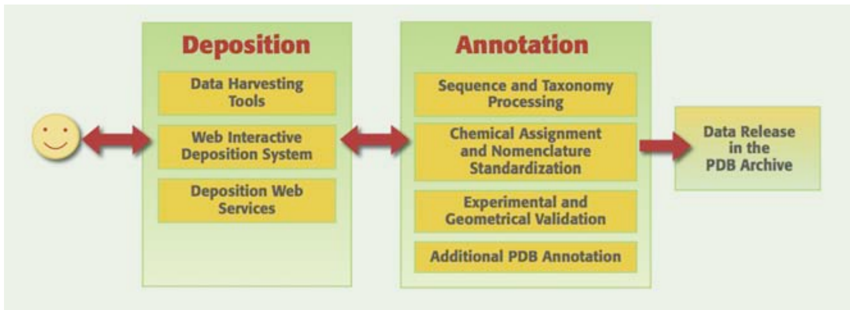
Efficiency

- Greater automation of routine depositor and annotator tasks to support increase throughput and our deeper annotation objectives

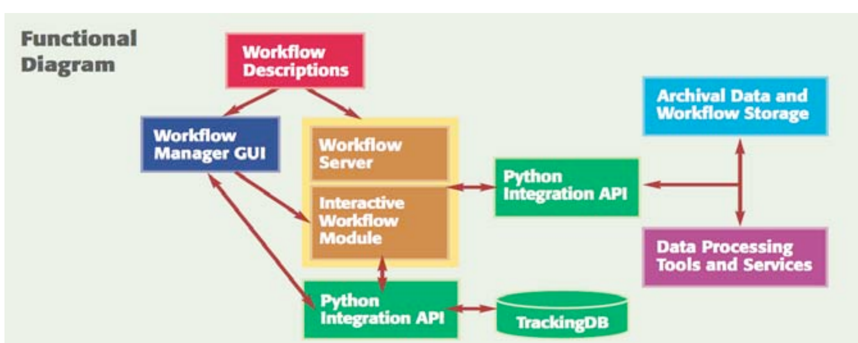
Quality

- Integration of enhanced validation
- Interfaces that provide user feedback
- Improved standardization in annotation by moving from unified data processing practices to a fully unified worldwide software system

System Components



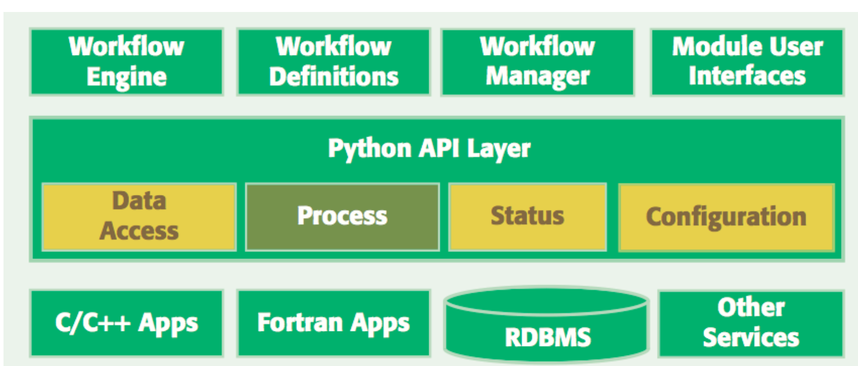
Workflow System Architecture



Workflow System Key Features

- Interactive and distributed batch execution modes
- Reusable workflows are defined XML and translated into Python scripts
- Workflows executed by a workflow engine or an interactive module
- Completion status and tracking details maintained in a relational database
- All data stored in a standardized file system capable of cross-site replication

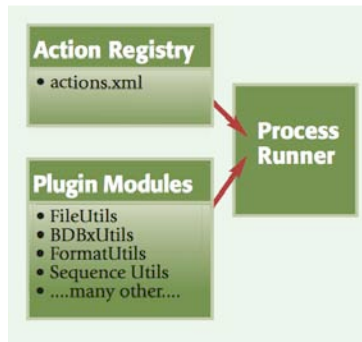
Software Architecture



Application Program Definition

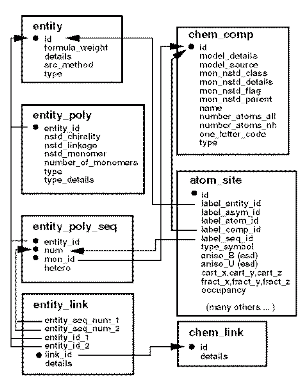
Python API Plug-in Functionality

- Extensibility in describing application functionality
- Programs and tools are defined through an XML format registry
- The registry contains required inputs, outputs, user and internal parameters, and the name of the class and method to be run for each application



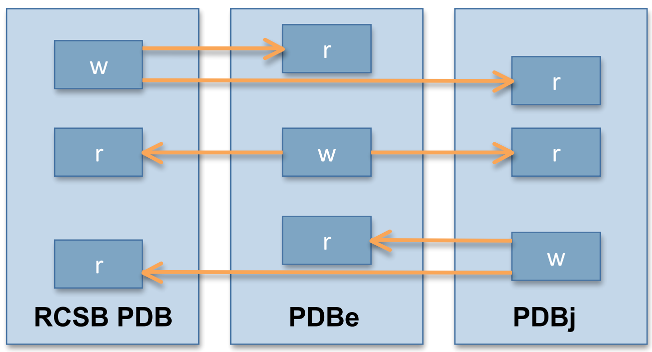
Archival Data Representation

- PDB Exchange Data Dictionary provides framework for representing data
- Supports X-ray, NMR, 3D-Electron Microscopy, SAXS and hybrid methods
- Provides a software-accessible description of the PDB data hierarchy that is used to perform detailed data validation



Physical Data Storage and Sharing

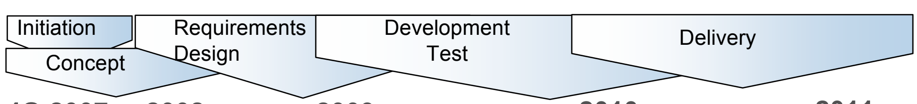
- Architecture provides for archival and working storage
- Worldwide hardware based replication and synchronization



NetApp

w writable data
r read only data
arrows=direction of copy

wwPDB Common D&A Tool Project Timeline



4Q 2007 **2008**
- Concept
- Define deliverables
- Initial design
- Process definition
- Data model definition

2009
- Requirements elaboration
- Data flow documentation
- Technical design
- Technical proof of concept

2010
- Sequence Module
- Ligand Chopper
- Ligand Module
- WF infrastructure
- Deposition Interface design
- Validation module in progress

2011
- D&A system delivery

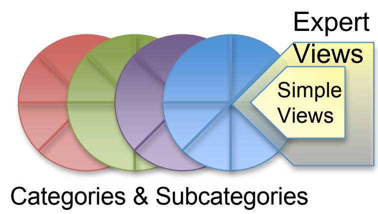
Data Distribution and Query "Data Out"

RCSB PDB AC

October 2, 2010

Philip E. Bourne
Peter W. Rose

Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



Strategy: Enable new scientific views of the archive, through the RCSB PDB website, that reflect structural biology and support both expert and novice access pathways through categorization of the PDB archive. This strategy will drive all activities including web development, enhanced annotation and outreach design.

The result will be more effective access to the archive content and search functionality.

New Layouts and Views

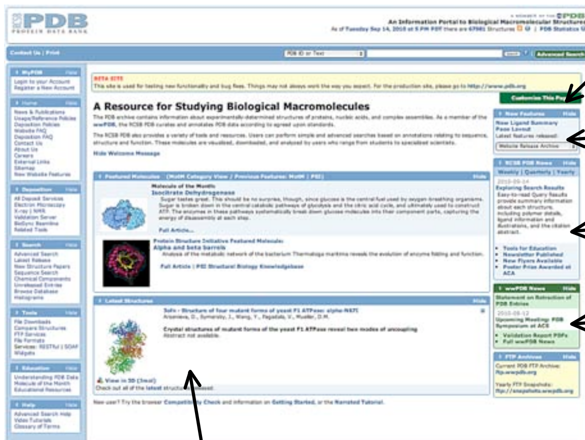
Data Out

www.pdb.org

New Home Page Layout

Objective: Accommodate the preferences of a broad user base

Site is composed of widgets that can be hidden or rearranged



Page customization menu

New site * features

RCSB PDB news

wwPDB news

Latest structure widget

* Feedback from AC

Data Out

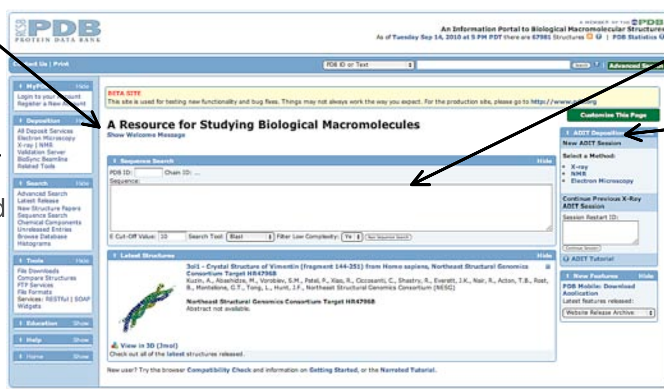
www.pdb.org

Example of a Customized Home Page

Welcome message hidden *

Menu items rearranged

Menu items collapsed



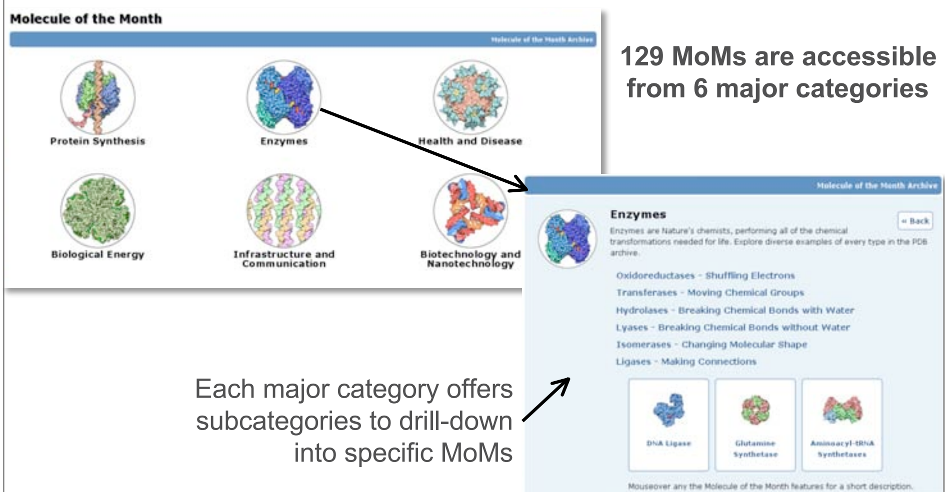
Sequence search widget

ADIT deposition widget

* Feedback from AC

Molecule of the Month – Category View

Objective: Beginnings of a structural view of biology



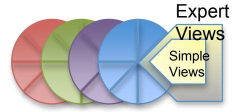
129 MoMs are accessible from 6 major categories

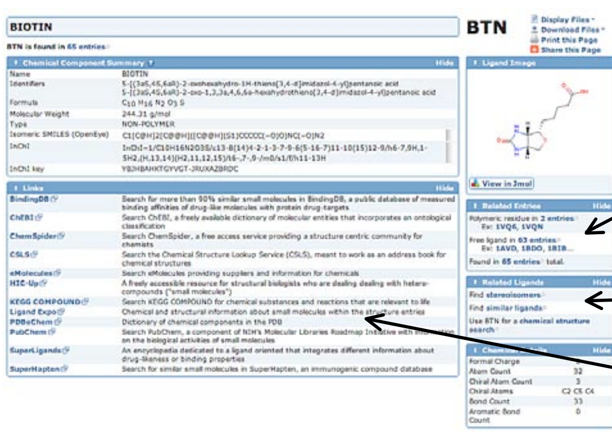
Each major category offers subcategories to drill-down into specific MoMs

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

Ligand Summary Page

Objective: Beginnings of a drug view





Links to entries that contain ligand

Links to related ligands

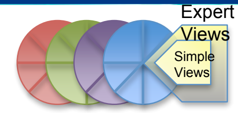
Ligand related external resource links

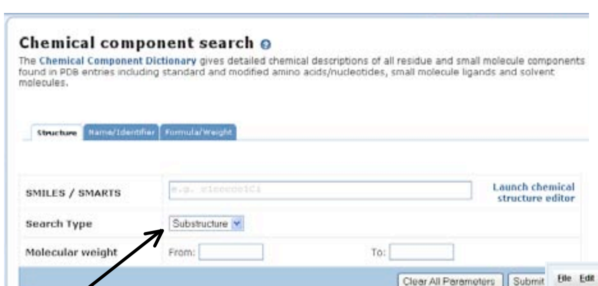
Query and Reporting Tools

Data Out
www.pdb.org

Chemical Components Search

Objective: Beginnings of a drug view





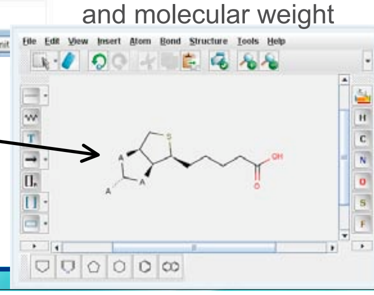
Search for ligands or modified residues by

- Chemical structure (SMILES, SMARTS)
- Name and identifiers (InChI)
- Chemical composition and molecular weight

Search types:

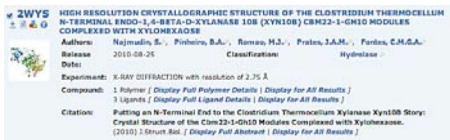
- Exact
- Substructure
- Superstructure
- Similarity

Chemical structure search using Marvin applet (supports atom and bond wildcards)

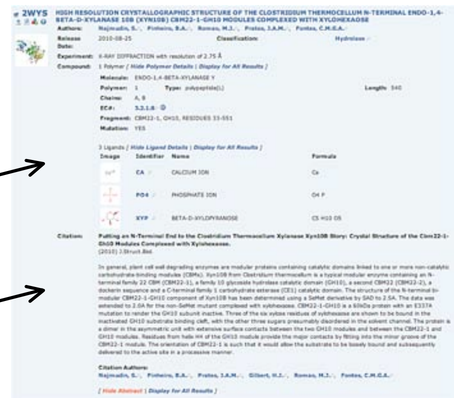


Customizable Query Results Page

Condensed view for
rapid browsing (default) *



Expanded view



Polymer and ligand
details exposed

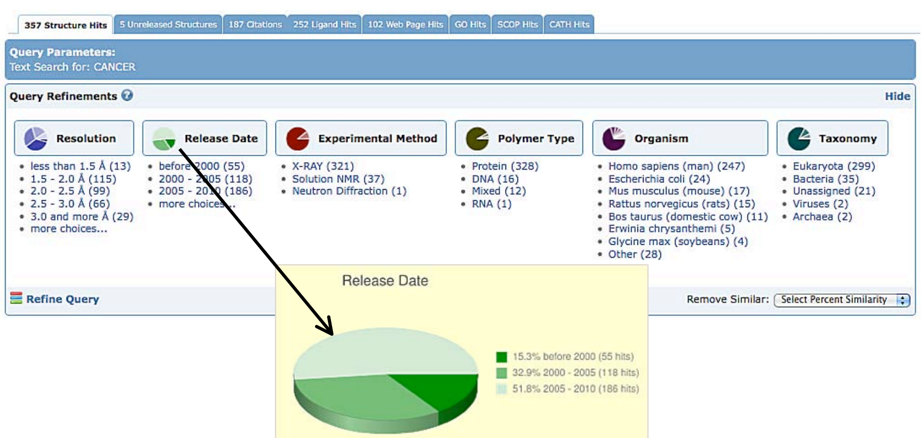
Abstract expanded

* Feedback from AC

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

Query Refinement through Drill-down

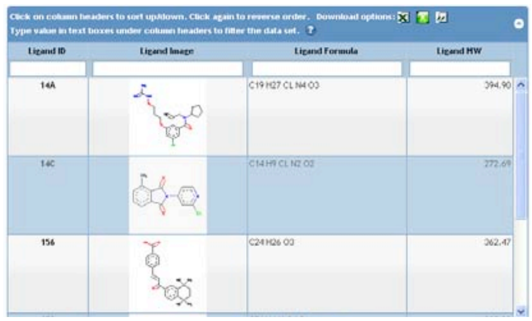
Objective: More intuitive results for all users



Improved Tabular Reports

Objective: Provide exact reports for any user

- Custom and predefined reports
- Sorting and advanced filtering
- Column customization
- Export to Excel, CSV
- Scalable to large tables
- Page navigation
- Resizable
- New fields added by request

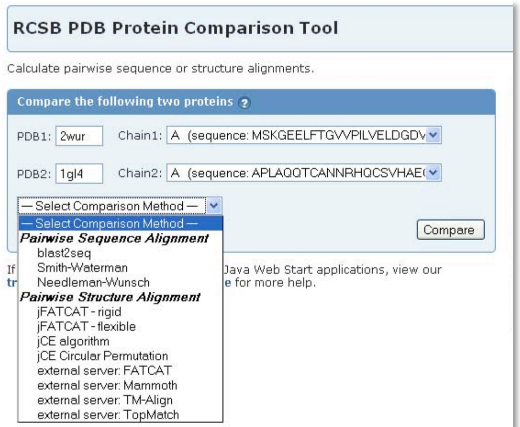


Example: Ligand report

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

Sequence and Structure Analysis and Visualization

Pair-wise Sequence and Structure Comparison



Pre-calculated Protein Structure Alignments at the RCSB PDB Website, *Bioinformatics* in press

All by All Structural Alignment

Objective: Find novel relationships

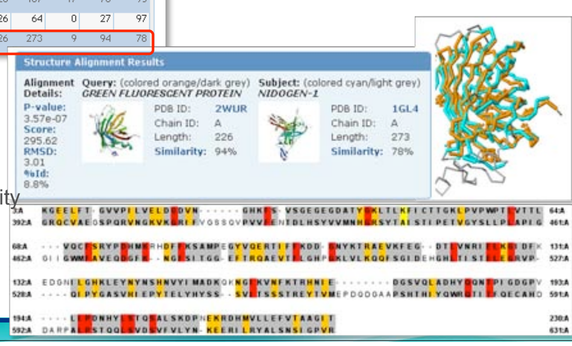
2WUR.A (chain 1) vs. representatives of other sequence clusters (chain 2)

Rank	Result	Chain 2	Title	P-value	Score	Rmsd	Len1	Len2	%ID	%Cov1	%Cov2
1	view	2G2S.B	Green fluorescent p	0.0	478.36	0.93	226	165	96	73	100
2	view	2JAD.A	YELLOW FLUORESCI	0.0	665.32	1.01	226	346	96	100	65
3	view	3E5T.A	Red fluorescent pro	0.0	525.00	1.87	226	228	20	97	96
4	view	3EVP.A	Circular-permutate	0.0	407.39	0.35	226	223	99	61	62
5	view	3GB3.A	KillerRed	0.0	598.80	1.26	226	229	24	98	97
6	view	2G6Y.D	green fluorescent p	7.77E-16	489.59	2.22	226	214	18	93	98
7	view	3EVU.A	Myosin light chain k	2.89E-15	407.23	0.52	226	397	99	62	35
8	view	2A50.D	GFP-like non-fluores	3.06E-12	365.21	2.00	226	167	17	70	95
9	view	2G2S.A	Green fluorescent p	7.95E-10	167.91	0.22	226	64	0	27	97
10	view	1GL4.A	NIDOGEN-1	3.57E-7	295.62	3.01	226	273	9	94	78

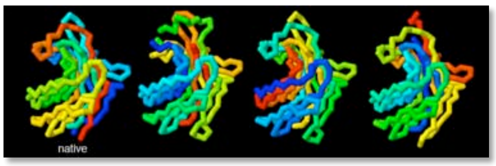
Representative chains from 40% sequence identity clusters are aligned with jFATCAT

Example: Green Fluorescent Protein

- Nidogen-1: similar 11-stranded
- beta-barrel and internal helices
- 3 Å RMSD, only 9% sequence identity
- Nidogen-1: component of basement membrane, no chromophore
- GFP and NID-1 may share common ancestor

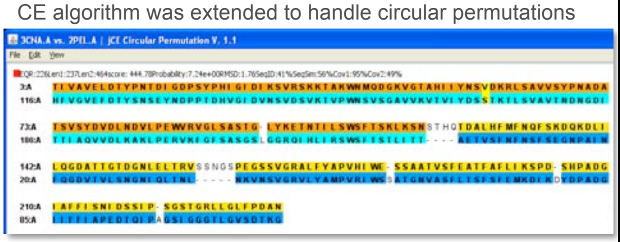


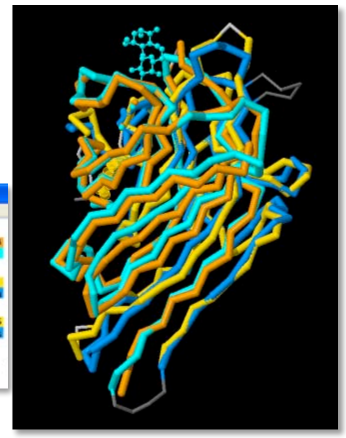
Structural Alignments with Circular Permutations



Example: Concanavalin A and circular permutations (MoM)

CE algorithm was extended to handle circular permutations





Pre-calculated Protein Structure Alignments at the RCSB PDB Website, *Bioinformatics* in press

Structure Visualization

Objective: Support large molecules and novice users

Visualization options on Jmol page

Composite view of split entries

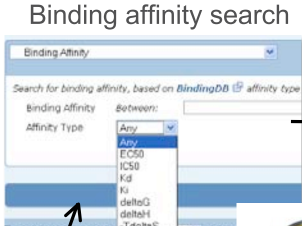
Integration With Other Resources

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

Binding Affinity

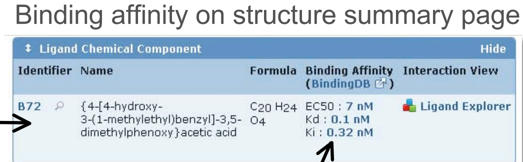
Objective: Link structural with energetic data

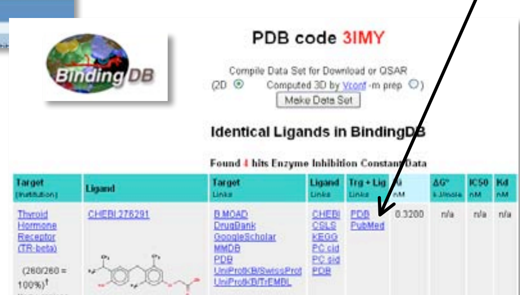
Binding affinity search



Inhibition constants and thermodynamic binding data

Binding affinity on structure summary page





Bi-directional links between RCSB PDB and BindingDB

BindingDB
www.bindingdb.org

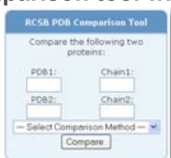
M. Gilson, et al. (2007) *Nucleic Acids Res.* **35**, D198-D201.

Web Widgets

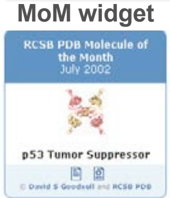
Objective: Enrich other websites with PDB data & tools

Web Widgets: Snippets of code that can be embedded in websites to access RCSB PDB functionality

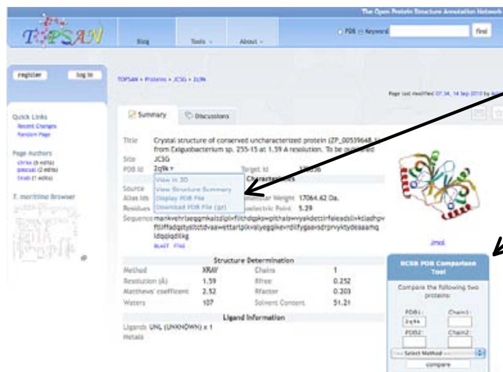
Comparison tool widget



MoM widget



Example: Widgets on TOPSAN site



Tag library Widget

Comparison tool widget

Will Widgets and Semantic Tagging Change Computational Biology? 2010 *PLoS Comp. Biol.* e1000673

Performance Improvements

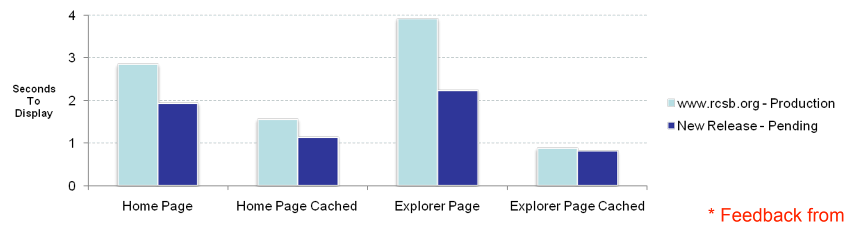
Website Performance Improvements *

Back-end

- Back-end tuning and use of multilevel caching in the areas of searches, query results, explorer pages and hierarchical views
- Result: faster data delivery

Front end

- Cleaner JavaScript and CSS
- Inline image data
- Compressed content
- Result: 25% - 40% increase in render performance



* Feedback from AC

Data Out
www.pdb.org

PDBMobile

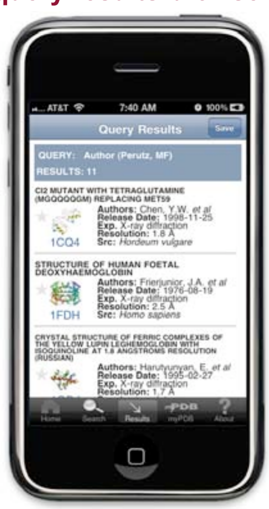
PDBMobile

Objective: Broaden user base through accessibility

- Fast, low bandwidth data access
- Initially supports iPhone iOS 4.1
- Future versions will support Android, Blackberry OS6 and others
- HTML 5-based web application
- Client-side database stores data for offline-access
- Tight integration with MyPDB

PDBMobile

Search interface and query results browser

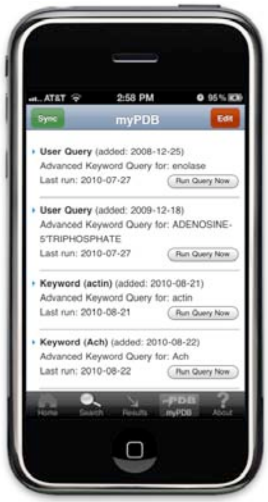


- All returned entries viewable on single page
- Search returns only PDB IDs
- Uses web services

PDBMobile

Tight integration with MyPDB

- Access to saved queries
- Add/delete queries
- Flag interesting entries
- Add personal structure annotations



Data Out
www.pdb.org

Plans for the Next Year and Beyond

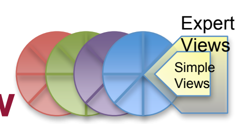
Overall: Meet User Needs Through Appropriate Views

Motivation: Different users come to the PDB with different skill levels, different expectations and different devices.

Objective: Support the work habits of major user groups to maximize their understanding of biology from a structural perspective.

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

What Views? A Simple View and a Drug View



Simple view

- Identify the content for the overarching 6-8 categories that will represent the full PDB archive
- Define the technical strategy and implement categorization of the full archive

Drug view

- Search by drug name and type
- Retrieve by class of receptor

Pragmatic Goals

- Changes to the infrastructure
- Support of new types of data analysis
- New query and reporting features
- Better support for mobile devices and MyPDB

RCSB PDB PROTEIN DATA BANK
Data Out
www.pdb.org

Changes to the Infrastructure

- Adopt middle layer to support new view and query capabilities (entity-based views)
- Expand web services
- Implement improvements requested by users
- Deploy archive remediation releases
- Upgrade hardware

Support of New Types of Data Analysis

- Support for electron density maps
- Effective use of domain information
  - Comparative view of domain assignments by different algorithms
  - All by all structural alignment of protein domains
- Structure comparison of related structures
  - Expand all by all structural alignment to entries within sequence clusters (compare homologs, active, inactive, apo, holo forms)
- Retrieval of similar functional sites
  - SMAP approach based on geometric and evolutionary relationships

New Query and Reporting Features

- Advanced Search
  - Develop search for posttranslational modifications
  - Add functionality for Boolean operations (AND, OR, NOT)
  - Develop search capabilities for PDR (Peptide Reference Dictionary)
- Query refinement
  - Expand drill-down functionality for entries, entities (sequence results), and ligands
- Data reporting
  - Multiple sequence alignments for entity (sequence) search results
  - Display of post-translational modifications in sequence view

Better Support for Mobile Devices and MyPDB

- PDBMobile
  - Deploy alpha release for iPhone
  - Productionize based on user feedback
  - Develop view for iPad form factor
  - Deploy on Android and other HTML 5 compatible devices
- MOMMobile
  - Develop simple structural biology view for the mobile phone
- MyPDB
  - Develop capabilities for personal structure annotation

Outreach and Impact

RCSB PDB AC
October 2, 2010

Christine Zardecki
Andreas Prlic

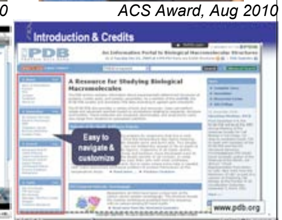

Rutgers Symposium, May 2010


NMR VTF, Sept 2009


San Diego Science Festival, Mar 2010


ACS Award, Aug 2010


Quick Video Tutorials



Online Tutorial Suite

Outreach & Education Goals

- RCSB PDB resource should meet its mission in the interest of science, medicine and education
- RCSB PDB is defined by, designed for, and owned by the communities it serves

RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

International User Communities

- **Biologists** (in fields such as structural biology, biochemistry, genetics, pharmacology)
- **Other scientists** (in fields such as bioinformatics, software developers for data analysis and visualization)
- **Students** and **Educators** (all levels)
- **Media** writers, illustrators, textbook authors
- **General public**

Community Interactions

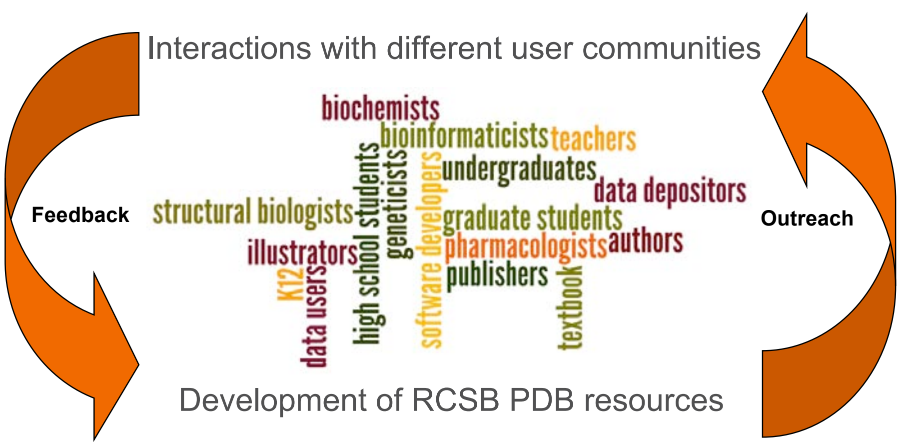
- Electronic help desks, discussion groups
  - New tracking system
- Demonstrations and presentations at professional meetings
- Personal interactions
- Exhibit booths
  - New meetings, improved materials and tracking systems
- Workshops, Posters
- Surveys

ISMB, Aug 2010

PDB Depositors' Lunch, ACA 2010

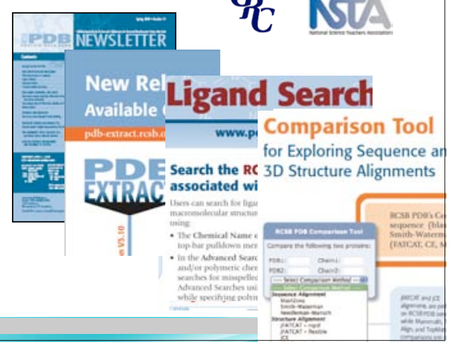
RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

The Outreach Cycle



Tell them, tell them, tell them again

- International scientific meetings and workshops
- Electronic news, RSS feeds, support pages, tutorials, listserv
- Printed and online publications (annual report, newsletter, flyers, brochures)



RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

Educational Activities and Resources

Teachers

- Exhibitions at NJ Science Convention, National Science Teachers Association's Convention
- Presentations for educators

K-12 students

- NJ Science Olympiad, Princeton Science Expo, school visits and tours

Graduate and undergraduate

- Courses at UCSD, Rutgers
- Poster prize
- Internships

General public

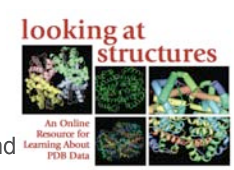
- Rutgers Day, RU Alumni Weekend
- San Diego Science Festival

OpenHelix
Tutorials

Poster Prizes

TAKE ME TO RUTGERS DAY

Events



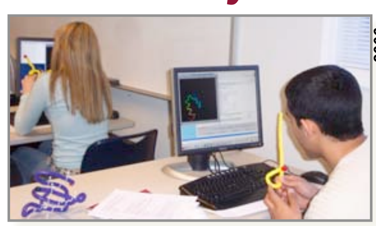
Online and printed resources

Recent Initiatives

Molecular View of Human Anatomy

2008

Learn the basics of molecular structure, PDB, and a given theme

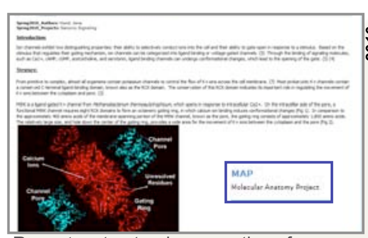


2006

Explore molecular structures related to the course theme

2008

Present using a molecular structural perspective of assigned topics



2010

Report a structural perspective of assigned/selected topics online

Students Exploring Molecular Structures (SEMS) Trial Courses

Courses at Rutgers

- **Undergraduate** *Molecular View of Human Anatomy* (2006, 2008, 2010) explored digestive system, cancer and AIDS, nervous system
- **Graduate** *Biophysical Chemistry* (2006, 2008)
- **Summer internships** (2006, 2008) explored digestive system, endocrine system

Planned Courses (2011-2012)

- Rutgers University
- King's College, PA
- Georgetown University, DC
- Wellesley College, MA

RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

Rubrics for Evaluation

Criteria	*Type of Learning*	*Student Ability Scoring Criteria*	*Score*
1	Knowledge	Recognizes building blocks and polymers of basic biological macromolecules. Recognizes structural features and conformation of proteins and nucleic acids.	1-5
2	Knowledge	Understands basic principles of bio-macromolecular interactions (covalent and non-covalent) and can recognize them in any given molecule or complex.	1-5
3	Knowledge	Understands the basis of biomolecular structure determination; recognizes the difference between different methods used and what can be learned from these structures	1-5
4	Skill	Can access, query and identify relevant molecular structures from the PDB	1-5
5	Skill	Can use appropriate visualization software to visualize molecular structures from the PDB. Should be able to select specific regions of the structure to highlight shape, interactions and other important details.	1-5
6	Skill	Can create clear labeled figures with legends to explain structure-function relationships and tell a molecular story	1-5
7	Knowledge/ skill	Can describe structure in words (written/oral) and provide appropriate attributions	1-5
8	Problem solving	Can search for additional information about the molecule in literature, databases and other authoritative resources	1-5
9	Application/ Creative thinking	Can compare structures of related molecules. Can relate molecular structure to biochemical, genetic or other known data.	1-5
10	Creative thinking	Can recognize unreported details about structure and discuss its implication on structure and/or function	1-5

What Students Learned

- Structural perspective of course theme
- Active learning skills
  - Visualization of molecules
- Self-learning skills/abilities
  - Research about a topic
  - Read scientific literature
  - Presentation skills
  - Write scholarly articles
- Application of curricular knowledge to comprehend reaction mechanism and regulation, etc.

Outreach and Impact
www.pdb.org

Journal Collaborations

- Coordination of *Instructions to Authors*
- Coordinating PDB release with online publication
  - Initially from NPG and IUCr journals
  - Now *JMB* (Top PDB Journal), *PNAS*, *Proteins*
  - In progress: *FEBS Journal*
- Validation Reports

Published 454 entries in 2009

Published 663 entries in 2009

Published 123 entries in 2009

Google Students

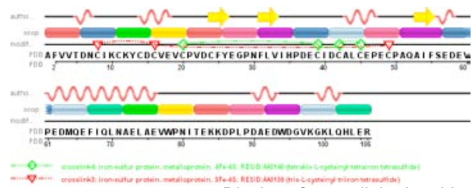
Global program that offers student developers stipends to write code for various open source software projects

Jianjiong Gao, graduate student in Computer Science, University of Missouri-Columbia

Mark Chapman, graduate student in Computer Sciences, University of Wisconsin-Madison

- Students work remotely
- Weekly Skypes, many emails

- Goal of developing new tools
  - New Multiple Sequence Alignment algorithm will be used as part of Comparison Tool
  - Identification of modified residues will be used in Sequence Tab



Display of cross-linked residues

Expanding current initiatives

National Science Teachers Association

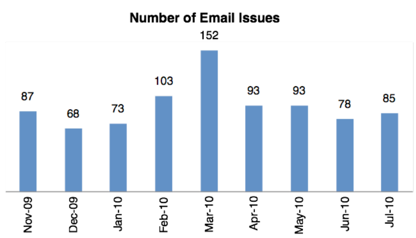
Science Olympiad Protein Modeling offered in more states, no longer trial event

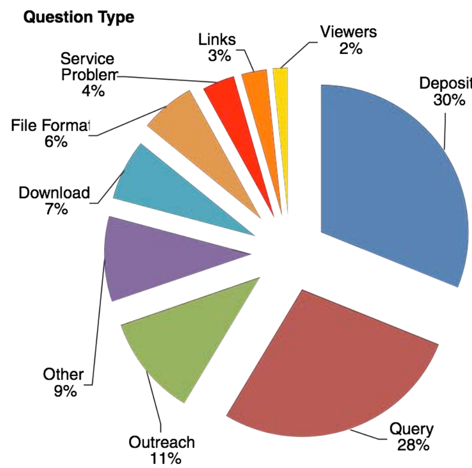
RU Chemistry Society Outreach
Getting others to spread the RCSB PDB word

Help Desk (info@rcsb.org)

New email-tracking software implemented last fall

- 832 emails sent to info@rcsb.org (11/1/09 – 7/31/10)
- 632 unique users



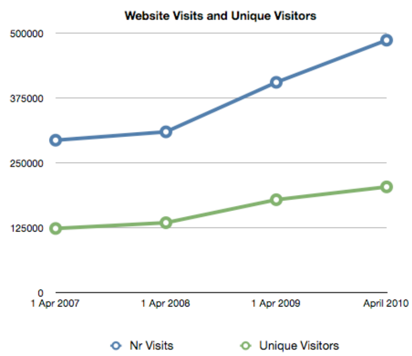


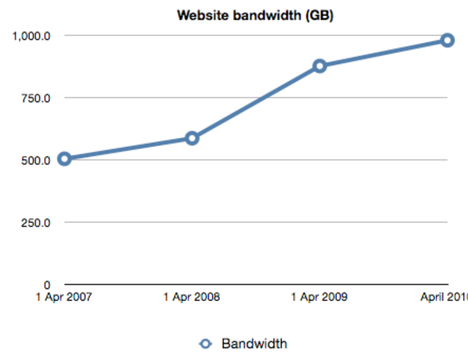
RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

Impact

RCSB PDB Website Usage

Number of visits and page views is growing faster than number of unique visitors





Non-Bounce Visits

- We have a number of short website visits (1 page viewed)
- Whenever we can, we use the stats for visits that look at more than just one page
- These are the “non-bounce” visits

15% Growth in a Year

Apr 17, 2010 - May 16, 2010
Comparing to: Apr 18, 2009 - May 17, 2009

Site Usage

320,593 Visits
Previous: 280,749 (+14.19%)

2,973,525 Pageviews
Previous: 2,478,028 (+20.00%)

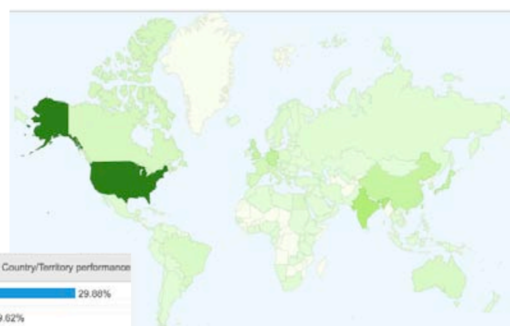
9.28 Pages/Visit
Previous: 8.83 (+5.08%)

37.14% % New Visits
Previous: 31.97% (+16.18%)

RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

Who is Using the RCSB PDB Globally?

- 320K visits (*) from 152 countries/territories per month



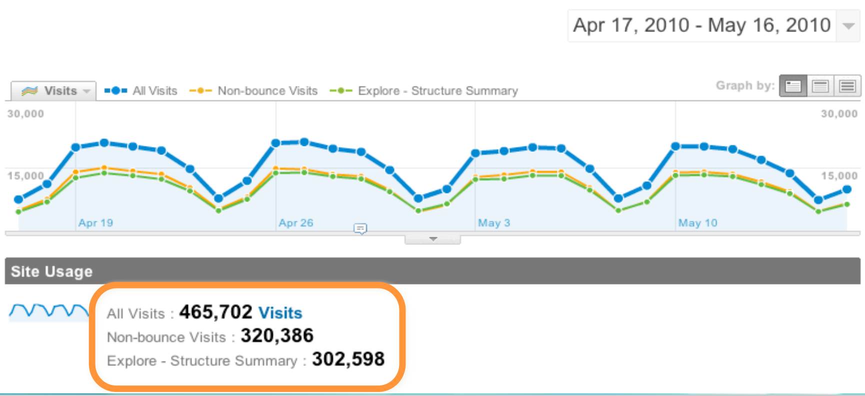
	Detail Level: Country/Territory	Visits	Individual Country/Territory performance
1.	United States	95,747	29.88%
2.	India	30,833	9.62%
3.	Germany	18,959	5.92%
4.	Japan	18,054	5.64%
5.	China	17,868	5.58%
6.	United Kingdom	16,366	5.11%
7.	France	9,535	2.98%
8.	Italy	9,164	2.86%
9.	Canada	7,982	2.49%
10.	Spain	6,931	2.16%

Visits from Apr. 17 – May 16, 2010 that include at least two page views, total visits = 465K

Who are the Major User Groups?

General audience

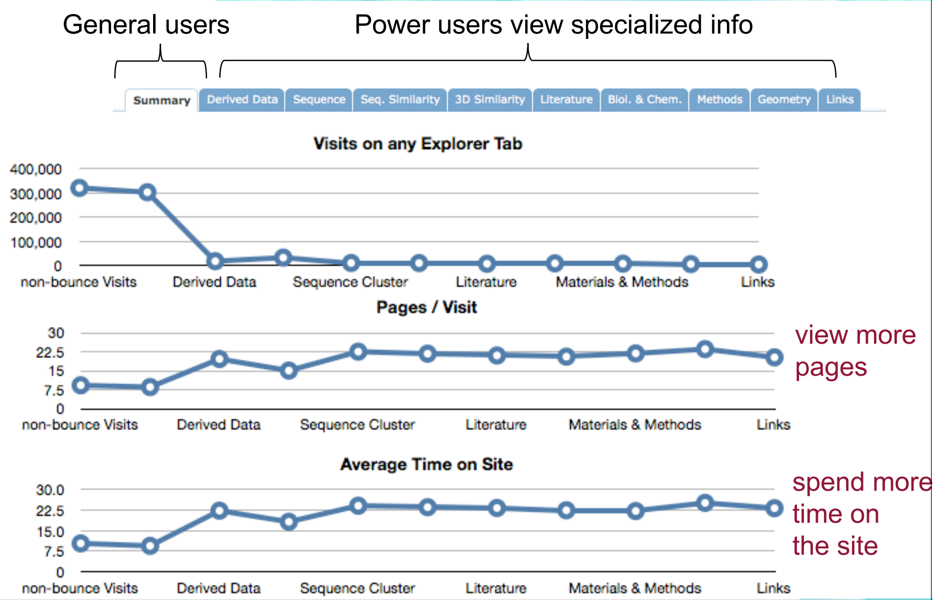
- Almost all visitors go to the Structure Summary Page



RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

"Power Users"

- ~ 12% of visitors (~ 40K visits per month)
- Use advanced search & tools, view 3D structures & ligands
- Seek advanced information on Structure Summary tabs



Educational Users

- ~10 % of visitors (~ 30 K visits per month)
- View *Molecule of the Month*, *Understanding PDB Data*, and Educational Resources pages

Growth of MyPDB

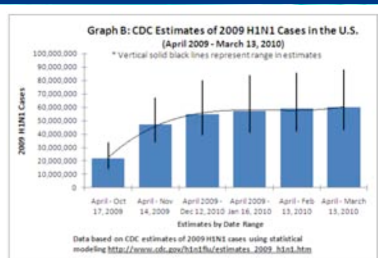
Number of registered users

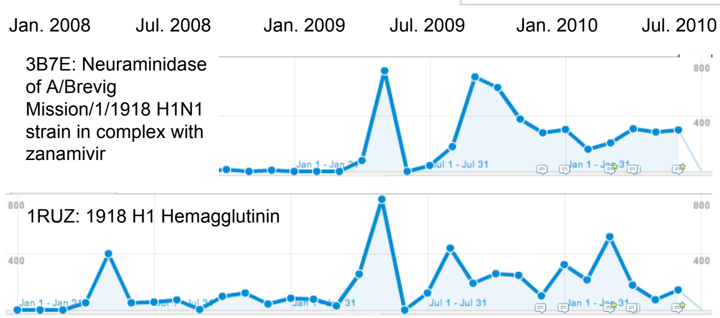
Registration Date

RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

RCSB PDB as a Research Tool for Influenza

Structure Summary page activity for H1N1 Influenza related structures





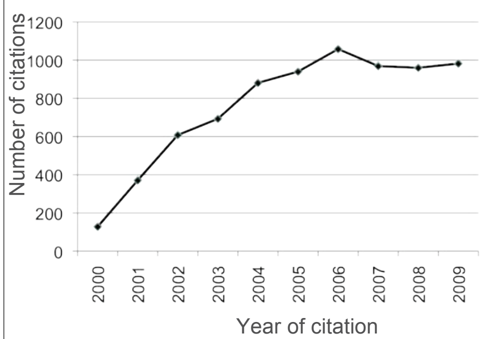
* http://www.cdc.gov/h1n1flu/estimates/April_March_13.htm

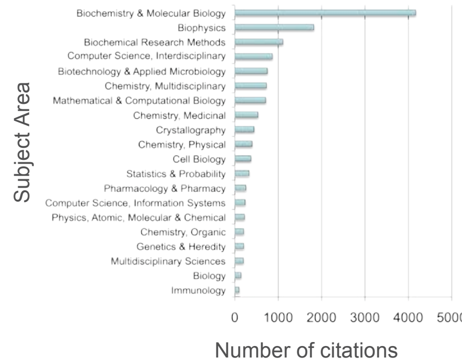
Outreach and Impact

www.pdb.org

Impact of RCSB PDB Reference

Berman et al., Nucleic Acids Res. (2000)





Cited more than 8000 times by

- 6921 articles
- 661 reviews
- 455 proceedings papers

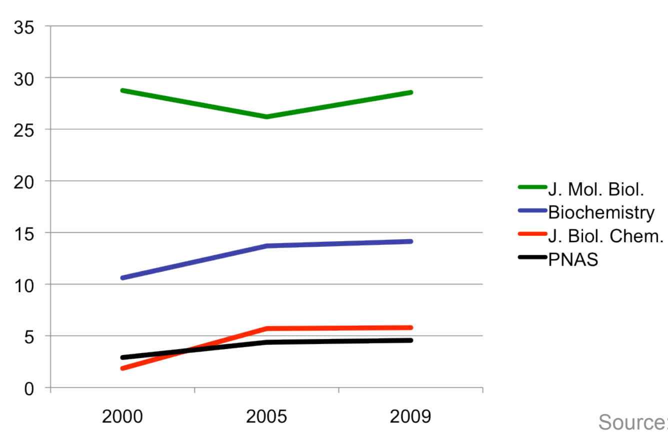
Source: ISI Web of Knowledge

Outreach and Impact

www.pdb.org

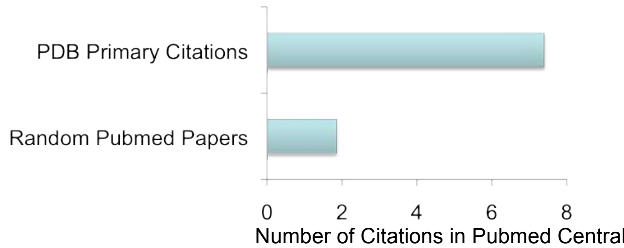
Percentage of Journal Articles Describing PDB structures

Percentage of articles that are PDB primary citations



Source: PubMed

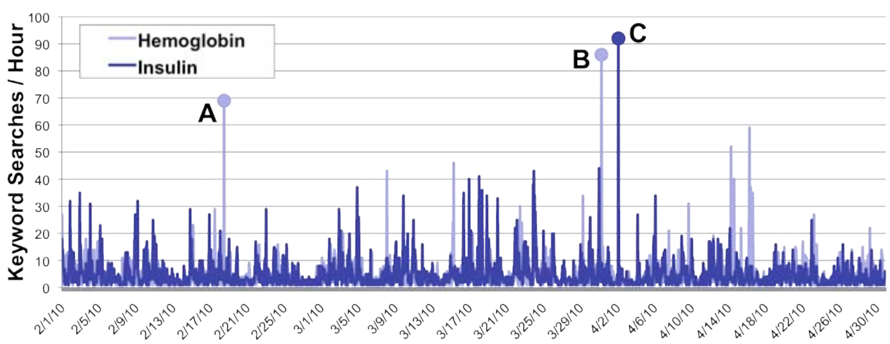
Impact of PDB Primary Citations



2262 PDB primary citations published in 2005
vs. 10,000 random papers published in 2005
(out of 693,092 in PubMed)

Source: PubMed

Evidence of Classroom Usage



Large coordinated classroom searches at worldwide universities
A. Universidad Nacional Mayor de San Marcos (Lima, Peru)
B. Universita' degli Studi del Piemonte Orientale (Piemonte, Italy)
C. Bits Pilani – K. K. Birla Goa Campus (Mormugao, India)

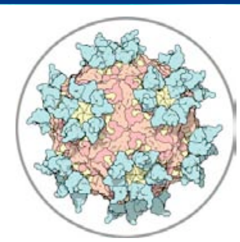
Summary

- Personalized outreach to numerous groups is effective, but not always scalable
  - We can't go to every professional society meeting
  - Can't visit every high school
- Website is definitely being heavily used, but is not always personalized
  - In-depth use is by advanced/experienced researchers ("power users")
  - Education pages/Molecule of the Month reaches a broad community in a broad way

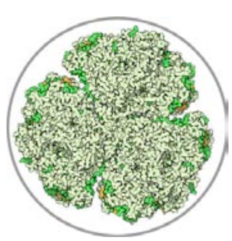
RCSB PDB PROTEIN DATA BANK
Outreach and Impact
www.pdb.org

Protein Synthesis

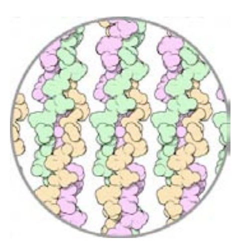
Enzymes



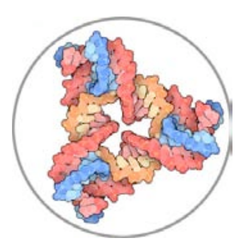
Health & Disease



Biological Energy



Infrastructure & Communication

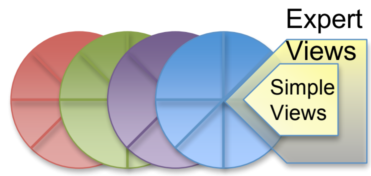


Biotechnology & Nanotechnology

Outreach and Impact

www.pdb.org

Strategic Goal: To create contextual views of the archive that will foster awareness of, and insight into, the structural basis of biology



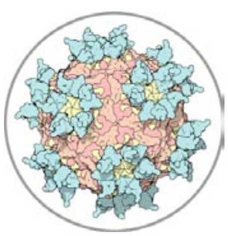
Strategy: Enable new scientific views of the archive, through the RCSB PDB website, that reflect structural biology and support both expert and novice access pathways through categorization of the PDB archive. This strategy will drive all activities including web development, enhanced annotation and outreach design.

The result will be more effective access to the archive content and search functionality.

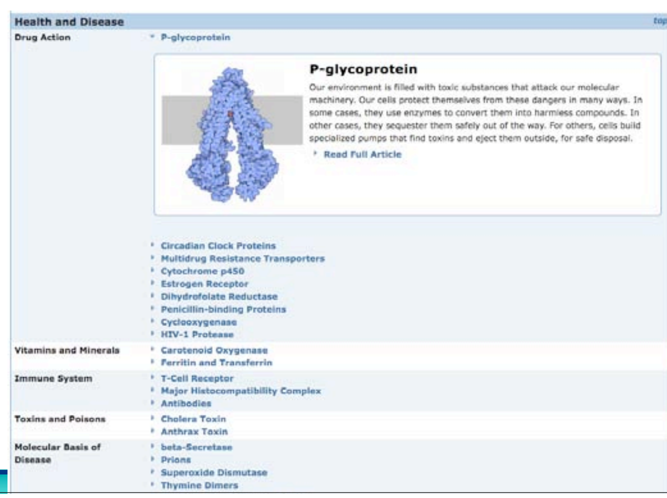
Outreach and Impact

www.pdb.org

First pass: Categorizing all Molecule of the Month Articles



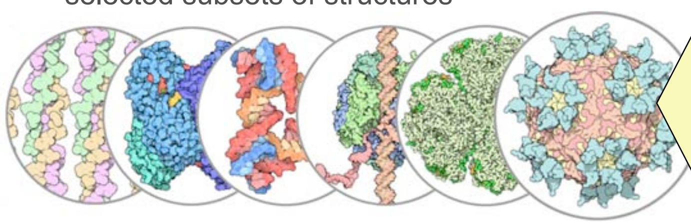
Health & Disease



For High-level Users…

Expert View

- Divide PDB to fit categories
  - Are there more categories needed?
  - Use categories to cross reference annotation
- Provide same services for all categories
  - Current functionalities for searching and reporting for pre-selected subsets of structures

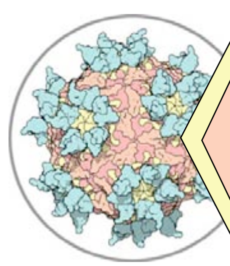


Expert view on Health & Disease

…to Non-expert Users

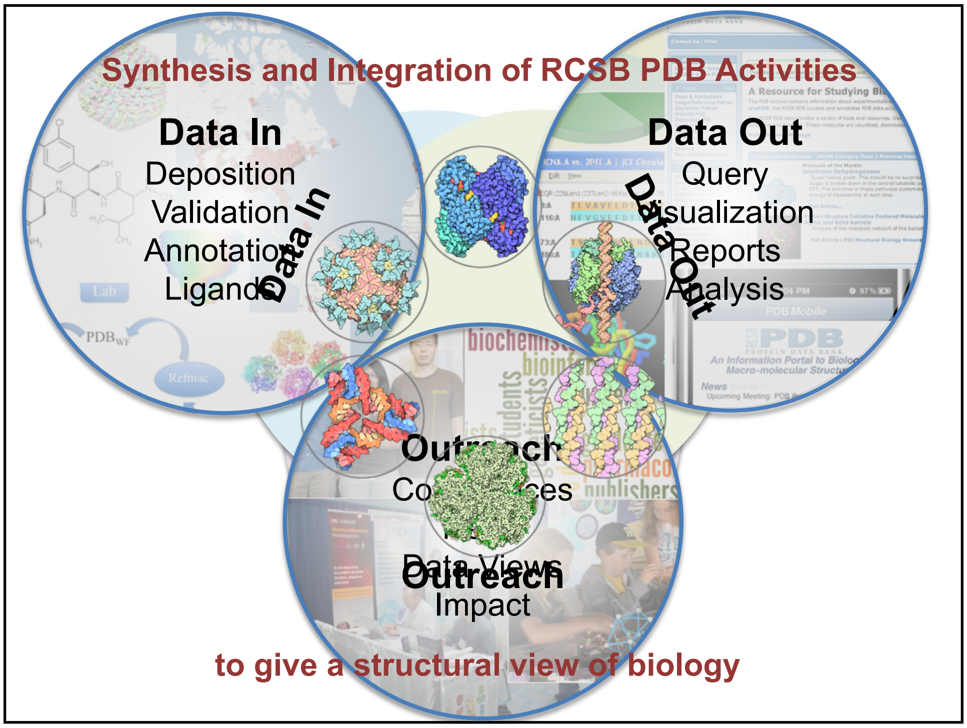
Simple View

- By selecting a subset of structures, users will access only entries referenced in *Molecule of the Month* articles
- Automatically integrates structures for users
- Could form basis for quick home page



Simple view on Health & Disease

Expert view on Health & Disease


Synthesis and Integration of RCSB PDB Activities
Data In
Deposition
Validation
Annotation
Ligands
Data Out
Query
Visualization
Reports
Analysis
Outreach
Data Views
Impact
to give a structural view of biology