Abstracting Knowledge from Protein Structures for Biology in the 21st Century

PDB40 Symposium

CSHL
October 2011

Janet Thornton
EMBL-EBI
Overview

• Personal Recollections of PDB

• Abstracting knowledge from structures for biology in the past and today

• Thoughts about the Future of PDB

• Thanks
Personal Recollections of the PDB: 1974 - 1995

- 12” tapes about every 3 months from Brookhaven via Daresbury to Oxford Lab in ~1974

- Growth in number of entries (‘70s)

- Validation 1989 CCP4 ‘Errors in Protein Structures’ / PDBClean/PROCHECK

- Visits to Brookhaven (Tom Koetzle, Frances Bernstein & Enrique Abola) as part of Scientific Advisory Board

- Challenges of data increase – move to RCSB: Helen, Phil & Gary
Personal Recollections of the PDB: 1995 onwards

- Establishing PDBe – grant from Wellcome Trust (for 4 staff) to EMBL- EBI:
  - 1995 – recruitment of Kim Henrick & Geoff Barton
  - Building relationships between PDBe & RCSB/PDBj/BMRB 1995 - 2005
  - Kim & colleagues started to build the EMDB (2002)
  - Establishment of wwPDB
  - Recruiting Gerard (Kleywegt) – 2009
    ‘Bringing Structure to Biology’
Abstracting Knowledge from the PDB

• The knowledge contributed by an individual protein structure about how this particular protein performs its biological function remains the most important aspect of knowledge in the PDB e.g. Von Willebrand Factor

• BUT additional knowledge in many areas can also be abstracted by combining information over many structures. In practice most proteins interact with many other molecules, either as multimers or as parts of pathways

∑ Information over all or subset of PDB entries to generate knowledge
Abstracting Knowledge from PDB: Historical perspective

- Practical knowledge e.g. Which proteins are likely to crystallise
- Basics Principles of Protein Structure (physics/chemistry)
- The Universe of Proteins & evolutionary relationships
- Structure to Function
1970’s Basic Principles of Protein Structure (Understanding Sequence to Structure)

- Properties of amino acids eg helix propensities
- Basic geometry of pp chain, e.g. phi, psi values
- Hydrophobic Core
- Secondary Structures
  - Helices - geometry; length, curvature; packing
  - Strands – twist; geometry; residue pairs
  - Turns – types; residue preferences
- Chirality
  - Twists of sheets, Right handed βαβ, Barrels
- Tools for ‘describing’ protein structures
  - Secondary Structure Assignment - DSSP
  - Hydrogen bonds - HBPlus
  - Accessibility - NACCESS
1980’s The Universe of Protein Structures from the PDB

- Interactions:
  - Amino acid packing
  - Tertiary packing – helix; sheet
- Domains & multi-domain architectures
- Folds
- Evolution – conserved structures
- New Tools
  - Visualisation
  - Homology Modelling
  - Simulations
  - Electrostatics
1990s Folds; Classification; Interactions

- Protein Structure Classifications
  - CATH & SCOP

- Interactions
  - Protein-protein
  - Protein-Ligand
  - Protein-DNA

- New Tools:
  - Structure Comparison eg DALI
  - Patch Analysis for PPI
  - Docking
  - Fold Recognition - Threading
Many of Tools now provided by PDB as searches

- PDBeMotif – to identify motifs
- PDBePISA – to assign multimeric status in crystal
- PDBeFold – to find all similar folds in PDB
## Structural Genomics Projects ~2000

Taken from [www.isgo.org](http://www.isgo.org)

<table>
<thead>
<tr>
<th>Project</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontario Centre for <strong>SG</strong></td>
<td>Canada</td>
</tr>
<tr>
<td>Montreal-Kingston Bacterial <strong>SG</strong> Initiative</td>
<td>Canada</td>
</tr>
<tr>
<td>Montreal Network for Pharmaco-Proteomics and <strong>SG</strong></td>
<td>Canada</td>
</tr>
<tr>
<td>CyberCell Project</td>
<td>Canada</td>
</tr>
<tr>
<td>Structural Proteomics in Europe (SPINE)</td>
<td>Europe</td>
</tr>
<tr>
<td><strong>SG</strong> of <em>Mycobacterium</em> pathogens</td>
<td>France</td>
</tr>
<tr>
<td><strong>SG</strong> of Eukaryotes</td>
<td>France</td>
</tr>
<tr>
<td>Yeast <strong>SG</strong></td>
<td>France</td>
</tr>
<tr>
<td><strong>SG</strong> of Orphan <em>E. coli</em> Genes</td>
<td>France</td>
</tr>
<tr>
<td>Protein Structure Factory</td>
<td>Germany</td>
</tr>
<tr>
<td>RIKEN <strong>SG</strong>/Proteomics Initiative</td>
<td>Japan</td>
</tr>
<tr>
<td>National Project on Protein Structural and Functional Analyses (7 centers)</td>
<td>Japan</td>
</tr>
<tr>
<td>Biological Information Research Center (BIRC)</td>
<td>Korea</td>
</tr>
<tr>
<td>The Korean Structural Proteomics Research Organization</td>
<td>Korea</td>
</tr>
<tr>
<td>National Centers for Competence in Research (NCCR)</td>
<td>Switzerland</td>
</tr>
<tr>
<td>North West <strong>SG</strong> Centre</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Oxford Protein Production Facility</td>
<td>UK</td>
</tr>
<tr>
<td>Cambridge Group</td>
<td>UK</td>
</tr>
<tr>
<td>New York <strong>SG</strong> Research Consortium</td>
<td>USA</td>
</tr>
<tr>
<td>Midwest Center for <strong>SG</strong></td>
<td>USA</td>
</tr>
<tr>
<td>Berkeley <strong>SG</strong> Center</td>
<td>USA</td>
</tr>
<tr>
<td>Northeast <strong>SG</strong> Consortium</td>
<td>USA</td>
</tr>
<tr>
<td>TB <strong>SG</strong> Consortium</td>
<td>USA</td>
</tr>
<tr>
<td>Southeast Collaboratory for <strong>SG</strong></td>
<td>USA</td>
</tr>
<tr>
<td>Joint Center for <strong>SG</strong></td>
<td>USA</td>
</tr>
<tr>
<td><strong>SG</strong> of Pathogenic Protozoa Consortium</td>
<td>USA</td>
</tr>
<tr>
<td>Center for Eukaryotic <strong>SG</strong></td>
<td>USA</td>
</tr>
<tr>
<td>Structure 2 Function Project</td>
<td>USA</td>
</tr>
</tbody>
</table>
From Structure to Function

Protein Structure

Molecular Function
biological multimeric state

MULTIMERS

INTERACTIONS

FOLD

SURFACE

3D STRUCTURE

ELECTROSTATICS

CLUSTERS

LIGANDS

MUTANTS & SNPs

ligand & functional sites
catalytic clusters, mechanisms & motifs
enzyme active sites
Fold & Function

- No direct correlation between fold & function, though some tendencies
  - DNA binding proteins tend to be helical
  - Haem binding proteins tend to be helical
  - Enzymes tend to adopt $\alpha\beta$ folds
  - Immune-related proteins tend to be $\beta$-sheet structures e.g. Ab
  - Membrane proteins are predominantly helical – apart from porins
However identifying sequence or structural similarity (i.e. identifying an evolutionary relationship) is the most powerful route to function assignment.

BUT members of the same protein superfamily often have a related but not identical function.
MICROEVOLUTION BY ARTIFICIAL SELECTION

Rock pigeon

Grey wolf

Cabbage (condensed shoot)

Kohlrabi (swollen stem & leaf bases)

Broccoli (flower cluster)

Brussels sprouts (Lateral buds)

Wild cabbage
Aspartate Amino Transferase Superfamily

- Aspartate Aminotransferase (2.6.1.1)
- 2,2-Dialkylglycine Decarboxylase (4.1.1.64)
- Tyrosine Phenolylase (4.1.99.2)
- Ornithine Decarboxylase (4.1.1.17)
SDR Family
Short chain dehydrogenase/reductase family

>60 in humans

Catalytic Tetrad: S,Y,K,N

Different Functions:
Oxidoreductases E.C. 1.1 & 1.3;
Lyases E.C. 4.3;
Isomerases E.C. 5.1

Many structures solved
Many different substrates
Understanding Enzyme Families and Evolution
Understanding Enzyme Families & Evolution

- Data
  - Protein Sequences
  - Protein Structures with ligands!
  - Substrate Knowledge (promiscuity)
    - in vitro
    - In vivo
  - Reaction mechanisms

- Computational tools for:
  - Sequence comparison
  - Structure comparison
  - Small molecule comparison
  - Reaction comparison

- Then we need to integrate and visualise all these data!!
The pipeline

Structure and sequence alignments for enzyme families -> Phylogenetic trees

Annotate with functional information and small molecule data (e.g., substrates, mechanism)
Phosphatidylinositol-Phosphodiesterase (PIP) Superfamily

Covalent Catalysis

General Acid/Base Catalysis

Bacteria

Eukaryotes

Hydrolysis occurs outside the enzyme

Hydrolysis occurs in the active site

EC 3.1.4.41

EC 4.6.1.13

EC 3.1.4.11
EC 3.1.4.46
Phosphatidylinositol-Phosphodiesterase Superfamily

G1

E.C. Code | Substrate | Product | Reaction
--- | --- | --- | ---
None | None | None
3.1.4.46 | | | |
017_MYCTU P96236 | | | |
008_THETN Q8RB32 (2pz0) 2pz0B00 MACIE | | | |
014_BACSU O07592 | | | |
015_TREPA O30405 | | | |
005_ECOLI P09394 (1ydy) 1ydyA00 MACIE | | | |
007_STAAN Q7A6H7 (2p76) 2p76A00 MACIE | | | |
006_STAEE A6QFH0 (2oog) 2oogD00 MACIE | | | |
018 YEAST Q08959 | | | |
003_SHIFL Q83PU8 (2otd) 2otdA01 MACIE | | | *
019_BOVIN Q3T0T0 | | | |
021_RAT Q9JL55 | | | |
020_HUMAN Q9HCC8 | | | |
004_CYTH3 Q11WD4 (3ch0) 3ch0A00 MACIE | | | |
022_ARATH Q9STZ3 | | | |
000_RAT P10688 (1dix) 1dixA02 MACIE | | | *

G2

E.C. Code | Substrate | Product | Reaction
--- | --- | --- | ---
4.6.1.13 | | | |
001_LISMO P34024 (2pcl) 2pclA00 MACIE | | | |
016_STAEE P45723 | | | |
009_BACTU P08954 (2or2) 2or2A00 MACIE | | | |
010_BACCE P14262 (2ptd) 2ptdA00 MACIE | | | |

G3

E.C. Code | Substrate | Product | Reaction
--- | --- | --- | ---
3.1.4.41 | | | |
002_LOXLA Q8I914 (1xx1) 1xx1A00 MACIE | | | |
023_LOXIN Q1W694 | | | |
### Phosphatidylinositol-Phosphodiesterase Superfamily

<table>
<thead>
<tr>
<th>E.C. Number</th>
<th>Substrate</th>
<th>Multi-domain Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.4.46</td>
<td><img src="image1" alt="Substrate" /></td>
<td><img src="image2" alt="Architecture" /></td>
</tr>
<tr>
<td>3.1.4.1</td>
<td><img src="image3" alt="Substrate" /></td>
<td><img src="image4" alt="Architecture" /></td>
</tr>
<tr>
<td>3.1.4.44</td>
<td><img src="image5" alt="Substrate" /></td>
<td><img src="image6" alt="Architecture" /></td>
</tr>
<tr>
<td>3.1.4.43</td>
<td><img src="image7" alt="Substrate" /></td>
<td><img src="image8" alt="Architecture" /></td>
</tr>
<tr>
<td>3.1.4.46</td>
<td><img src="image9" alt="Substrate" /></td>
<td><img src="image10" alt="Architecture" /></td>
</tr>
<tr>
<td>3.1.4.11</td>
<td><img src="image11" alt="Substrate" /></td>
<td><img src="image12" alt="Architecture" /></td>
</tr>
</tbody>
</table>

- **Difference in substrate**: Indicates changes in substrate structure.
- **Difference in product**: Indicates changes in product structure.
- **Difference in multi-domain architecture & substrate**: Indicates changes in the domain architecture.

**G1**

- **Hydrolytically removes 5'-nucleotides successively from the 3'-hydroxy termini of 3'-hydroxy-terminated oligonucleotides**

- **Known structure with bound cognate ligand shows active site located in single domain; second domain not contributing to functional change**

* Not in archshema as not in reviewed uniprotkb
✝ Not in Funtree as filtered out by sequence similarity
3.1.4.46
Hydrolytically removes 5’-nucleotides successively from the 3’-hydroxy termini of 3’-hydroxy-terminated oligonucleotides

3.1.4.1

3.1.4.44

3.1.4.43

3.1.4.46

4.6.1.13
Loss of metal co-factor

4.6.1.14

Difference in substrate

Difference in substrate

Difference in substrate

Difference in product

Difference in multi-domain architecture & substrate

Difference in product

* Not in archshema as not in reviewed uniprotkb
† Not in Funtree as filtered out by sequence similarity

E.C. Number | Substrate | Multi-domain Architecture
---|---|---
3.1.4.46 | Hydrolytically removes 5’-nucleotides successively from the 3’-hydroxy termini of 3’-hydroxy-terminated oligonucleotides | 
3.1.4.1 | | 
3.1.4.44 | | 
3.1.4.43 | | 
3.1.4.46 | | 

Product

Known structure with bound cognate ligand shows active site located in single domain; second domain not contributing to functional change

<table>
<thead>
<tr>
<th>G1</th>
<th>G2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in substrate</td>
<td>Loss of metal co-factor</td>
</tr>
</tbody>
</table>
3.1.4.46
Hydrolytically removes 5'-nucleotides successively from the 3'-hydroxy termini of 3'-hydroxy-terminated oligonucleotides

3.1.4.1

3.1.4.44

3.1.4.43

3.1.4.46
Known structure with bound cognate ligand shows active site located in single domain; second domain not contributing to functional change

4.6.1.13
phospholipid

4.6.1.14
sphingolipid

Difference in mechanism & substrate

Difference in multi-domain architecture & substrate

Loss of metal co-factor

Not in archshema as not in reviewed uniprotkb
Not in Funtree as filtered out by sequence similarity

* E.C. Number

Substrate

Multi-domain Architecture

G1

G2

G3
Phosphatidylinositol-Phosphodiesterase Superfamily

Eukaryotes (clade 1)

Bacteria (clade 2)

Spider venom (clade 3)

Glycerophosphodiester

1-phosphtidyl-1D-my-o-inositol

Sphingomyelin
Phosphatidylinositol-Phosphodiesterase Superfamily

Legend
Spider venom (Clade 3) – red
Eukaryote (Clade 1) – blue
Enzyme Domains & Superfamilies

To test we started with an analysis of 6 superfamilies (based on SFLD database from Babbitt group):

- Haloacid dehalogenase
- Terpene Cyclases
- Amidohydrolase
- Crotonase
- Enolase
- Vicinal Oxygen Chelate

Now we have processed 276 Superfamilies

The superfamilies were chosen using MACiE to identify domains with known catalytic residues.
Data Overview

The number of E.C. Codes within a superfamily

The number of ligands within a superfamily
Changes in enzyme function:

• Which changes in enzyme function are observed?

• At which level of E.C. Code?

• How do we represent these changes?

E.C. Exchange Matrix
### E.C. Changes Using Phylogenetic Trees

#### Percentage of changes (total number of counts)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E.C. 1</td>
<td>36.3%</td>
<td>1.3%</td>
<td>0.8%</td>
<td>0.6%</td>
<td>0.7%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td>(1212)</td>
<td>(42)</td>
<td>(28)</td>
<td>(19)</td>
<td>(24)</td>
<td>(7)</td>
</tr>
<tr>
<td>E.C. 2</td>
<td>12.6%</td>
<td>1.6%</td>
<td>1.6%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>0.2%</td>
</tr>
<tr>
<td></td>
<td>(421)</td>
<td>(52)</td>
<td>(52)</td>
<td>(18)</td>
<td>(16)</td>
<td>(8)</td>
</tr>
<tr>
<td>E.C. 3</td>
<td>25.1%</td>
<td>0.8%</td>
<td>0.5%</td>
<td>0.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(837)</td>
<td>(27)</td>
<td>(17)</td>
<td>(8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.C. 4</td>
<td></td>
<td>9.4%</td>
<td>1.0%</td>
<td>0.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(314)</td>
<td>(33)</td>
<td>(13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.C. 5</td>
<td></td>
<td></td>
<td>2.4%</td>
<td>0.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(79)</td>
<td>(4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E.C. 6</td>
<td></td>
<td></td>
<td></td>
<td>3.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(112)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Total Number within class changes**: 2967 (89%)
- **Total Number between class changes**: 360 (11%)
CONCLUSIONS

- New functions emerge by local domain evolution and domain fusions
- Evolution of enzyme function occurs within most superfamilies
  - Changes within a class dominate – i.e., changes of specificity
  - Changes between EC primary classes do occur, but much more rarely – some changes are more common than expected
- Small number of families cover majority of reactions
  - Small no. of primordial enzymes sufficient for life?
- Most changes in reaction chemistry are observed in very distantly related enzymes (ancient changes?)
  - Changes in specificity at leaves of trees
  - Changes in reaction chemistry at ‘root’ of trees
Challenges for the PDB (from Gerard)

• Growth
  • Number, size, complexity of entries
  • Hybrid, low-resolution methods
  • From molecular to cellular structural biology
  • User base!

• Validation

• Integration

• From structural biology archive to biomedical resource
  • Best-practice models versus published models
  • New ways of accessing and using structural information
Growth of EBI Databases 2000-2010*

All resources are growing rapidly

Data doubling every 5 months

12 petabytes data storage

CHALLENGE: DATA => KNOWLEDGE
More Data

- Structural data:
  - More data
  - RNA
  - Membrane proteins
  - Protein complexes
  - FEL Data (Dynamics)

- Other data
- Integration of data
- ??

NGS Data
Human Variation Data
Links to disease phenotypes
HT Cell Biology
HT Light microscopy
EM Tomography
DNA in 3D
Large protein machines

Uroporphyrinogen decarboxylase (1uro)
Heme biosynthesis pathway
Porphyria cutanea tarda
Data Integration: PDB Sequences
SIFTS

Used by:
- wwPDB
- UniProt
- Pfam
- PDBe
- RSCB
- SCOP
- CATH
- PDBsum
- …
Plea for more functional data in PDB to facilitate knowledge extraction:
Capturing knowledge learnt from structure into the PDB, using agreed standards, vocabularies and ontologies:

**Simple things:**
- Experimental protocols
- Function of protein
- Function of ligand e.g. inhibitor/crystallisation aid
- Functional highlights of structure – biological consequences
- Role of dynamic movement
- Relationship to other structures in PDB

**More complex:**
- Protein localisation
- Catalytic site for enzyme
- Binding site for receptor
- Mechanism of enzyme
- Effects of Mutations
- Interaction partners/pathway context
- Disease relationships
THANKS to

- All Structural Biologists, who deposit in PDB
- Original Founders of PDB
- Current and past leaders of PDB
- All staff of wwPDB