PDBe education activities

• Part of EMBL-EBI training program
  • Mainly for graduate/researchers
  • Webinar, e-learning resources, YouTube channel, workshops – half or full day

• Schools project –
  • Collaboration with local art societies and schools in Cambridge, and beyond,
  • Explore the biomolecular world through the creation of artwork
  • Dissemination of this knowledge - calendar, exhibitions
Semantic crowdsourced annotation

Sameer Velankar
How do we generate new knowledge?

Prior knowledge → Curiosity → New questions → Experiment → New data →

Synthesis and interpretation → New model or hypothesis → Predictions →

Publish results
From data to knowledge

Molecular data
Information
Knowledge

Mutation X disrupts enzyme function, which causes Y disease
Global scientific output doubles every 9/10 years

http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html
Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references; Bornmann and Mutz; DOI: 10.1186/1479-7364-5-1-17
What the papers say: Text mining for genomics and systems biology
Molecular and cellular structures – exciting times

- CLEM
- 3D-SEM
- SXT
- ET
- SAXS
- EM
- X-ray
- NMR
- EM

Organism - Virus - Complex - Molecule - Chemical Entity

- Infected Cell
- Assembly
- Chains
- PDB (1971)
- EMDB (2002)
- EMPIAR (2014)
wwPDB mission

• Manage and provide access to the PDB archive
  • Follows FAIR data principle: findable, accessible, interoperable and reproducible (Wilkinson et al. (2016) Scientific Data, DOI: 10.1038/sdata.2016.18)
  • Deposition, validation and biocuration of structure data
  • Establish common data standards
  • Data quality – (Validation) Task Forces
  • Develop policies and procedures
Bringing structure to biology

Molecular data

Information

Knowledge

Mutation X disrupts enzyme function, which causes Y disease

“Coordinates by themselves just specify shape and are not necessarily of intrinsic biological value, unless they can be related to other information”

*Integrative database analysis in structural genomics, Mark Gerstein, Nature Structural Biology 7, 960, 2000*
Infected Cell

Virus Assembly

Complex Chains

Molecule

Chemical Entity

- Cellular organelle
- Organism
- Strain/variant
- Carbohydrate chain
- Immune system evasion

- Species
- Tissue/cell type
- Protein-protein interface
- Tertiary protein assembly
- Proteins bind and recognise cell surface
- Membrane bound protein
- Quality of Structure
- Chirality
- Fit to data
- Environment

Annotations and Biological context
Annotations in the PDB archive

Sequence and Taxonomy

Stereochemistry

Chemical nomenclature

Binding site

Assembly

Validation

PDBe
Data/information visualisation tools

Biological Context

Molecular and cellular structures

Data integration

Knowledge

Display Annotation

Interactive viewers

Data/information visualisation tools
Annotations and Biological context

- Cross-reference information
  - UniProt, Pfam, InterPro, GO, SCOP, CATH, PubMed, NCBI taxonomy

- Value-added information
  - Gene names, homologous protein information

- Standardisation of names and synonyms for each macromolecule
Knowledgebase

Wikipedia - A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems which were the first knowledge-based systems.

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.

Provide information in text format but better categorized—similar to articles.
Scientific articles are tailored to present information in human-readable aliquots. Although the Internet has revolutionized the way our society thinks about information, the traditional text-based framework of the scientific article remains largely unchanged.

Publishing perishing? Towards tomorrow's information architecture
Michael R Seringhaus and Mark B Gerstein
BMC Bioinformatics 2007, 8:17 doi:10.1186/1471-2105-8-17
A Future Vision

The optimal information architecture for biology would capture a broad range of data in digital format and facilitate database deposit alongside manuscript publication. It would index all full-text journal articles, associate keywords and identifiers with database records, and link textbooks, laboratory Web sites and high-level commentary. It would provide multiple levels of peer-review, community comment and annotation, and search results tailored to individual user profiles. This vast network of information would be interrelated, linked and accessed via a single seamless portal.

Publishing perishing? Towards tomorrow's information architecture
Michael R Seringhaus and Mark B Gerstein
BMC Bioinformatics 2007, 8:17 doi:10.1186/1471-2105-8-17
Author / Journal Info
- AuthorList: Kawasaki L et al.
- Title: The Gbeta(KISte4p) subunit of...
- Journal: Yeast 22(12) 947-956
- Date: 2005-09

Results
- **K. lactis** (species)
  - **KISte4** (gene)
    - **KISte4p** (protein)
      - **CLONED**
        - Available at...
      - **SEQUENCED**
        - Sequence ATGTAATAGGCTAC....
      - **MUTANTS**
        - **DELETION**
          - **FUNCTIONAL ASSAYS**
            - Sterile in both MATa and MATα
            - No defect in vegetative growth
          - **STRAIN INFORMATION**
            - Available at...
        - **INTERACTIONS**
          - **TWO-HYBRID**
            - KISte4p
            - Control / no partner
            - KISte4p*
            - KISte4p2p
            - ScGpa1p
Potato salad

Potato salad is a dish made from boiled potatoes and a variety of other ingredients. It is generally considered a side dish, as it usually accompanies the main course. [Wikipedia]

Nutrition Facts

Potato salad

<table>
<thead>
<tr>
<th>Amount Per 1 cup (250 g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calories 357</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>% Daily Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Fat 20 g</td>
<td>30%</td>
</tr>
<tr>
<td>Saturated fat 3.6 g</td>
<td>18%</td>
</tr>
<tr>
<td>Polyunsaturated fat 9 g</td>
<td></td>
</tr>
<tr>
<td>Monounsaturated fat 6 g</td>
<td></td>
</tr>
<tr>
<td>Cholesterol 170 mg</td>
<td>56%</td>
</tr>
<tr>
<td>Sodium 1,323 mg</td>
<td>55%</td>
</tr>
<tr>
<td>Potassium 635 mg</td>
<td>18%</td>
</tr>
<tr>
<td>Total Carbohydrate 28 g</td>
<td>9%</td>
</tr>
<tr>
<td>Dietary fiber 3.2 g</td>
<td>12%</td>
</tr>
<tr>
<td>Protein 7 g</td>
<td>14%</td>
</tr>
<tr>
<td>Vitamin A 7%</td>
<td>Vitamin C 41%</td>
</tr>
<tr>
<td>Calcium 4%</td>
<td>Iron 8%</td>
</tr>
<tr>
<td>Vitamin D 0%</td>
<td>Vitamin B-6 20%</td>
</tr>
<tr>
<td>Vitamin B-12 0%</td>
<td>Magnesium 9%</td>
</tr>
</tbody>
</table>

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs.

1. Bring a large pot of salted water to a boil. Add potatoes and cook until tender but still firm, about 15 minutes. ...
2. Place eggs in a saucepan and cover with cold water. ...
3. In a large bowl, combine the potatoes, eggs, celery, onion, relish, garlic salt, celery salt, mustard, pepper and mayonnaise.

Old Fashioned Potato Salad Recipe - Allrecipes.com
allrecipes.com/recipe/16729/old-fashioned-potato-salad/
Structured data markup for web pages
Classic potato salad

**Nutrition facts:**
- **144 kcal**

**Ingredients:**
- **800g small new potato**
- **3 shallot**

Team with Christmas leftovers or summer BBQ favourites. Either way, Matt Tebbutt’s Classic potato salad is hard resist.
Classic potato salad

Nutrition facts:
- 144 kcal

Ingredients:
- 800g small new potato
- 3 shallot

Title

Recipe

Nutrition
Classic potato salad

Nutrition facts:
- **Calories:** 144 kcal

**Ingredients:**
- 800g small new potato
- 3 shallot

Structured data markup for web pages

RDFa
JSON-LD
Microdata

<div itemscope itemtype="http://schema.org/Recipe">
  <h1 itemprop="name">Classic potato salad</h1>
  <div itemprop="nutrition" itemscope itemtype="http://schema.org/NutritionInformation">
    Nutrition facts:
    <span itemprop="calories">144 kcal</span>,
  </div>
</div>

Ingredients:
- *recipeIngredient*>800g small new potato</span>
- <span *recipeIngredient*>3 shallot</span>
Bringing structure to biology

“Coordinates by themselves just specify shape and are not necessarily of intrinsic biological value, unless they can be related to other information”

*Integrative database analysis in structural genomics, Mark Gerstein, Nature Structural Biology 7, 960, 2000*
Molecular and cellular structures – exciting times

- CLEM
- 3D-SEM
- SXT
- ET
- SAXS
- EM
- X-ray
- NMR
- EM

Organism
- Infected Cell
- Virus
- Assembly
- Complex
- Chains
- Molecule
- Chemical Entity

PDB (1971)
EMDB (2002)
EMPIAR (2014)

PDBe
EMBL-EBI
3DSEM reconstruction
HIV virion in infected macrophages

Tomographic reconstruction
the features of individual SIV virus particles can be identified

3D rendering of data from EMDB
entry 5018; Liu et al., 2008

Fitted atomic model
structure of a HIV-1 gp120 trimer;
PDB entry 3dno; Liu et al., 2008
The hierarchy within each PDB entry

- **File contents**
- **Chains**
- **Residues**
- **Atoms**
- **Quaternary structures**
- **Unique molecules**
- **Chains**
Acetylcholinesterase

Key Enzyme in the Nervous System

Solution of the three-dimensional (3D) structure of *Torpedo californica* acetylcholinesterase (TaChE) in 1991 opened up new horizons in research on an enzyme that had already been the subject of intensive investigation.[1] The unanticipated structure of this extremely rapid enzyme, in which the active site was found to be buried at the bottom of a deep and narrow gorge, lined by 14 aromatic residues (colored dark magenta), led to a revision of the views then held concerning substrate traffic, recognition and hydrolysis.[2] To understand how those aromatic residues behave with the enzyme, see Flexibility of aromatic residues in acetylcholinesterase. Solution of the 3D structure of acetylcholinesterase led to a series of theoretical and experimental studies, which took advantage of recent advances in theoretical techniques for treatment of proteins, such as molecular dynamics and electrostatics and to site-directed mutagenesis, utilizing suitable expression systems. Acetylcholinesterase hydrolyzes the neurotransmitter acetylcholine (ACh), producing choline and an acetate group. ACh directly binds Ser200 (via its nucleophilic Oγ atom) within the catalytic triad (Ser200, His440, and Glu327) (ACh/TaChE structure 1eas). The residues Trp84 and Phe330 are also important in the ligand recognition.[3] After this binding acetylcholinesterase hydrolyzes ACh. See also: ACHE inhibitors and substrates

Cholinesterase
Acetylcholinesterase: Treatment of Alzheimer's disease
Acetylcholinesterase (Hebrew)
Quips – Acetylcholinesterase: A gorge-ous enzyme

Additional Resources

For additional information, see:
Alzheimer's Disease
ACHE inhibitors and substrates

External Links

- Acetylcholinesterase Tutorial by Karl Oberholser, Messiah College

Display Interactive Model
Torpedo californica AChE (PDB code 1eas)
Solution of the three-dimensional (3D) structure of *Torpedo californica* acetylcholinesterase (TcAChE) in 1991 opened up new horizons in research on an enzyme that had already been the subject of intensive investigation.[1] The unanticipated structure of this extremely rapid enzyme, in which the active site was found to be buried at the bottom of a deep and narrow gorge, lined by 14 aromatic residues (colored dark magenta), led to a revision of the views then held concerning substrate traffic, recognition and hydrolysis.[2] To understand how those aromatic residues behave with the enzyme, see *Flexibility of aromatic residues in acetylcholinesterase*. Solution of the 3D structure of acetylcholinesterase led to a series of theoretical and experimental studies, which took advantage of recent advances in theoretical techniques for treatment of proteins, such as molecular dynamics and electrostatics and to site-directed mutagenesis, utilizing suitable expression systems. *Acetylcholinesterase* hydrolyzes the neurotransmitter *acetylcholine* (ACh), producing choline and an acetate group. ACh directly binds Ser200 (via its nucleophilic Oy atom) within the catalytic triad (Ser200, His440, and Glu327) (ACh/TcAChE structure 2ace). The residues Trp84 and Phe330 are also important in the ligand recognition.[3] After this binding acetylcholinesterase hydrolyzes ACh. See also:

### Flexibility of aromatic residues in the active-gorge of AChE

The high aromatic content of the deep and narrow active-site gorge of acetylcholinesterase (AChE) is a remarkable feature of this enzyme. There are 14 conserved aromatic amino acids lined along the gorge of *Torpedo californica* AChE (TcAChE), F120, F288, F290, F330, F331, W84, W233, W279, W432, Y70, Y121, Y130, Y334, and Y442. The side-chain conformational analyses based on the multiple available crystal structures and molecular dynamics (MD) simulation trajectories show that the degree of flexibility of these 14 aromatic side chains is diverse. While those of F330 and W279 are both very flexible, the side-chain conformations of F120, W233, W432, Y70, Y121, F288, F290 and F331 appear to be fixed. Residues located on, or adjacent to the omega-loop (C67-C94), viz. W84, Y130, Y442, and Y334, display different flexibilities in the MD simulations and in the crystal structures.
Information that can be tagged on proteopedia

- Biological entities - proteins, nucleic acids, large assemblies
- Relationship between different states of biological entities (next slide)
- Biological assemblies – interface residues
- Chemical entities – inhibitors, substrate, cofactor, product, drug
- Ligand binding site – binding site residues and role of different residues
- Role of specific residues (stabilization of residue with disallowed Ramachandran angles)
- Disease association
- Appropriate reference for each tagged information
Multiple states of macromolecular machines

1 dataset → multiple coordinate sets

Nguyen et al., 2016, Nature, 530, 298-302

b) In silico classification

c) Multiple maps and models
How to make these data accessible?

External users
Jmol, PyMOL, JalView, ...

PDBe production process
images, entry pages, ...

EBI resources
Reactome, UniProt, ...

Coordinate selection

PISA

EMDB

PDB

SIFTS

Validation

REST API
FunPDBe – community-driven project to provide functional annotations

- Bioinformatics and Biological Resource Fund project (3 years)
- Identify predicted annotations from collaborating groups
- Develop standards for representing annotation data
- Implement a central resource for data delivery
FunPDBe – community-driven project to provide functional annotations

- Year 1 – Functional-site prediction (Prof. Christine Orengo; co-PI)
- Year 2 - Annotating known functional sites and biological assemblies (Prof. Janet Thornton; co-PI)
- Year 3 - Effects of genetic variants on structure/function of macromolecules (Prof. Mike Sternberg; co-PI)

- Collaborating groups - Blundell, Barton, Martin, Fraternali, Levy, Teichmann, Wass, Mitchell, Al Lazakani
Data selection
Data selection – “Enzyme: Hydrolases”
Resolution and Rfactor (Best structure)

- Dibenzothiophene desulfurization enzyme B
  - Resolution 1.8 Å
  - $R_{\text{work}} = 0.18$, $R_{\text{free}} = 0.2$
  - Wild type enzyme

- Dibenzothiophene desulfurization enzyme B
  - Resolution 1.8 Å
  - $R_{\text{work}} = 0.17$, $R_{\text{free}} = 0.2$
  - Mutant in complex with substrate

2de2

2de4
Data selection – “best structure”

Protein: (Neo)pullulanase

2z1k  Crystal Structure of Ttha1563 from Thermus thermophilus HB8

Niwa H, Kuramitsu S, Matsunaga E, Yokoyama S, RIKEN Structural Genomics/Proteomics Initiative (RSGI), Shimada A

Source organism: Thermus thermophilus HB8 [300852]
Assembly composition: polysaccharide/protein complex
Interacting compound: P04

Protein: (S)-2-haloacid dehalogenase

1qq5  STRUCTURE OF L-2-HALOACID DEHALOGENASE FROM XANTHOBABER AUTOTROPHICUS

Dijkstra BW, Ridder IS, Rozeboom HJ, Kalk KH

Source organism: Xanthobacter autotrophicus [280]
Assembly composition: protein only structure
Interacting compound: FMT

Other entries (7)
Data selection – “best structure”
Data presentation
3D images – most used visualisation

• Where is “my protein” in this assembly?

• How many unique macromolecules are there in this complex?

• What is the overall shape of the complex?

• Where is this sequence or structure domain in the 3D structure?
Interactive web components

- Web components for visualization of PDB and related protein structural data
- PDB Topology Viewer
- Sequence feature View
- UniPDB Viewer
- PDB_REDO
- PDB Residue Interactions
- PDB 3D Complex

http://www.ebi.ac.uk/pdbe/pdb-component-library/
Data access mechanism

Archive → Dynamic selection and transfer → Users

**PDB**

Limited Bandwidth

Interactive visualization

Data file

Coordinates: **PDBe Coordinate Server**
Annotations: **PDBe REST API**
Electron density maps: **PDBe Density Server**
Compression: **MMTF (RCSB PDB)**
Visualization: **PDB Component Library and 3D viewers**
LiteMol

Browser-based 3D visualisation plug-in

Fast – quickly visualise even the largest structures
Small – 400 kB and no other dependencies
Open Source – written in JavaScript

Common 3D atomic models, assemblies, symmetry
Density maps and cryo-EM data
Validation data, sequence annotation
Solution

- **CoordinateServer**: 3D coordinate data
- **DensityServer**: Experimental data (density)
- **PDBe**: Annotation API
- **BinaryCIF**: Data compression
- **LiteMol**: Visualisation
CoordinateServer + BinaryCIF

Size of HIV-1 Capsid (2.44M atoms, PDB id 3j3q) in MB, gzip compressed

- CIF: 41.78 MB
- BinaryCIF: 12.17 MB
- BinaryCIF "cartoons": 1.54 MB
DensityServer + BinaryCIF (down-sampling)

Cryo-EM structure
Zika virus

1.6 GB
CIF
CCP4
DensityServer

1.03 MB
BinaryCIF
Map down-sampling: Zika virus (EMD-8116)

- Level 1 - compressed from 1.6GB gzip to 9MB gzip
- Level 2 - compressed from 1.6GB gzip to 1MB gzip
Visualising model, biological context and experimental evidence

- 3D structure
- Electron density
- Annotation (ligand validation)

Full files: 20,313 KB
LiteMol: 40 KB

508× less
LiteMol – new WebGL-based viewer

- Does not require Java
- Reads mmCIF files, CCP4 maps and EMDB maps
- Basic support for PDB format
- Integration with PDBe API
- Support for large structures
  - Delivery of coordinates, ligand environment
Interactive 1D, 2D and 3D views

pdbe.org/4tpk