

## The wonderful world of structure archiving

### What's happening and what's next?

Gerard Kleywegt, PDBe, EMBL-EBI

Winter School on Structural Biology, CEITEC, Brno, 13 February 2015

Summary...

More! Bigger!

Blobbier!

Better! Cooler!

### Outline

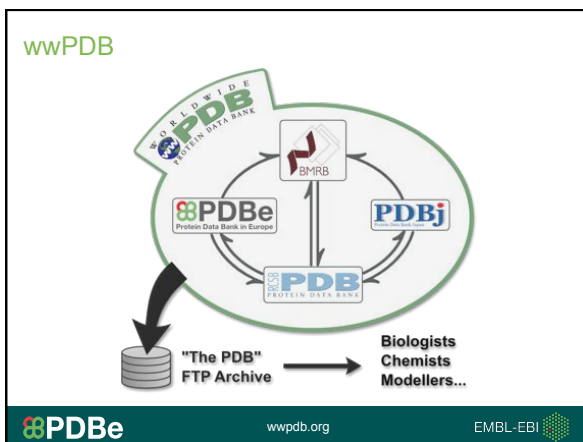
- PDB in 2014/2015
  - More! (100,000+ entries)
  - Bigger! (Large structures & "Formageddon")
  - Better! (D&A & validation)
- PDBe in 2015
  - Cooler! (Nifty new things coming this year)
- The future
  - Blobbier! (Cellular imaging and hybrid methods)

### Who's who again?

PDBe, wwPDB, EMDataBank

### PDBe at a glance

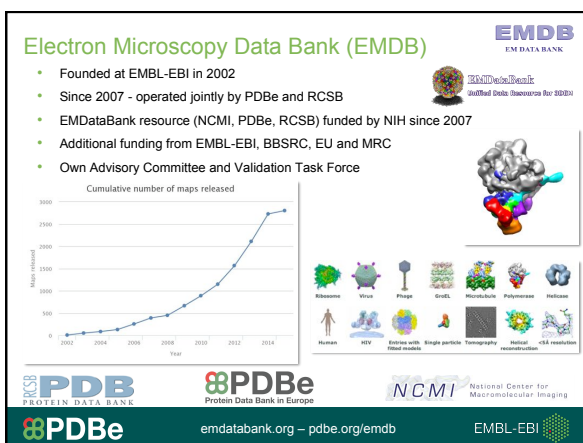
- Mission: **Bringing Structure to Biology**
- Founding partner of Worldwide Protein Data Bank (wwPDB)
- Birthplace of Electron Microscopy Data Bank (EMDB)
- Founding partner of EMDataBank
- Major activities:
  - Deposition and annotation site for structural data on biomacromolecules & complexes (X-ray, NMR, EM)
  - Integrated resource to serve structural data and information
  - Liaise with structural biology community
- Guided by advisory bodies made up of community experts
  - PDBe, wwPDB and EMDataBank advisory committees
  - (Validation) Task Forces (method-specific) & *ad-hoc* working groups



wwPDB partnership

- Collaborate on "data in"
  - Policy issues
  - Weekly releases
  - Validation standards
  - Format specifications
  - Chemical Component Dictionary
  - Deposition and annotation procedures
  - Archive quality and remediation
  - Journal interactions
  - Community interactions
- Friendly competition on "data out"
  - Serving PDB data with added-value
  - PDB-based services
  - Other services, resources and activities

PDBe  
wwpdb.org



Roles of PDBe, wwPDB and EMDataBank

Roles	PDBe	wwPDB	EMDataBank
Community interactions	CCP4, CCPN, CCP-EM	(V)TFs, IUCr, journals, ...	EM-VTF, workshops, portal
Challenges	CAPRI, CASD-NMR	(CASP)	EM modelling
Formats	(3D cellular imaging data?)	PDB, PDBx, working groups	Maps, FSC, segmentations
Data models & ontologies	Crystallisation ontology, CCPN	PDBx	EMDB data model, EMX
New methods	(SXT? 3DSEM? CLEM?)	SAS? Hybrids?	(?)
Deposition, annotation, validation, archiving, distribution	(3D cellular imaging archive?)	PDB, BMRB	EMDB
Integration	SIFTS and more	PDB annotation	EMDB annotation
Advanced services exposing structural information	Many!	-	-

PDBe  
Gutmanas et al., Acta Cryst. D69, 710 (2013)  
EMBL-EBI

PDB in 2014/2015

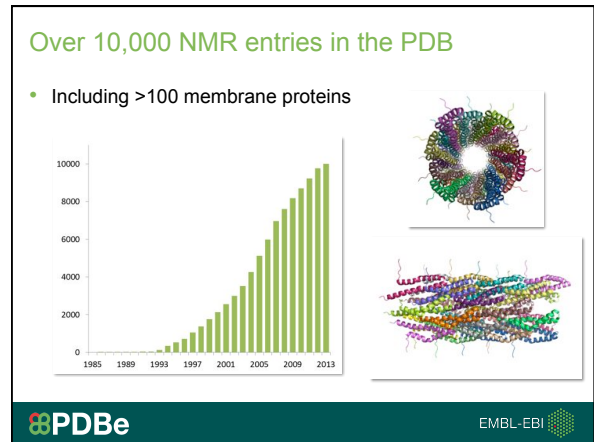
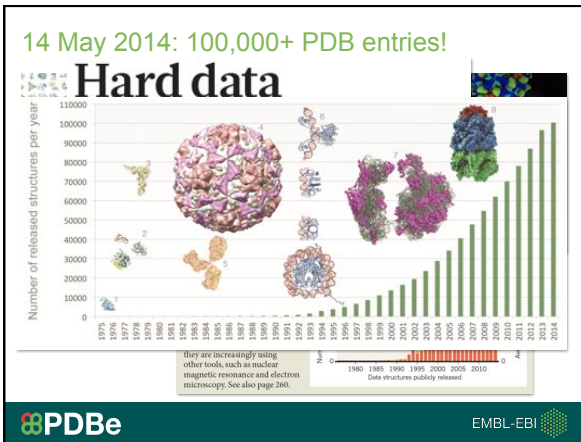
More! Bigger! Better!

PDBe  
EMBL-EBI

PDB in 2014/2015

- More!
  - Over 100,000 entries in the active archive
  - More than 10,000 of these are NMR entries
- Bigger!
  - Large structures now released intact (not SPLIT)
  - Required move to mmCIF/PDBx ("*Formageddon*")
- Better!
  - New common Deposition & Annotation system
  - Validation reports
  - Archive remediation

PDBe  
EMBL-EBI



### Large structures now released intact

- Limitations of PDB format necessitated SPLIT entries
- mmCIF/PDBx format does not have these limitations
- Workshop at EMBL-EBI in 2011 – decision to support PDBx in major refinement packages and to switch to PDBx as the distribution format for the PDB archive

mmCIF: [wwpdb.org](http://wwpdb.org)

EMBL-EBI

### Large structures to be released intact

- 2013
  - Large structures can be deposited intact (ADIT, AutoDep)
  - Distribution as PDBx and PDBML in separate ftp area, and also as SPLIT entries in regular archive
- 2014
  - New wwPDB Deposition & Annotation system designed to handle large structures
  - July: previously SPLIT entries reunited and distributed in parallel in separate ftp area
  - 10 December: "Formageddon"!
    - Switched to PDBx as archive distribution format
    - SPLIT entries replaced by reunited entries

EMBL-EBI

### mmCIF/PDBx workshop for programmers

EMBL-EBI, November 2013

EMBL-EBI

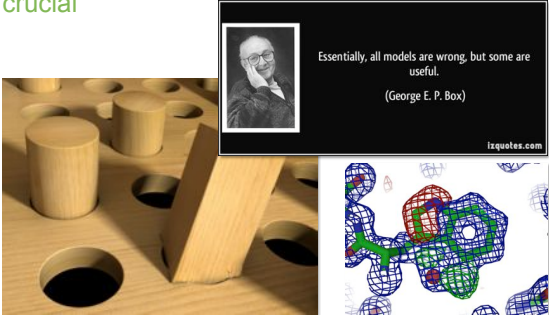
### New Deposition & Annotation system

- Designed to handle large structures from any combination of techniques (X-ray, NMR, EM)
- Jointly developed by wwPDB partners
- Replaces AutoDep, ADIT, EMDep
- Validation integral part of deposition and annotation
- X-ray module in production since January 2014
- X-ray/NMR/EM/neutron coming in 2015

<http://deposit.wwpdb.org/deposition>

EMBL-EBI

### Validation of structural data and models is crucial

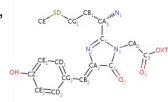


Essentially, all models are wrong, but some are useful.  
(George E. P. Box)

EMBL-EBI


### Archive remediation

- Never-ending process to improve the content, description and consistency of data in the PDB archive
- Incidental: affecting one or a few entries, often acting on information from users
- Archive-wide
  - In the past: literature citations, sequence references, taxonomy assignments, peptide ligands (inhibitors, antibiotics)
  - Future: carbohydrates, protein modifications, metal-containing ligands



EMBL-EBI

### PDBe in 2015



Cooler!

EMBL-EBI

### PDBe in 2015 – lots of cool stuff coming soon!

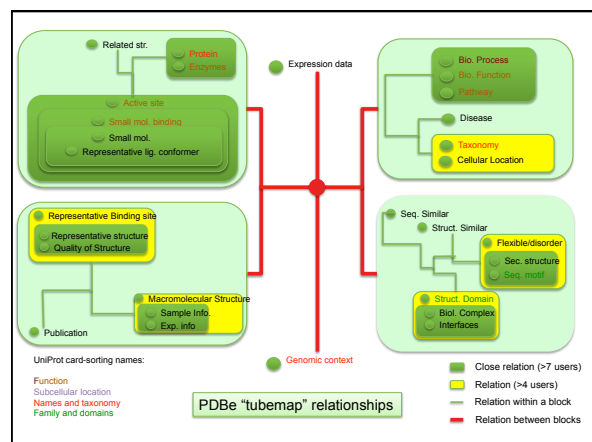
- Cooler!
  - Redesigned PDB and EMDB entry pages, based on extensive user-testing
  - Several unique features to enhance the content of the archives
  - New search/browse system that makes finding things easier and allows for further analysis
  - Redesigned “corporate” web site (*pdb.e.org*)
  - For developers: all our data will be available through an API
  - A bit later: validation portal to make analysing and validating X-ray, NMR and EM structures easier
  - Lots of 3DEM-related activity, too!

EMBL-EBI

### PDBe website needed a make-over

- Many improvements:
  - Quality of underlying data (literature, assemblies, ...)
  - Unique data (e.g., through mining full-text articles)
  - Data access through API (PDB, EMDB, SIFTS, PISA, validation)
  - Data discovery and analysis (search/browse)
  - User-driven redesign of entry pages
  - User interfaces (e.g., linked 1<sup>st</sup>/2<sup>nd</sup>/3<sup>rd</sup> structure viewer)
  - Javascript-based 3D viewers (instead of Java applets)
  - Improved “corporate” website

EMBL-EBI





**3unb**  
 Released: 2012-02-29  
 Last modified: 2012-02-13  
 X-ray diffraction  
 1.6 Å resolution

**Function | Biology**  
 Entry contents: 14 distinct polypeptide chains heterotrimeric 20-mer  
 Biochemical function: hydrolase activity  
 Biological process: response to organic nitrogen  
 Cellular component: proteasome complex  
 Pathway: ubiquitination  
 Disease: described adult disease

**Structure analysis**  
 Domains: Proteasome subunit,  $\alpha$  is terminal signature  
 Proteasome subunit, Proteasome beta subunit, C-terminal  
 Architecture: heterotrimeric bundle  
 3-unb 3unb

### Citation information

- Primary citation: figures with captions if full-text publication
- Reviews/articles that reference primary citation or mention PDB entry in full text (data-mining by Europe PMC)

**PDBe** EMBL-EBI

### Entry images

- Display "preferred" assembly  $\rightarrow$  smallest assembly that contains all entities (i.e., distinct molecules)
- If no such assembly, use the one with the most entities
  - ... and find out if annotation is correct or not

**PDBe** PDB entry 3MIN EMBL-EBI

### Entry/assembly images – virus structures

**PDBe** EMBL-EBI

**PDBe • 2clq**  
 Released: 09 May 2006  
 X-ray diffraction  
 2.18 Å resolution

**Function and Biology**  
 Reaction catalyzed: ATP + protein + ADP + phosphate  
 Biochemical function: transferase activity, transferring phosphate-containing groups  
 Biological process: protein phosphorylation  
 Cellular component: not assigned

**Structure analysis**  
 Assembly: homo dimer (preferred)  
 Composition: 1 distinct polypeptide molecule  
 Entry contents: 1 distinct polypeptide molecule  
 Macromolecule: Mitogen-activated protein kinase kinase kinase B

**Visualisation**  
 Molecule: UniProt, Pfam, Chain A, Sec. Str., CATH  
 + More chains: CATH 3.30.200.20

**2clq:A** Annotations

### PDBe – new search/browse system

- Provides suggestions based on reference data (auto-complete)
  - E.g., enzyme names, GO terms, protein names, CATH classification
- Supports categorisation of result set (“faceting” – like Amazon)
  - By method of structure determination, release date, resolution, ligands, author, CATH domain, ...
- Supports multiple views of the result set
  - Entry view, molecule view (more to come)

### Auto-complete interface

### Search results – entry view

### Search results – entry view

### Search results – macromolecule view

### Search results – compound view

Entries: Macromolecules | **Compounds** | Protein families

Interaction 1 to 19 of 19

Compound: **ACT : ACETATE ION**

Interacting with:

Protein: **Alcohol dehydrogenase class 4 mu/sigma chain**

Best example found in:

**1d1t** MUTANT OF HUMAN SIGMA ALCOHOL DEHYDROGENASE WITH LEUCINE AT POSITION 141

Interacting compounds: ZN ACT CAC NAD

Also occurring in entries (2)

---

Compound: **ACT : ACETATE ION**

Interacting with:

Protein: **Cellular retinoic acid-binding protein 2**

Best example found in:

**2f56** Crystal Structure of Apo-Cellular Retinoic Acid Binding Protein Type II At 1.26 Angstroms Resolution

Interacting compounds: CL ACT NA

Also occurring in entries (5)

---

Compound: **AZE : ALL-TRANS AXEROPHTHENE**

Interacting with:

Protein: **Cellular retinoic acid-binding protein 2**

Only example found in:

**2g7b** Crystal Structure of the R132KR111L121E mutant of Cellular Retinoic Acid Binding Protein Type II In Complex With All-Trans Retinal At 1.18 Angstroms Resolution

### Search results – protein-family view

Entries: Macromolecules | Compounds | **Protein families**

Protein family: **ADH\_N**

Protein family: **ADH\_N**

Occurring in macromolecule:

Protein: **Alcohol dehydrogenase class 4 mu/sigma chain**

Best example found in:

**1d1t** MUTANT OF HUMAN SIGMA ALCOHOL DEHYDROGENASE WITH LEUCINE AT POSITION 141

Interacting compounds: ZN ACT CAC NAD

Also occurring in entries (2)

---

Protein family: **ADH\_cinc\_N**

Occurring in macromolecule:

Protein: **Alcohol dehydrogenase class 4 mu/sigma chain**

Best example found in:

**1d1t** MUTANT OF HUMAN SIGMA ALCOHOL DEHYDROGENASE WITH LEUCINE AT POSITION 141

Interacting compounds: ZN ACT CAC NAD

Also occurring in entries (2)

---

Protein family: **Lipocalin**

Occurring in macromolecule:

Protein: **Cellular retinoic acid-binding protein 2**

Best example found in:

**2g7b** Crystal Structure of the R132KR111L121E mutant of Cellular Retinoic Acid Binding Protein Type II In Complex With All-Trans Retinal At 1.18 Angstroms Resolution

Interacting compounds: AZE NA

Also occurring in entries (18)

### PDBe – new “corporate” web site

Protein Data Bank in Europe

Home | Search | Structure | Biology | Services | Research | Training | About PDBe

News | Publications | Events | PDBe: Protein Data Bank in Europe

Featured structure: Crystal structure of hemagglutinin from an H5N1 influenza virus, PDB: 5D3G

News: PDB passes 100,000 structure milestone (28 May 2014), PDBe to incorporate OpenEye Cheminformatics software (20 May 2014), The road to 100,000 entries: building a community resource (5 Aug 2014 to 12 Aug 2014)

EMBL-EBI

### The PDBe API

- Application Programming Interface
- PDB, EMBD, CCD, SIFTS, PISA, SSM, validation, topology, search system
- Used to populate new PDBe entry pages
- Available to external developers

EMBL-EBI PDBe REST API

Programmatic access to PDBe data

REST calls based on PDB entry data

Summary: <http://www.ebi.ac.uk/pdbe/api/pdb/entry/summary/pdbid>

Molecules in the entry (alias /entry/entities): <http://www.ebi.ac.uk/pdbe/api/pdb/entry/molecules/pdbid>

EMBL-EBI Info: sameer@ebi.ac.uk

### Delivering, analysing and validating experimental data at PDBe

EMBL-EBI

### EMDB volume viewer

Map: Microtubules co-polymerized with doublecortin

Model(s): **2XRP**

Map Controls: Map of Solid Surface, Level: 0.7, Opacity: 0.5, Color: Yellow, Map(s) of 2XRP, Background: Light Grey

View Controls: Size (Å): 10, Depth (Å): 10

EMBL-EBI pdbe.org/emd-1788



### EMDB visual map analysis

Visual analysis of map: EMD-1788 - Doublecortin-stabilized microtubule at secondary structure resolution  
PDB model: 2Z97

The view of atomic maps provides information related to each 3D volume related to the EMDB. The atomic maps are shown in a 3D view, allowing the user to interactively explore the data. The map is shown in a 3D view, allowing the user to interactively explore the data. The map is shown in a 3D view, allowing the user to interactively explore the data.

Map density distribution

Atom induction

EMBL-EBI

pdbe.org/emd-1788

### Slice viewer for tomograms

Collaboration with OME (Dundee)

EMD-1053 - Untangling disordered kinets with electron tomography.  
Substructure: 3AM  
Resolution: 2.8  
PDB model: 1A55

EMBL-EBI

pdbe.org/emd-1053

### OLDERADO

- Helps you analyse NMR ensembles
- How many clusters?
- Representative models?
- Which rigid domains?
- pdbe.org/olderado

OLDERADO cluster and domain composition for PDB entry 2k4v

Summary of Oldero results

Most representative model: 8

Largest domain composition residues: Chain A: 5-13, 21-85, 95-108

RMSD of all models in ensemble: 7.100 Å

EMBL-EBI

### Vivaldi

pdbe.org/vivaldi

Vivaldi: Visualisation, analysis and validation of NMR entries

PDB ID: 1U6

Model selection

EMBL-EBI

### Vivaldi++ (design study)

Structure validation for PDB entry 2k4v

Interactive 3D viewer

Model selection

Charts

Summary

More graphs

EMBL-EBI

### Prototype/Mock-up - "EDS+"

Residue in focus

Full-entry view

Residue-centric view with electron density maps

Overall quality

Residue-level quality

Ramachandran plot

Residue group selector

EMBL-EBI



### Ligand-validation prototype

Item	Value	Mean	StdDev	Stdev	Z
BOND CAPCAD	1.476	1.361	0.000	1	0.86
ANGLE CACACACAO	116.848	112.817	0.395	31	+3.30
ANGLE CACACACAO	116.837	112.800	0.398	31	+3.34
BOND CAGGLAC	1.787	1.738	0.000	10000	+2.33
ANGLE CAPCADICAO	124.843	128.038	1.485	7	-2.07
ANGLE CACACACAO	102.888	112.650	3.738	383	-3.84
ANGLE CAGCAGCAT	122.235	116.595	3.148	30	+1.79
ANGLE CACACACAO	122.825	118.177	1.763	10000	+1.38
BOND CAGCAG	1.526	1.484	0.016	173	+0.22
BOND CALGAR	1.398	1.375	0.021	10000	+1.18
BOND CAGCAG	1.398	1.375	0.021	10000	+1.00
ANGLE CAGCAGICAO	118.838	121.253	1.708	329	-3.94
ANGLE CAGCAGCAG	118.881	121.072	1.831	10000	-0.91

### PDBe - validation portal

- Unified delivery of experimental data and validation information for X-ray, NMR and EM
- Visualisation (1D, 2D, 3D, linked)
- Selections
- Terminology
- Look-and-feel

### PDBe - validation portal

- If you can use one, you can use them all!
- Lowers barrier for non-experts
- Interactive pages will include
  - Interactive 3D viewer
  - Graphs, plots and tables, tightly coupled to 3D viewer
- Static entry pages will provide intermediate level of detail

(Bob Hanson of Jmol/JSmol fame)

### The future

Blobbier!

### The world is changing

- Biology
- Structural biology
- Bioinformatics
- ICT
- Funding landscape
- Emerging nations
- How about structural biology archives?

### Challenges facing archives

- Increasing size and complexity of structures and data
- More heterogeneous information at a range of scales
- Need to coordinate across disciplines
- Need to integrate structural data on scales from atoms to cells
- Need to integrate structure with other biological and chemical data
- Need to deliver appropriate structure data to non-experts (in context of their work)
- (Funding)

### Structural biology archives today

Technique	Models	Data
X-ray	PDB	PDB
NMR	PDB	BMRB + PDB
3DEM	PDB	EMDB

- Simple world
- 3 techniques
- 3 archives
- atomistic models

**Trends**

- Many techniques, ranging in scale from atoms to cells
- Many types of model: atomistic, "residue-istic", map segmentations (with contrast), envelopes (no contrast), geometric shapes
- Hybrid methods (integrative modelling): mixed models (atomistic/lobby; experimental/theoretical) and heterogeneous data with variable information content

**PDBe** EMBL-EBI

### Structural biology archives tomorrow?

Technique	Models	Data
X-ray	PDB	PDB
NMR	PDB	BMRB + PDB
3DEM	PDB + "ModelDB" + "BlobDB"	EMDB
SAXS/SANS	PDB + "SASDB"	"SASDB" + PDB
Theoretical models	"ModelDB"	n/a
Hybrid methods	PDB + "ModelDB" + "BlobDB"	"BlobDB"? PDB?
SXT, 3DSEM, ...	"BlobDB"	"3DCellDB"? EMDB?
CLEM	PDB + "ModelDB" + "BlobDB"	"3DCellDB"? EMDB?

- "ModelDB" = theoretical atomistic models?
- "BlobDB" = non-atomistic models and hybrid methods data?
- "SASDB" = SAXS/SANS data and models not archived elsewhere?
- "3DCellDB" = 3D cellular imaging data and segmentations?

**PDBe** EMBL-EBI

### The key will be "integration"

**PDBe** Image: Zeev-Ben-Mordehai et al., 2014 EMBL-EBI

### Integrating imaging and 3D structural data

• Enriching biology with structural information

Atoms   Molecules   Machines   Cells   Samples

← Scales / Methods →

NMR   X-ray   EM   SAXS   ET   SXT   3DSEM   CLEM

**PDBe** Instruct & Euro-Biolmaging EMBL-EBI

### Integrating imaging and 3D structural data

• Annotating and linking structure through biological information

Chemistry   Sequences   Variation   Interactions   Pathways

← Information / Resources →

ChEBI   ENA/UniProt   Ensembl   ChEMBL/IntAct   Reactome

**PDBe** Elixir EMBL-EBI

### Integrating imaging and 3D structural data

- Archive 3D cellular imaging data (EM, ET; later: SXT, 3DSEM, CLEM)
- Annotate using bioinformatics resources, classifications and ontologies (UniProt, GO)
- Link to 3D molecular structure data (X-ray, NMR, EM) in PDB and EMDB
- Provide tools to make the information easily accessible and to facilitate discovery

**PDBe** EMBL-EBI



If you see this slide, I've gone too far