



## The E-MSD Relational Database

An autonomous structural database capability and infrastructure services in Europe



<http://www.ebi.ac.uk/msd>



## The Macromolecular Structure Database (MSD) group

- Based at the European Bioinformatics Institute (EBI), an outstation of the European Molecular Biology Laboratory (EMBL) at Hinxton, UK
- Started in 1996 with the goal of providing an autonomous structural database capability in Europe
- The aims of the group are to provide:
  - a deposition site via which macromolecular structures can be added to the PDB (AutoDep)
  - a stable and clean repository of macromolecular structure data
  - services that allow users to access, search and retrieve structural data



<http://www.ebi.ac.uk/msd>



## Protein Data Bank

- Consists of three sites
  - RCSB (USA), PDB-j (Japan) and ePDB/MSD (UK)
- PDB is the single repository of all publicly available macromolecular structures
- Started in 1971, the archive now has around 44,000 entries and new entries are added weekly
- Structures are deposited by experimentalists and contents is freely available
- Historical format of the archive is flat-files with fixed line format, although an improved flat-file format (mmCIF) is available



<http://www.ebi.ac.uk/msd>



## wwPDB

The screenshot shows the wwPDB website interface. At the top, it says 'WORLDWIDE PDB PROTEIN DATA BANK'. Below that, there are navigation links: Home, wwPDB Charter Agreement, RCSB PDB, ePDB, PDBj, News, Contact Us. The main content area is titled 'wwPDB Members' and features a news item dated '16 Sept 2003' stating that RCSB, EBI and PDBj collaborate to form the worldwide Protein Data Bank (wwPDB) to manage the PDB archive. It also includes sections for 'WHAT IS THE MISSION OF THE wwPDB?', 'WHAT IS THE PDB ARCHIVE?', 'HOW CAN I ACCESS THE PROTEIN DATA BANK ARCHIVE?', and 'WILL THERE BE ANY CHANGES IN THE WAY I DEPOSIT'. The footer includes a 'Done' button and an 'Internet' icon.



<http://www.ebi.ac.uk/msd>



## Biological databases and Informatics

- The National Science Foundation (NSF) believes that future advances in the biological sciences will depend both upon the creation of new knowledge and upon effective management of proliferating information.

<http://www.nsf.gov/pubs/2005/nsf05577/nsf05577.html>



<http://www.ebi.ac.uk/msd>



## Role of Bioinformatics

- To Support Experimental Biology
  - To Collect and Archive Data
  - To provide Framework and Integration
  - To give Easy Access to Data
- To make New Discoveries through Data Analysis



<http://www.ebi.ac.uk/msd>



## Deposition of data



<http://www.ebi.ac.uk/msd>



## Three complex techniques



X-ray  
crystallography:  
synchrotron



cryo-EM:  
Electron  
microscope



NMR:  
High Field  
Spectrometer



<http://www.ebi.ac.uk/msd>



## EMDB

EM DATA BANK

*Maps/volumes* from cryo-electron microscopy are deposited in the EMDb at the EBI.

Depositions started June 2002

**EM Deposition:**

<http://www.ebi.ac.uk/msd-srv/emdep/index.html>

**EM Search:**

<http://www.ebi.ac.uk/msd-srv/emsearch/index.html>



<http://www.ebi.ac.uk/msd>



## EMDB

EM DATA BANK

### EMDB Deposition

#### Data Requirements:

Single MAP (Volume) per deposition

**mandatory**

(most map FORMATS accepted and converted to CCP4 via em2em)

Experimental Details

**mainly optional**

Layer-line data, masks,  
structure factors, images

**optional**



<http://www.ebi.ac.uk/msd>



## Depositions

Full deposition site from June 1999

15% of all submissions via the EBI.

Partnership with RCSB for a single PDB.

All EBI processed data at release date are sent to RCSB for central distribution.



<http://www.ebi.ac.uk/msd>



**Historically, the PDB files are the primary repository of Macromolecular Structure Data**



<http://www.ebi.ac.uk/msd>



## Problems with the legacy archive

- Macromolecular structures are very complex
- Existing PDB format is incapable of fully describing even existing structures
- Format is not readily extensible, to cope, for example, with structural genomics data
- Historical archive is non-uniform and poorly populated
- Search and retrieval of flat files is difficult and/or inaccurate

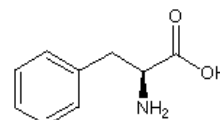


<http://www.ebi.ac.uk/msd>



## PHENYLALANINE

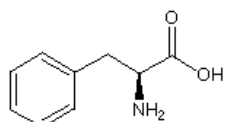
All looks normal – is it?



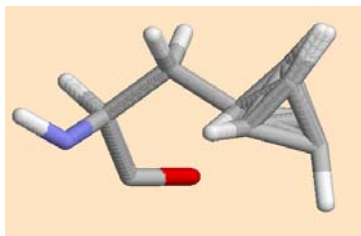
ATOM	2567	N	PHE	B	175	7.821	-25.530	-22.848	1.00	8.71
ATOM	2568	CA	PHE	B	175	8.845	-25.172	-21.877	1.00	9.41
ATOM	2569	C	PHE	B	175	9.449	-23.798	-22.169	1.00	10.02
ATOM	2570	O	PHE	B	175	10.664	-23.613	-22.103	1.00	10.37
ATOM	2571	CB	PHE	B	175	9.928	-26.251	-21.848	1.00	9.53
ATOM	2572	CG	PHE	B	175	10.969	-26.137	-22.982	1.00	10.03
ATOM	2573	CD1	PHE	B	175	12.356	-25.819	-22.988	1.00	10.51
ATOM	2574	CD2	PHE	B	175	11.725	-27.211	-23.402	1.00	10.25
ATOM	2575	CE1	PHE	B	175	11.821	-27.095	-22.869	1.00	11.17
ATOM	2576	CE2	PHE	B	175	12.282	-26.086	-24.008	1.00	10.95
ATOM	2577	CZ	PHE	B	175	10.953	-26.335	-23.622	1.00	11.38



<http://www.ebi.ac.uk/msd>



## PHENYLALANINE Not Quite an Outlier!!



<http://www.ebi.ac.uk/msd>



- Spelling errors abound, e.g. 23 versions of this humble bug
- Legacy format cannot enforce relationships between records, e.g. chain names may be inconsistently named within an entry

```
$COLI
COLI
E. COLI
E.COLI
ESCHERICHIA COLI
ESCHERICHI $COLI
ESCHERICHIA $COLI
ESCHERICHIA COLI
ESCHERICHIA COLI.
EXCHERICHIA COLI
EXPRESCHERICHIA COLI
```

ESCHERICHIA COLI

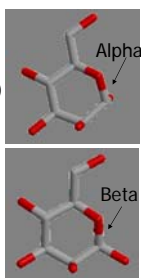


<http://www.ebi.ac.uk/msd>



## Ligand nomenclature

- Ligands are often named inconsistently or even entirely incorrectly, e.g.  $\alpha$ -D-mannose (MAN) vs  $\beta$ -D-mannose (BMA)



<http://www.ebi.ac.uk/msd>



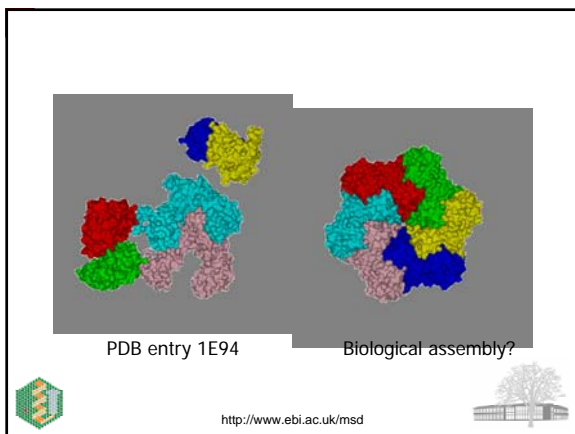
## PDB entry the deposited coordinates usually consist of the contents of the asymmetric unit

- The contents of the ASU define a single copy of the macromolecule.
- The contents of the ASU consist of more than one copy of the macromolecule.
- The contents of the ASU require crystallographic symmetry operations to be applied to generate the complete macromolecule(s).
- A combination of the above, including multiple copies and required symmetry transformations



<http://www.ebi.ac.uk/msd>





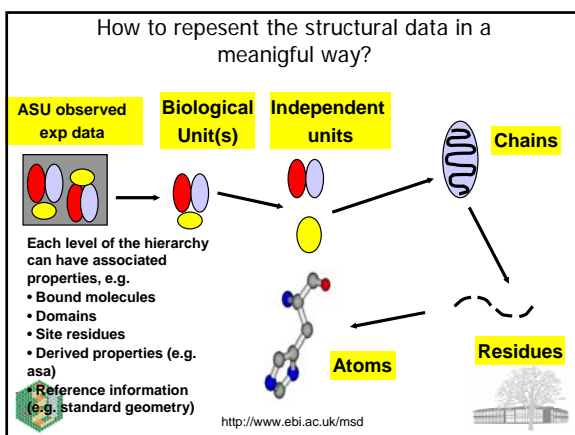
- ## How to improve data quality?
- **Introduce more checks**
    - Authentication of source
    - Authentication of most recent version
    - Validation of correct methodology used
    - Conformity to standards
    - Error checks
    - Consistency checks - to identify simple typos
    - Outlier detection - to identify suspect records
- <http://www.ebi.ac.uk/msd>

- **What happens when these checks fail?**
    - Raise issue with the depositor
  - **But the depositor might:**
    - be unavailable
    - not interested
    - not know the answer anyway
    - not be sure about which data have the problem
  - **The older the entry, the less likely the depositor can/will help**
- <http://www.ebi.ac.uk/msd>

## What is the solution?

Don't rush and define another format  
 Represent the structure data in a meaningful way (use data model)

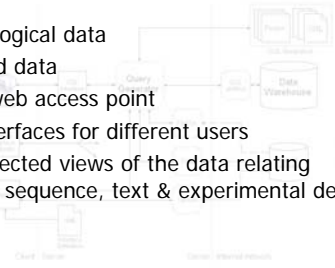
<http://www.ebi.ac.uk/msd>



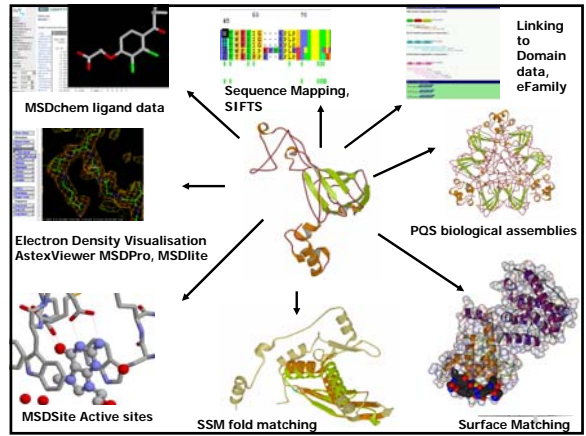
- ## Adopt standards
- Use NCBI taxonomy database to ensure correct organism names
  - Use Uniprot database to ensure correct protein description
  - Enzyme database
  - Annotated ligand information
- <http://www.ebi.ac.uk/msd>

## MSD provides...

- Clean biological data
- Integrated data
- A single web access point
- Query interfaces for different users
- Interconnected views of the data relating structure, sequence, text & experimental details



<http://www.ebi.ac.uk/msd>





EMBL-EBI

## Validation & Structure Quality

Jawahar Swaminathan



EMBL-EBI

## Ground rules for bioinformatics

- Don't always believe what programs tell you  
*they're often misleading & sometimes wrong!*
- Don't always believe what databases tell you  
*they're often misleading & sometimes wrong!*
- Don't always believe what lecturers tell you  
*they're often misleading & sometimes wrong!*
- In short, don't be a naive user
  - when computers are applied to biology, it is vital to understand the difference between mathematical & biological significance
  - computers don't do biology  
- they do sums quickly!



EMBL-EBI

## Validation

- **validation** n
- 1: the act of validating; finding or testing the truth of something [syn: [proof](#)]
- 2: the cognitive process of establishing a valid proof [syn: [establishment](#)]
- Assessing the quality of a model is called validation. Validation is something that needs to be done both by producers (crystallographers, NMR spectroscopists, electron microscopists, etc.) and users (biologists, enzymologists, medicinal chemists, etc.) of models.



EMBL-EBI

## Some Truths

- Never trust a structure at face value.
- Any structure is only as good as the experimental data which goes into its determination.
- Just because it is published in Nature does not mean the structure is not without flaws.



EMBL-EBI

## Errors in Structures

- Completely wrong
  - Wrong trace, incorrect fold of protein
  - Register errors, where trace of protein is not in keeping with sequence order.
- Partial errors
  - Incorrectly built loops.
  - Wrong residues built into the structure (i.e., Proline instead of Aspartic acid).
- Bad data quality
  - Bad geometry and stereochemistry.
  - Incorrect positioning of ligands etc due to lack of experimental evidence.



EMBL-EBI

## Some Quality Indicators

Some data quality indicators for structures are

1. Ramachandran Plot
2. Geometry and Stereochemistry
3. R-factor/FreeR-factor (Structures from X-ray crystallography)
4. Correlation between experimental data and structure
5. Resolution of the data upon which the structure is based (Structures from X-ray crystallography)

EMBL-EBI

## Ramachandran Plot

- A graph between the dihedral angles of an amino acid in a protein.
- Due to steric hindrance from amino acid side chains, only certain angles are allowed in a folded protein.
- A plot between the dihedral angles of individual amino acids in a protein can serve to indicate how well the structure has been determined.
- Any deviations from the allowed values are called Outliers and usually indicate bad geometry

Dihedral Angles

EMBL-EBI

## Ramachandran Plot

The Ramachandran Plot.

Ramachandran Plot p0b4hb5

Standard Plot showing where different secondary structures fit into the plot.

A real life example. All non-glycine residues are in allowed regions.

EMBL-EBI

## Validation

So what do you think about this ?

- Ideally, there should be no outliers in the Ramachandran plot, except for Glycine and Proline, which are "special" amino acids.
- However, there may be some rational explanation for outliers by the scientist depositing the structure. (Always refer to the publication!).
- Expect to find more than 85-90% of residues to fall into the red regions.

EMBL-EBI

## Geometry and Stereochemistry

- This is supposed to be Phenylalanine and should look like:

**BUT....**

EMBL-EBI

## Geometry and Stereochemistry

- This is supposed to be a sugar and should look like:

Irregular bonding at N1 and N5 of residue GTL in chain C 1

**BUT....**

EMBL-EBI

## Geometry and Stereochemistry

- Always look at the structure in graphical viewers.
- Look at the geometry section in PDB files (REMARK 500).
- Use tools like MSAnalysis, PDBSum to analyze structures.

<http://www.ebi.ac.uk/msd-as/MSDvalidate>

EMBL-EBI

## R-Factor/Correlation

- R-factor is a measure of the agreement between the crystallographic model and the experimental X-ray diffraction data.
- Free R-factor is calculated between the structure and a certain subset of the data excluded from the structure calculation process.
- In a good structure, the difference between R-factor and Free R-factor ( $\Delta R$ ) should be less than 5%.
- Correlation calculates the overall correlation between the structure and the data available.
- Good structure should have overall correlation in excess of 90%.

See <http://eds.bmc.uu.se/eds/> for experimental correlations in crystal structures

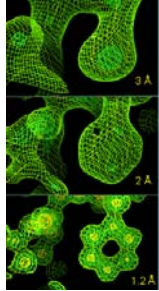
Look at the R-factors on the Atlas Pages in the tutorials !!!

EMBL-EBI

## Resolution

- Resolution is an indicator of the level of detail available in the data used for determining structures in X-ray crystallography.
- Higher resolution (lower number) means that there is more detail available.
  - Low resolution:  $<3.0\text{\AA}$
  - Medium resolution:  $1.8-3.0\text{\AA}$
  - High Resolution:  $1.0 - 1.8\text{\AA}$
  - Atomic Resolution:  $>1.0\text{\AA}$

Not all parts of the structure are at the same resolution...



EMBL-EBI

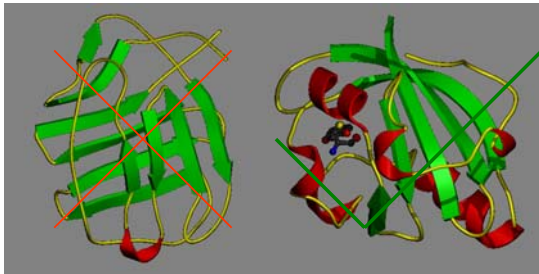
## So what do you look for...

- Higher resolution structures where more than one available
- Good geometry and stereochemistry (Look at the Ramachandran plot)
- Lower R-factor and  $\Delta R$  (FreeR-factor – Rfactor)
- High correlation coefficient between experimental data and structure.
- Complete structures (pay attention to the Sequence and how much of it is represented in the structure), with no sequence conflicts.
- Structures with ligands bound may be more useful for analysis than apo-form structures.

Note: These are general guidelines which may help you choose the best structure for your analysis where more than one structure for the same protein is available.

EMBL-EBI

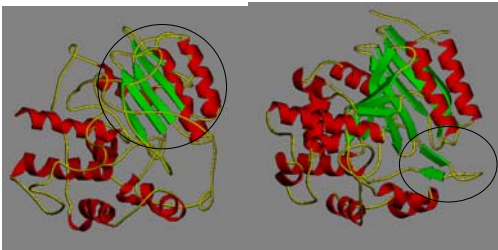
## Wrong Structures !!



PDB entry 1PHY                      PDB entry 2PHY

EMBL-EBI

## Wrong Structures



PDB entry 1PTE                      PDB entry 3PTE

EMBL-EBI

## General Evaluation Criteria

*Be sceptical and cynical!*

When you are searching for information you need to judge its quality and suitability.

Think critically about each piece of information you find and how you found it.

Relevance:

- Does the information you have found adequately support your research?
- Does it answer the question, or support one of your arguments?
- How general or specific is the information about the topic?



EMBL-EBI

## Validation

Some programs for Structure Validation:

- Procheck  
<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>
- Biotech Validation Suite:  
<http://biotech.ebi.ac.uk:8400/>
- WHATCHECK:  
<http://swift.cmbi.ru.nl/gv/whatcheck/>
- JCSG Validation:  
<http://www.jcsg.org/scripts/prod/validation1.cgi>
- MSDanalysis:  
<http://www.ebi.ac.uk/msd-as/MSDvalidate>

## Selected websites offering structure quality assessments

Name	Address	Description
MSDValidate	<a href="http://www.ebi.ac.uk/msd-as/MSDvalidate">http://www.ebi.ac.uk/msd-as/MSDvalidate</a>	Geometric Analysis of structures in PDB or your own PDB file.
RCSB PDB	<a href="http://www.rcsb.org/pdb">http://www.rcsb.org/pdb</a>	Provides detailed geometric information for all PDB entries in the archive.
PDBsum	<a href="http://www.ebi.ac.uk/thornton-srv/databases/pdbsum">http://www.ebi.ac.uk/thornton-srv/databases/pdbsum</a>	A pictorial atlas of PDB entries which includes quality indicators such as the Ramachandran plot and full PROCHECK analysis
PDBREPORT	<a href="http://swift.cmbi.ru.nl/gv/pdbreport/">http://swift.cmbi.ru.nl/gv/pdbreport/</a>	Detailed structure quality reports for each PDB entry calculated using WHAT_CHECK.
MolProbity	<a href="http://kinemage.biochem.duke.edu/molprobity/">http://kinemage.biochem.duke.edu/molprobity/</a>	Graphical contact and geometrical analysis of an uploaded PDB-style file or PDB entry.
Verify3D	<a href="http://nihserver.mbi.ucla.edu/Verify_3D/">http://nihserver.mbi.ucla.edu/Verify_3D/</a>	Provides a graphical analysis of quality of an uploaded PDB file.
AQUA	<a href="http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server">http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server</a>	An interactive program to assess the quality of structures determined by NMR spectroscopy using deposited experimental data and PDB files
EDS	<a href="http://eds.bmc.uu.se/eds">http://eds.bmc.uu.se/eds</a>	The Uppsala Electron Density Server offers statistics regarding the accuracy and quality of X-ray crystal structures based on the analysis of electron density maps.
Biotech Validation Suite	<a href="http://biotech.ebi.ac.uk:8400">http://biotech.ebi.ac.uk:8400</a>	A validation suite for protein structures, including full PROCHECK and WHAT_CHECK analyses.
JCSG Validation Tools	<a href="http://www.jcsg.org/scripts/prod/validation1.cgi">http://www.jcsg.org/scripts/prod/validation1.cgi</a>	Provides tools for validating structures and X-ray experimental data.
VADAR	<a href="http://redpoll.pharmacy.ualberta.ca/vadar">http://redpoll.pharmacy.ualberta.ca/vadar</a>	<b>V</b> olume, <b>A</b> rea, <b>D</b> ihedral <b>A</b> ngle <b>R</b> eporter for analyzing protein structure quality.
STAN	<a href="http://xray.bmc.uu.se/cgi-bin/gerard/rama_server.pl">http://xray.bmc.uu.se/cgi-bin/gerard/rama_server.pl</a>	Assesses protein structure geometry and associated waters.