Protein Structure

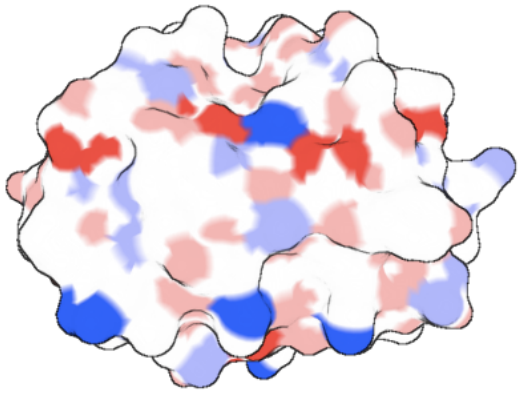
Sue Jones

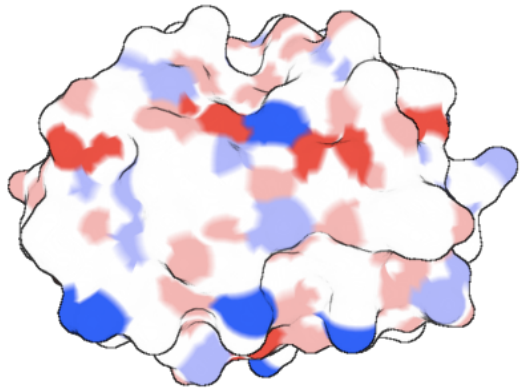
IBioIC Bioinformatics Section 4

Strathclyde University

16/17th March 2017

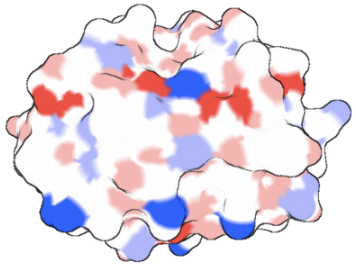
sue.jones@hutton.ac.uk





Outline

- What is protein structure?
- How do we know about protein structure?
- Where and how is protein structure data stored?
- Why is it important?
- Sequence/structure/function relationships
- What can we do with protein structure information?
- Visualising and analysing protein structure



Section 1

What is protein structure?

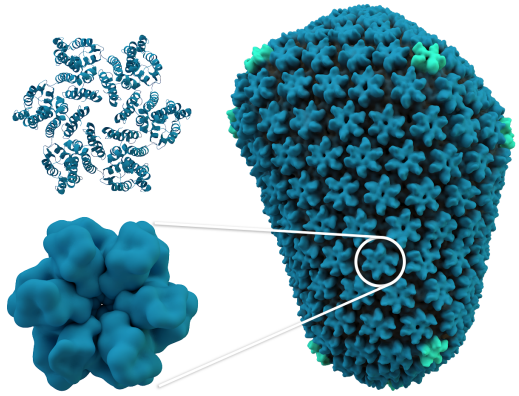
How do we know about protein structure?

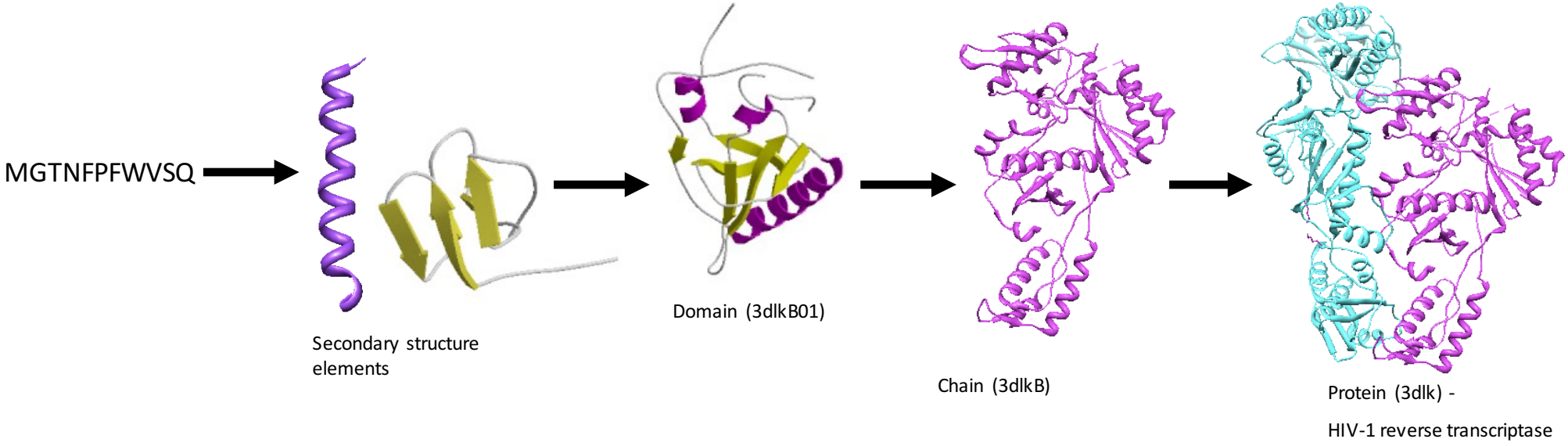
Where is structure information stored

Practical 01: Exploring the protein structure databank (RCSB)

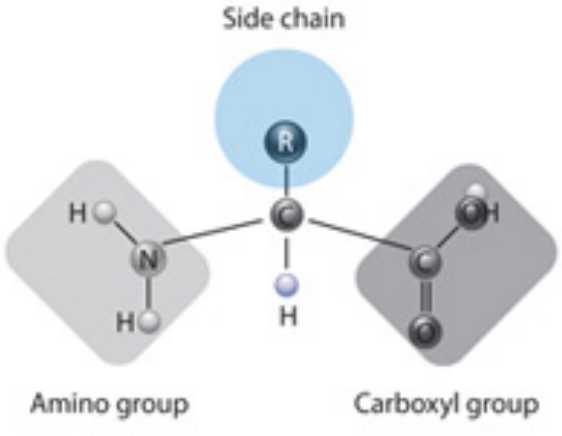
What is protein structure?

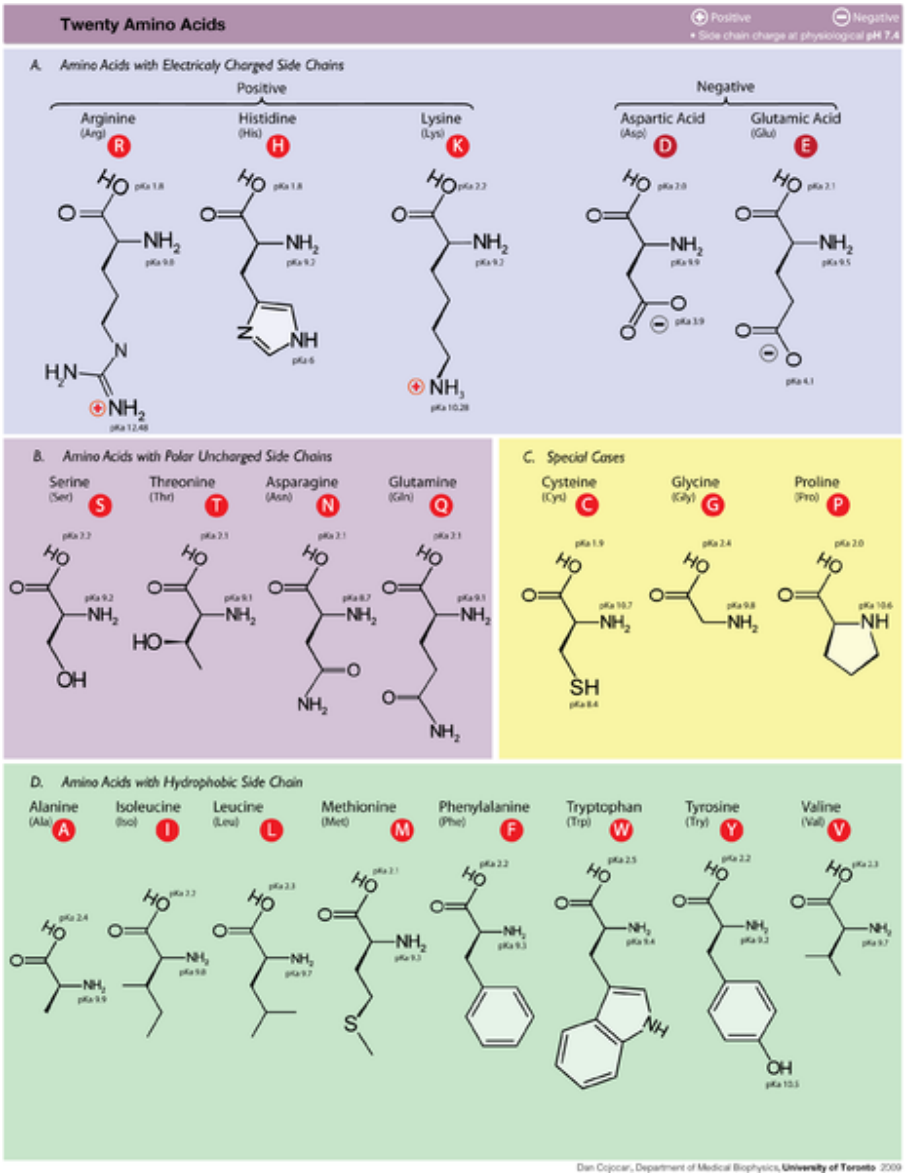
- Hierarchical organisation
- Primary (sequence), secondary (strands, helices), tertiary (fold)
- Quaternary: dimers, tetramers, n-mers





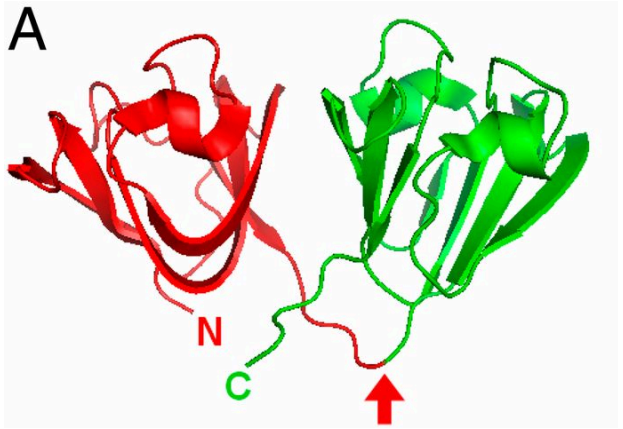
Amino acid structures: sidechains



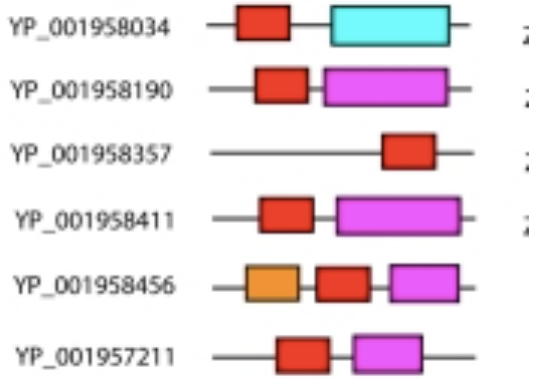


Protein Domains

- Domains are “units of evolution”
  - evolve and exist independently
  - often can fold independently
- ~ 40% PDB proteins multidomain ~60-80% of genes in known genomes code for multidomain proteins
- Domains ‘re-used’ in many different proteins
- Combined in different ways

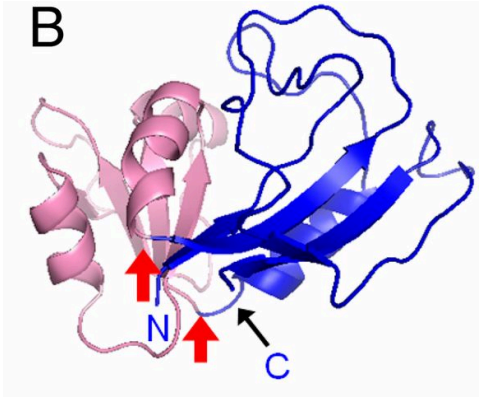


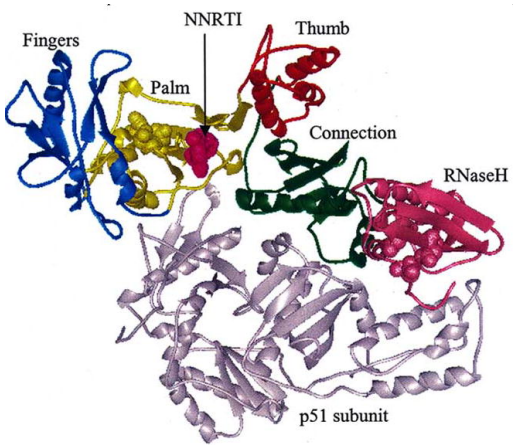
Human γD-crystallin (1HK0): 2 independently foldable domains connected by a single linker



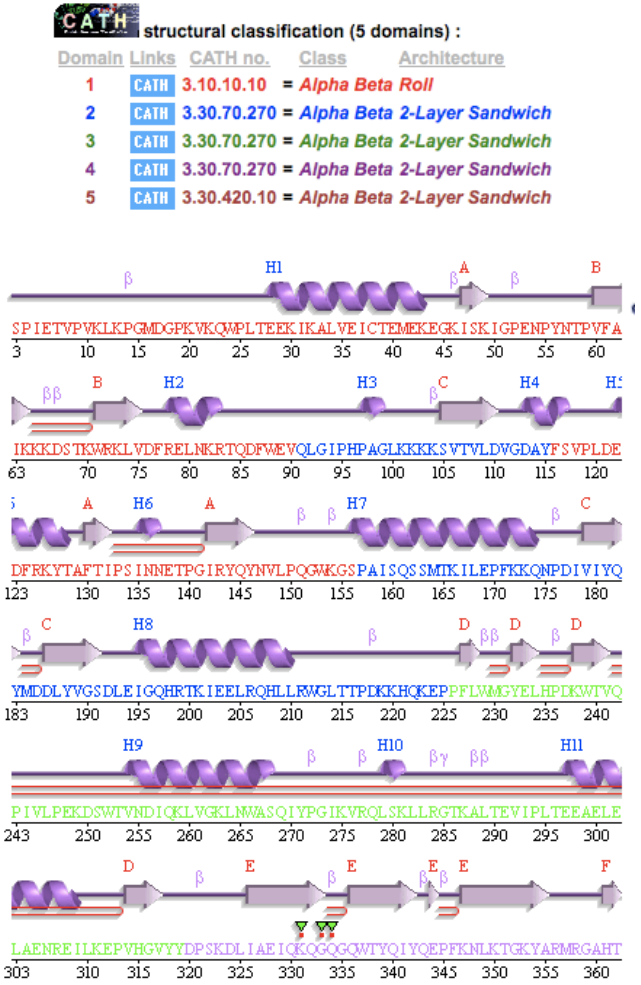
Domains can be discontinuous

- Protein folding brings together parts of the sequence that are not contiguous
- E.g. HIV1 reverse transcriptase (3dlk)
- E.g. Dihydrofolate reductase (1rx1) sequence domains vs structural domains





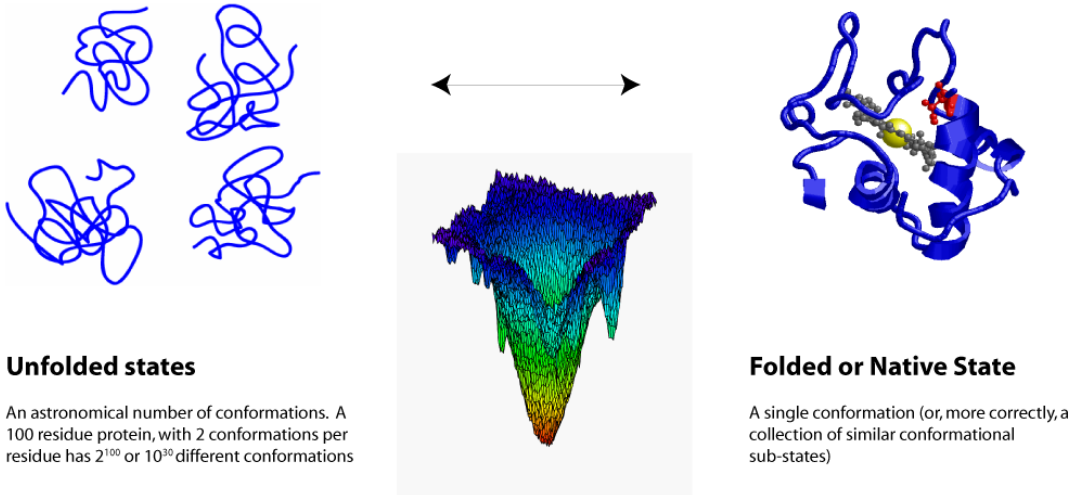
Takashi Inanami et al. PNAS 2014;111:15969-15974



Images from PDBsum

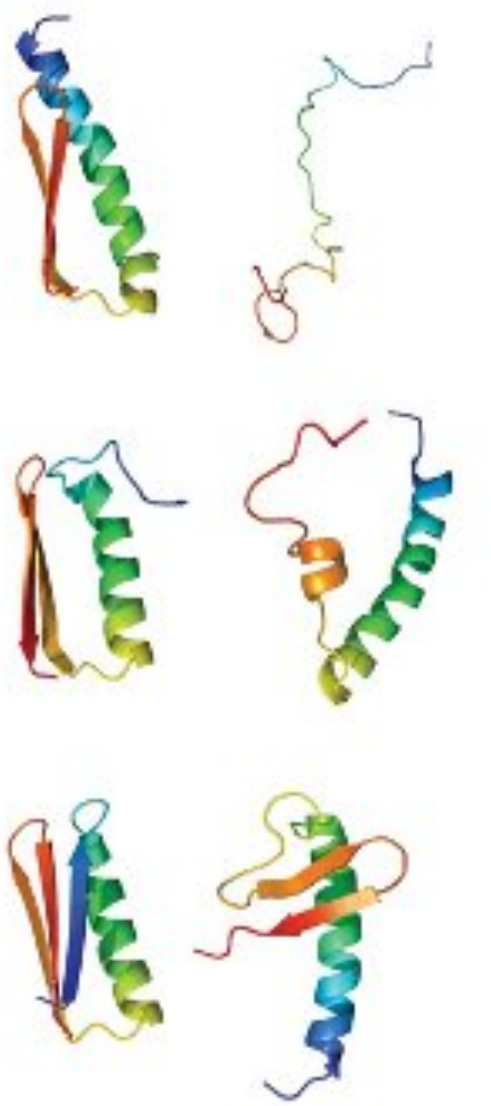
Protein folding: native state

- Protein usually fold into a ‘native conformation’: minimum free energy state
- energy determined by interactions: mainchain, sidechain, solvent
- Proteins evolved so that a single way of folding is thermodynamically much more stable than alternative ways



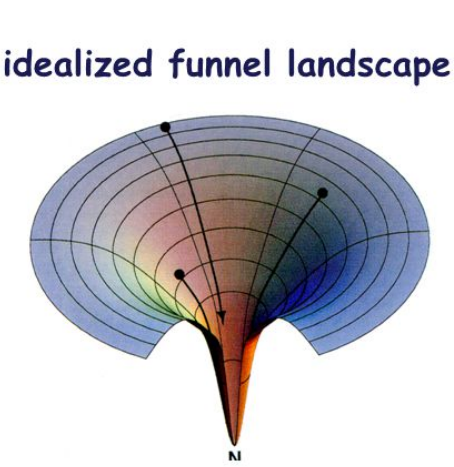
How to achieve a stable native state?

- Native conformation needs to satisfy many requirements:
- Hydrophobic residues: packed inside the core away from water (hydrophobic effect)
- Interior : densely packed to maximise interactions
- Polar groups on the inside : hydrogen bond to other groups (formation of helices and sheets)
- All residues must have stereochemically allowed conformations

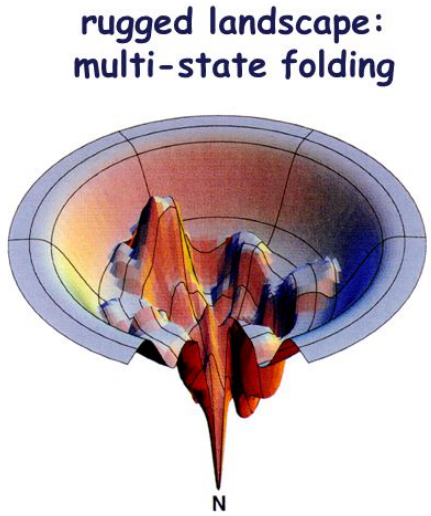


Protein folding: Folding funnels

- Number of conformations a protein needs to sample to fold to the native state is BIG: need longer than the age of the universe to fold
- Yet most proteins can fold in millisecond time scales (Levinthal paradox (60's))
- Folding is cooperative, i.e. steps towards native state are improvements on preceding steps
- Residues do not randomly search the conformational space



As the chain forms increasing numbers of intrachain contacts, and lowers its internal free energy, its conformational freedom is also reduced



A rugged landscape with kinetic traps, energy barriers, and some narrow throughway paths to native. Folding can be multistate

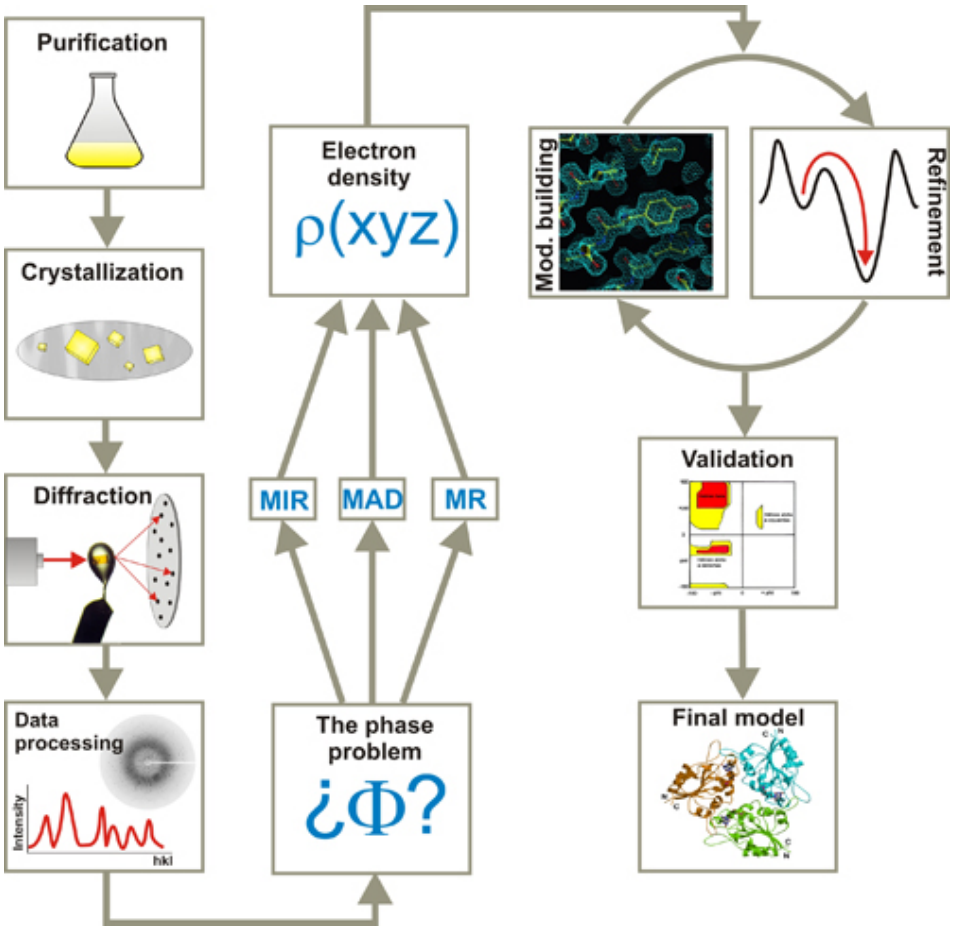
http://dillgroup.org/energy-landscapes.html

How do we know about protein structures?

- X-ray crystallography
- Nuclear Magnetic Resonance
- Electron Microscopy (CryoEM)
- Combination techniques

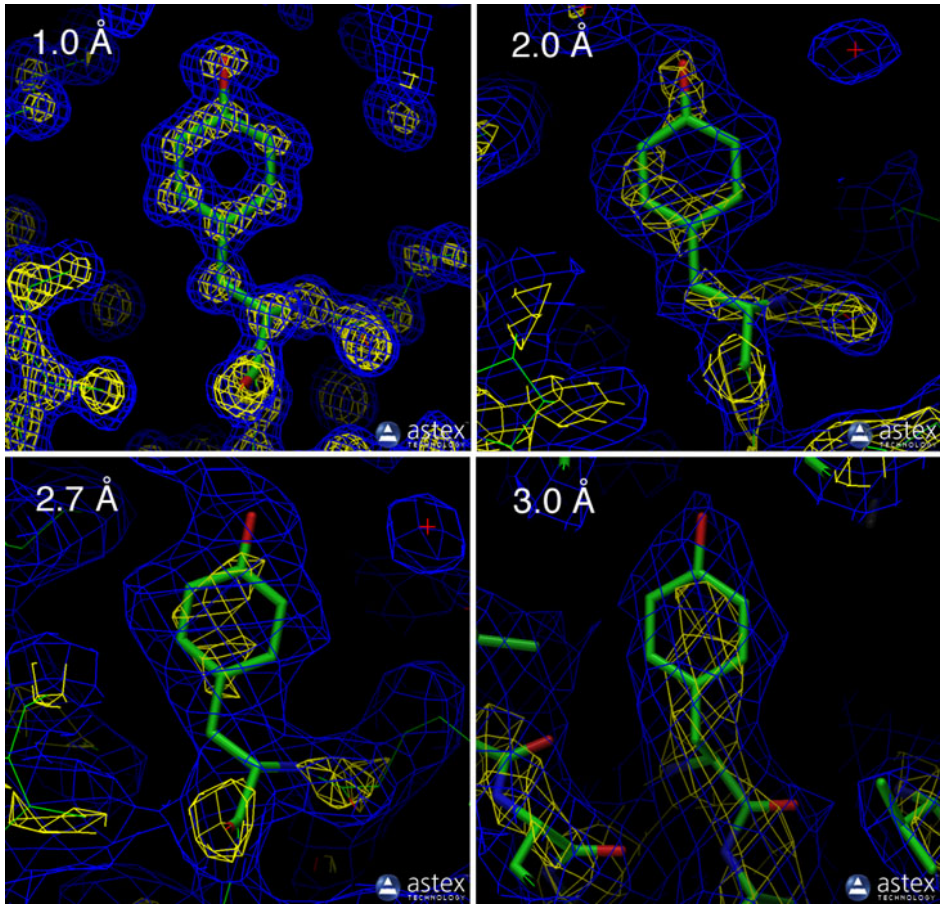
X-ray crystallography

- Crystals are formed by regular arrays of molecules that diffract x-rays in regular and predictable patterns
- Proteins fixed in a crystal
- Snapshot of single protein conformation



Quality of a structure - Resolution

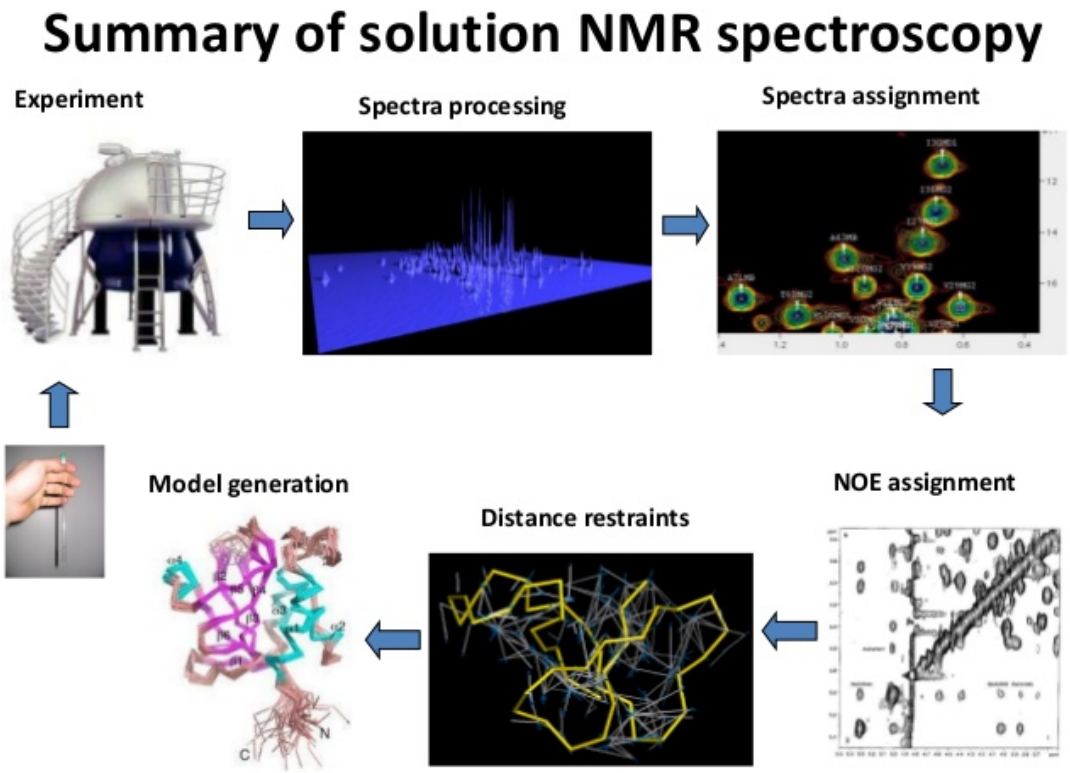
- Resolution: measure of the amount of detail present in the diffraction pattern and corresponding computed electron density map
- Lower the number the better the resolution



www.pdb.org

Solution NMR

- Nuclei of isotopes such as 1H, 13C and 15N carry magnetic dipoles
- Take up different orientations with different energies in the magnetic field of an NMR spectrometer
- nuclei in different environments resonate (vibrate) at different frequencies,
- plotting intensity against frequency gives a 1- dimensional NMR spectrum
- Multidimensional spectrum
- Proteins in solution

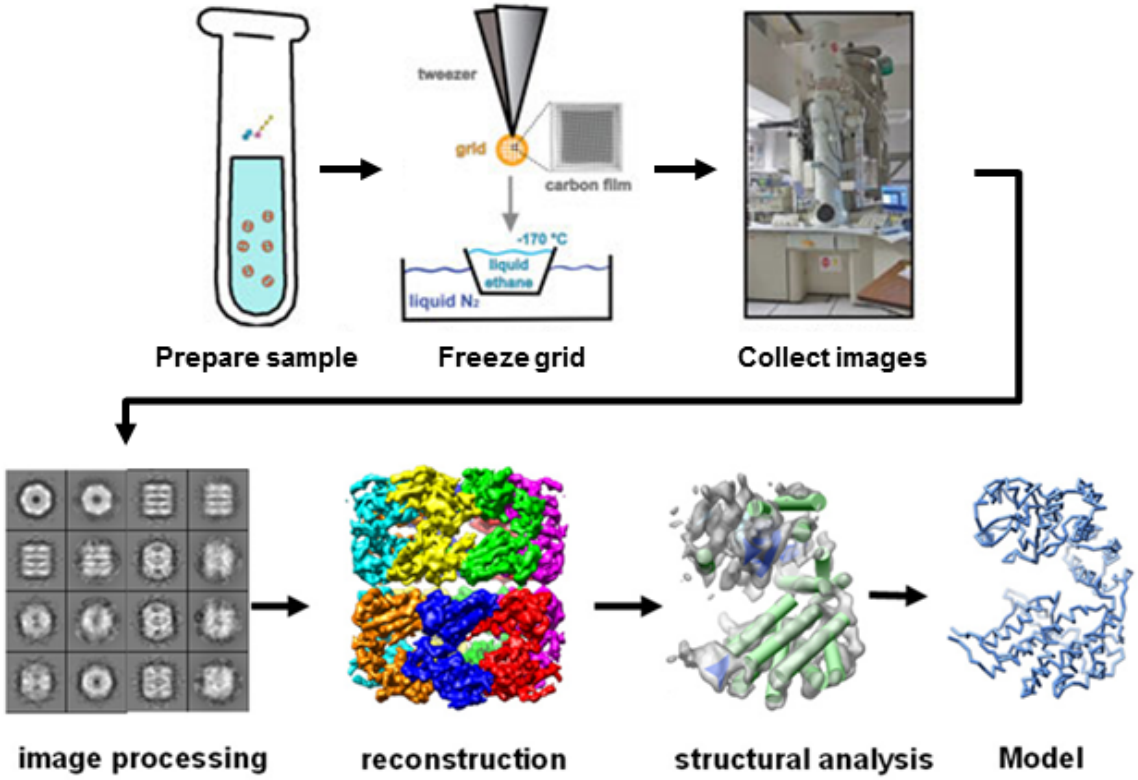


NMR: Strengths and limitations

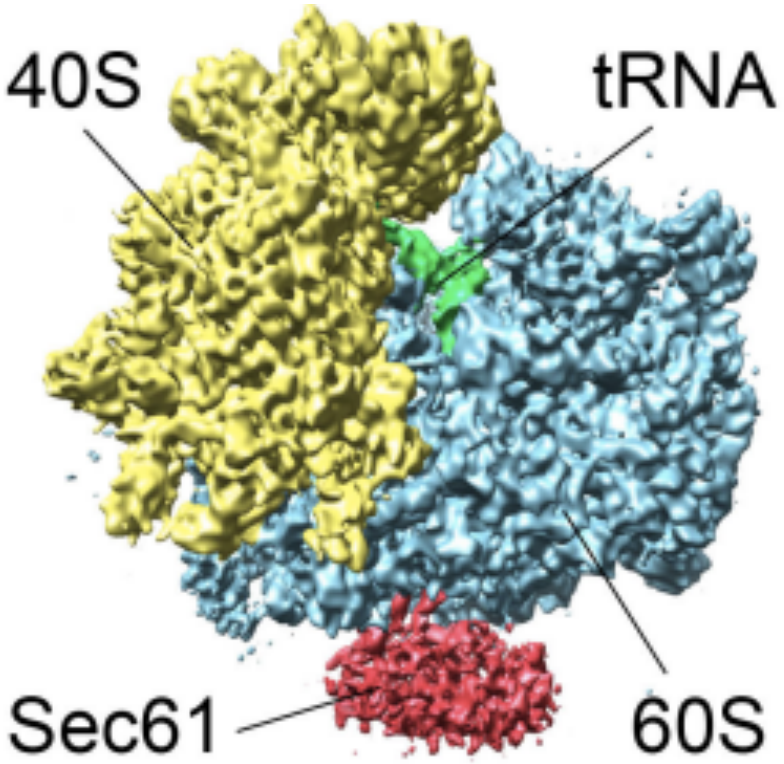
- Difficult to estimate the reliability of NMR structures from PDB
- Structures often excluded from high-quality datasets used to train and test software
- Proteins are studied in a more natural environment (in solution)
- Applicable to membrane proteins, disordered or unfolded proteins Dynamics of the protein can be studied
- Originally limited limited to ‘small’ proteins (<50 kDa)

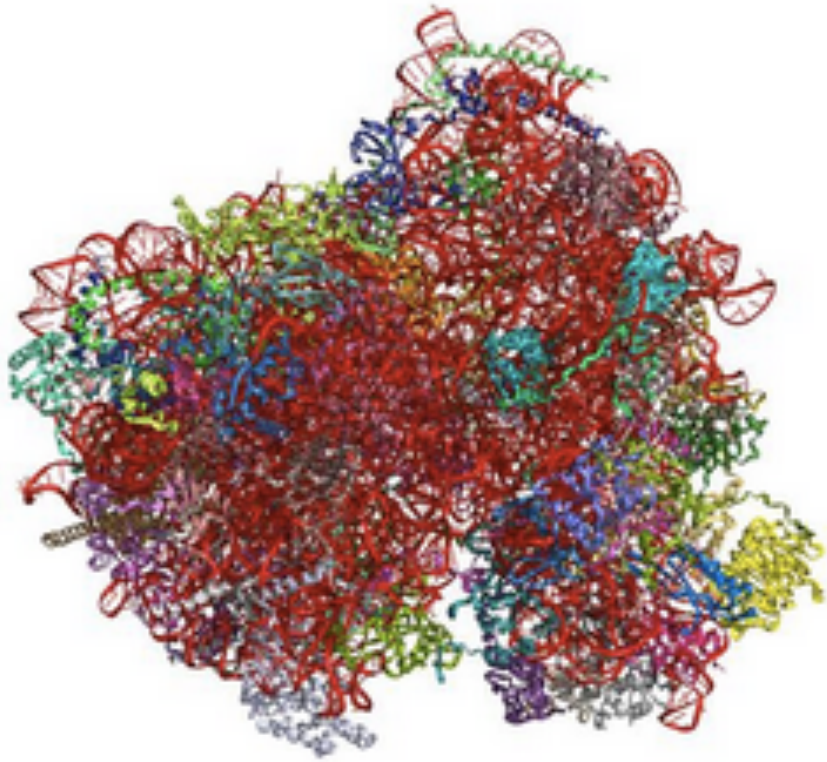
Cryo-Electron Microscopy

- Imaging radiation-sensitive molecules in transmission electron microscope using very low temperatures
- Initially poor resolution 10 Å
- Now resolution < 3.0 Å



Cryo-EM Success: Ribosome



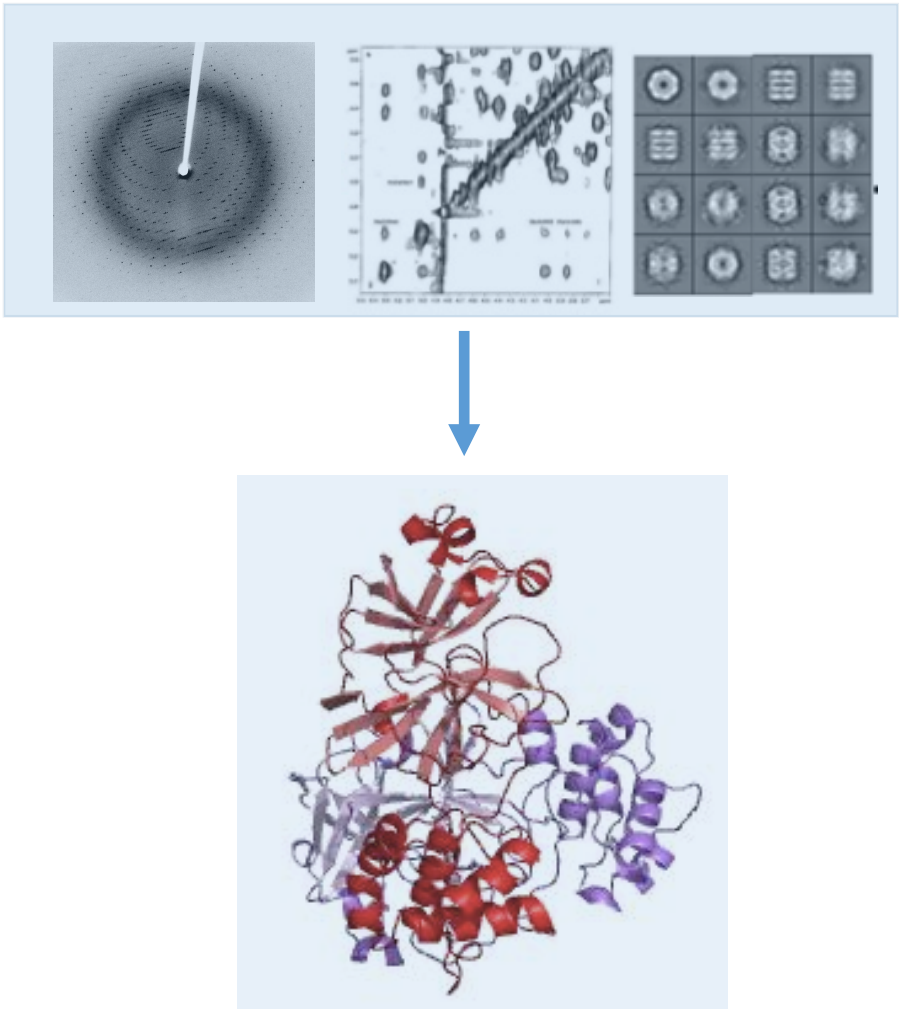


Human 80s ribosome (3.5 Å)

69 protein chains & 5 nucleic acids

Protein structures are models

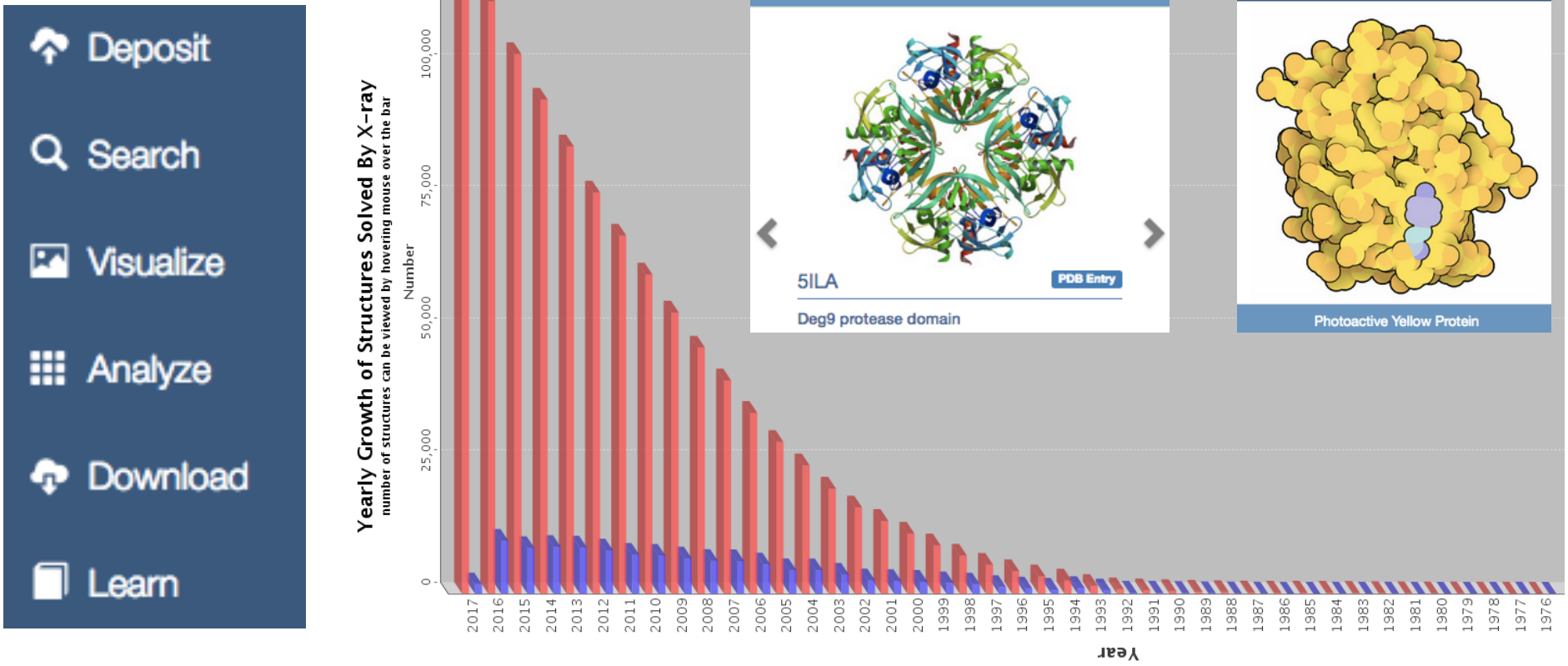
- Easy to consider a protein structure as the "true" three-dimensional structure of a protein
- In reality, every structure is a model that fits the experimental data to a reasonable degree
- Experimental data can also contain errors!
- Not all structures are equal!



Where are protein structure stored?

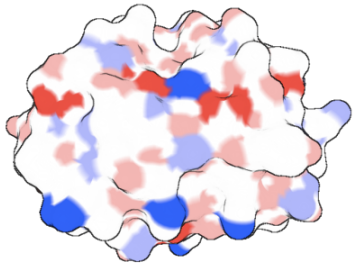
RCSB PDB PROTEIN DATA BANK
An Information Portal to 126447 Biological Macromolecular Structures

www.rcsb.org



Practical 04-01

Searching for data in the Protein structure database (RCSB)



Section 2

Protein structure file format

Protein structure visualisation

RCSB Protein Structure File Format

- New PDBx/mmCIF: ‘macro-molecular Crystallographic Information File'
- RCSB entries distributed in PDBx/mmCIF file format standard in 2014
- Old PDB format not been modified or extended since 2012: now frozen
- As PDBx/mmCIF format evolves PDB format files will become outdated
- 190 page file format description!
  ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_A4.pdf
- Based on context free simple grammar: Data are presented in either key-value or tabular form. It is much easier to parse than the record-oriented PDB format
- Parsing tools available in most languages and toolkits (bioPython, bioJava)

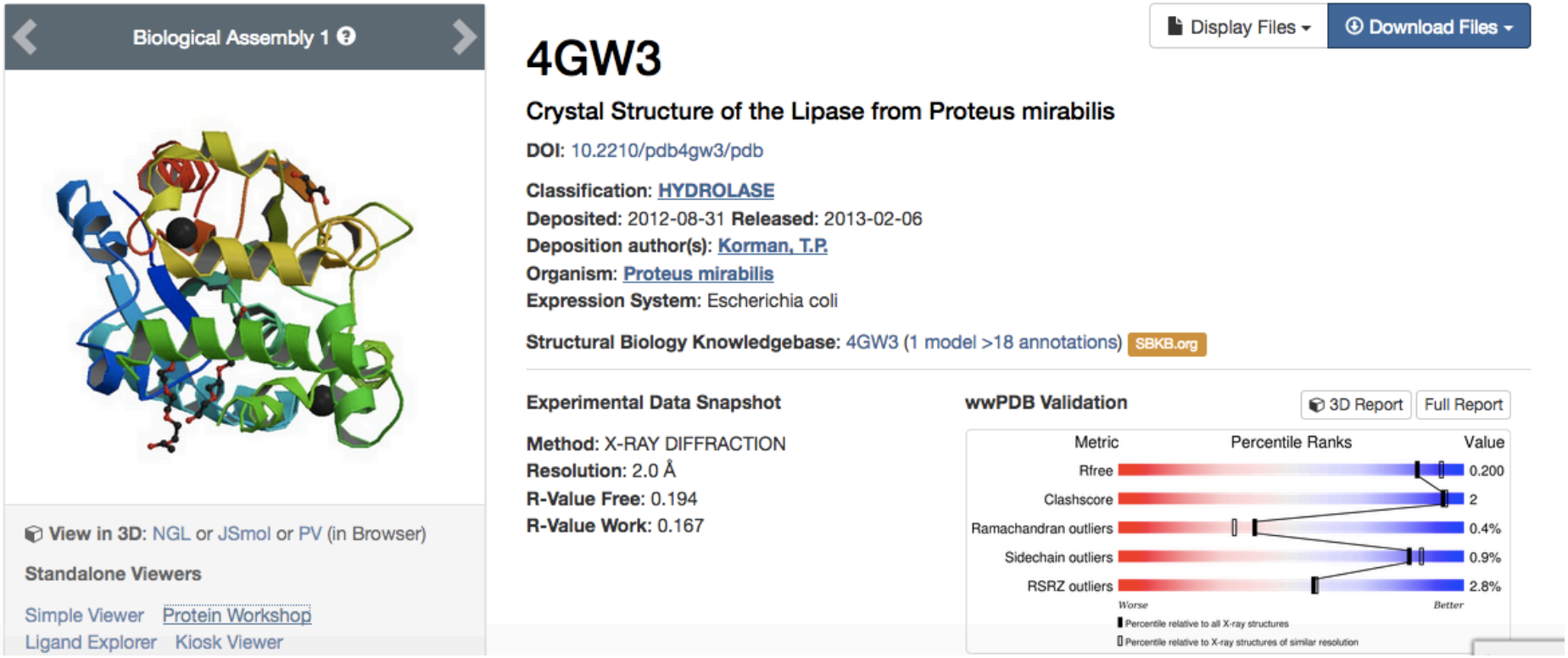
PDBx/mmCIF format

```
PDB   4GW3
RCSB RCSB074705
#
_database_PDB_rev.num                1
_database_PDB_rev.date               2013-02-06
_database_PDB_rev.date_original      2012-08-31
_database_PDB_rev.mod_type           0
_database_PDB_rev.replaces           4GW3
_database_PDB_rev.status             ?
#

_entity_src_gen.pdbx_gene_src_scientific_name      'Proteus mirabilis'
_entity_src_gen.pdbx_gene_src_ncbi_taxonomy_id     584

ATOM   1    N   N    . SER A 1 22  ? -1.562 23.864 -13.828 1.00 34.26 ? ? ? ? ? ? 2   SER A N   1
ATOM   2    C   CA   . SER A 1 22  ? -0.729 25.018 -14.279 1.00 34.62 ? ? ? ? ? ? 2   SER A CA  1
ATOM   3    C   C    . SER A 1 22  ? 0.681  24.588 -14.693 1.00 33.18 ? ? ? ? ? ? 2   SER A C   1
ATOM   4    O   O    . SER A 1 22  ? 1.233  25.136 -15.646 1.00 33.54 ? ? ? ? ? ? 2   SER A O   1
ATOM   5    C   CB   . SER A 1 22  ? -0.664 26.093 -13.194 1.00 35.21 ? ? ? ? ? ? 2   SER A CB  1
ATOM   6    O   OG   . SER A 1 22  ? -0.040 27.270 -13.676 1.00 36.35 ? ? ? ? ? ? 2   SER A OG  1
```

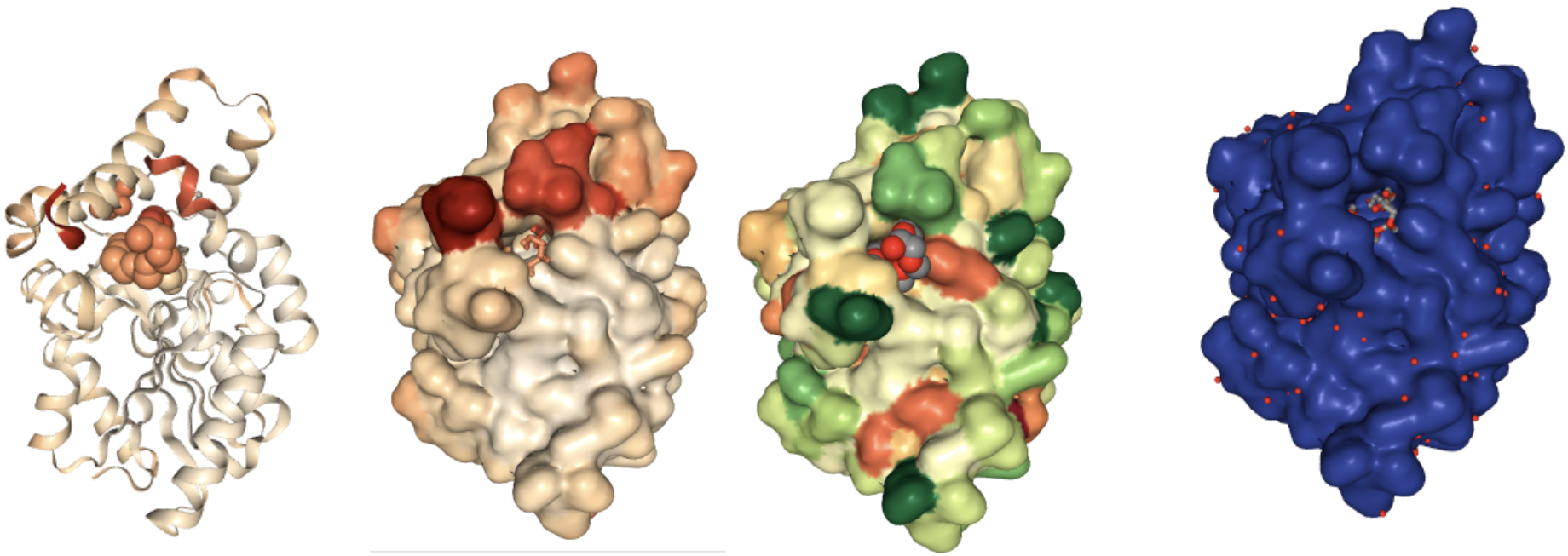
? Missing values: col 9: insertion code

Col 15: sigma X : standard uncertainty of coordinate (0.01% of PDB entries)

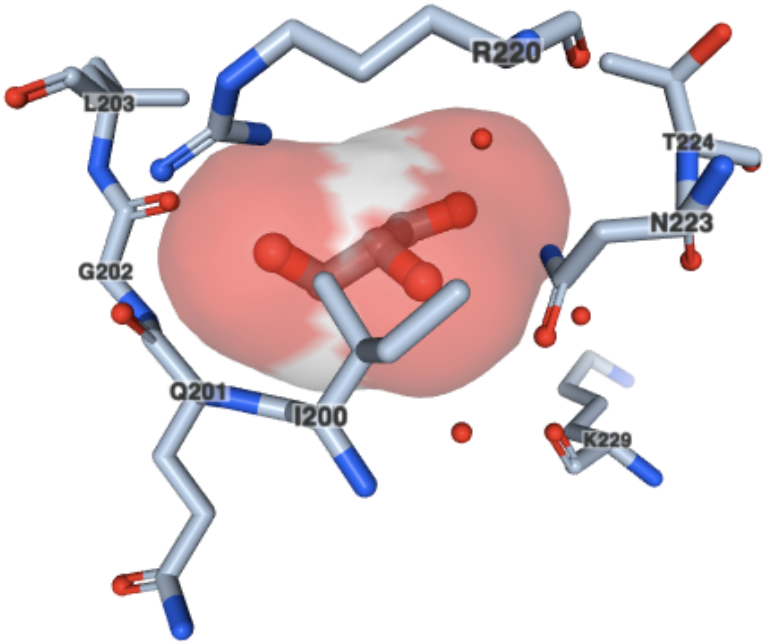
PM Lipase structures in the RCSB

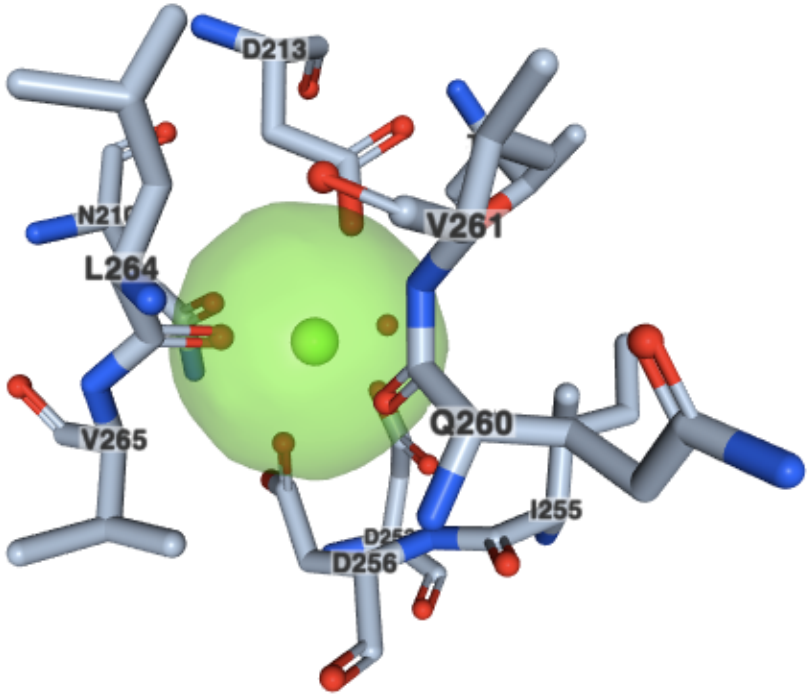


Protein structure visualisation



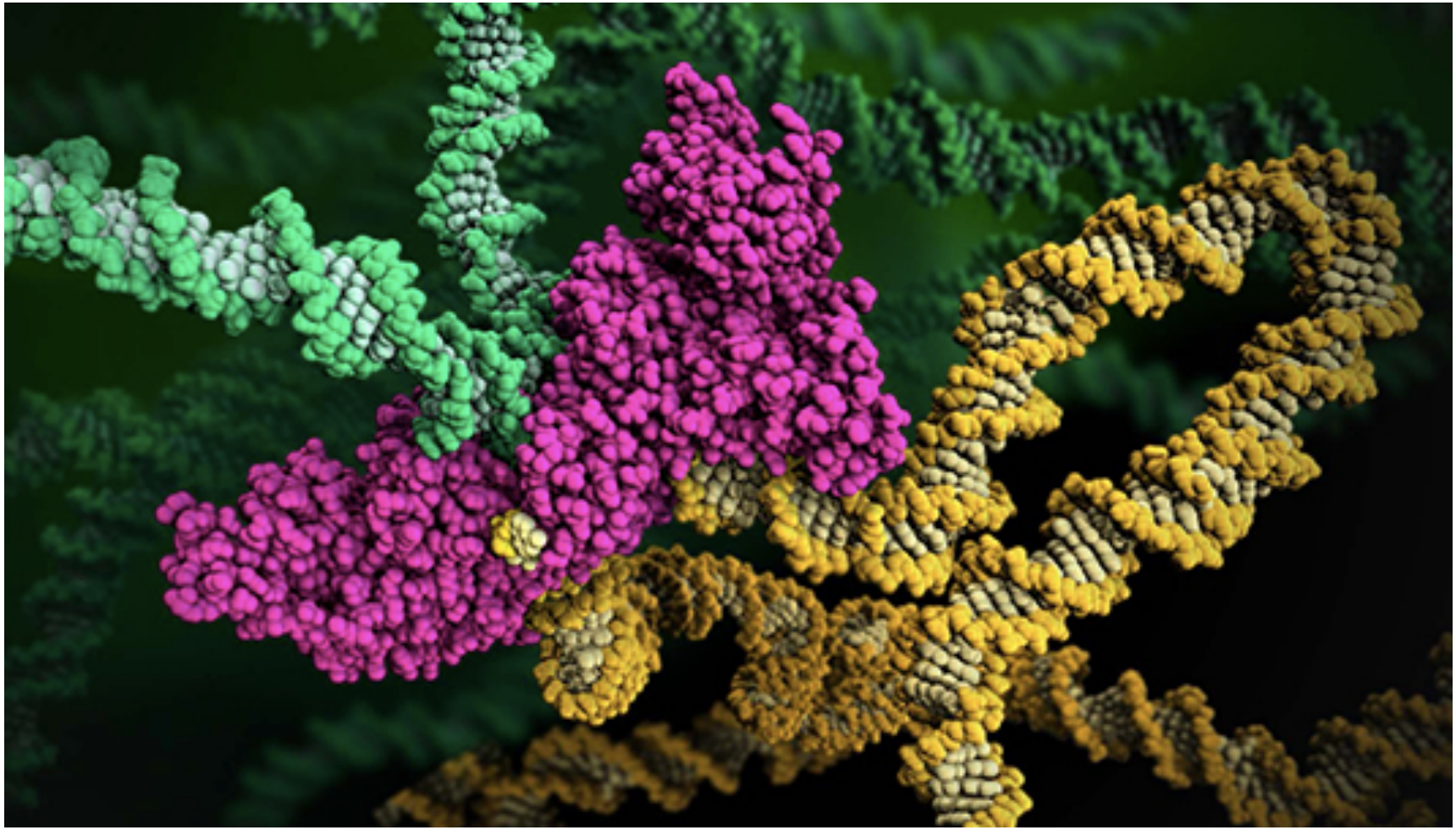
Protein structure visualisation





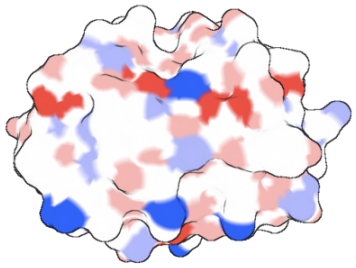
Red- glycerol, green calcium

Structures become art



Practical 04-02

Protein structure visualisation using NGL – a web-based viewer



Section 3

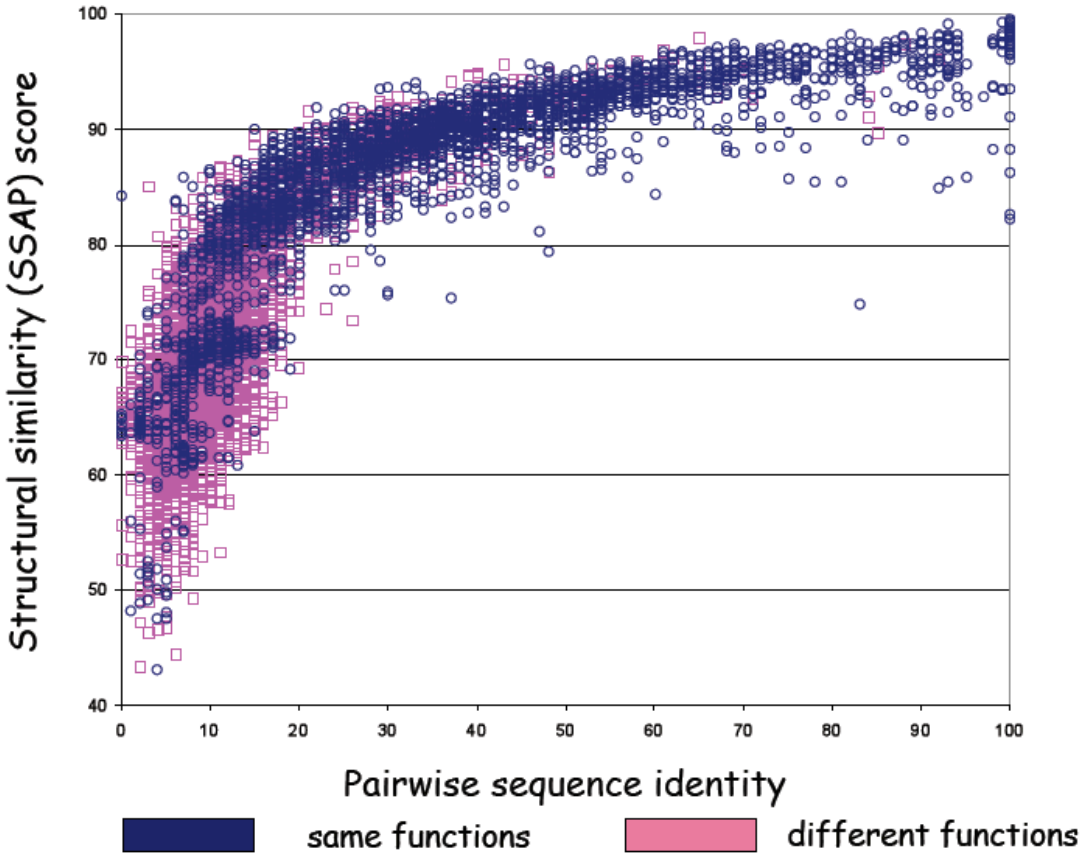
Sequence/structure/function relationships

Protein families

What can protein structure be used for?

Sequence/Structure/Function relationships

Proteins with low sequence similarity can have very similar structures

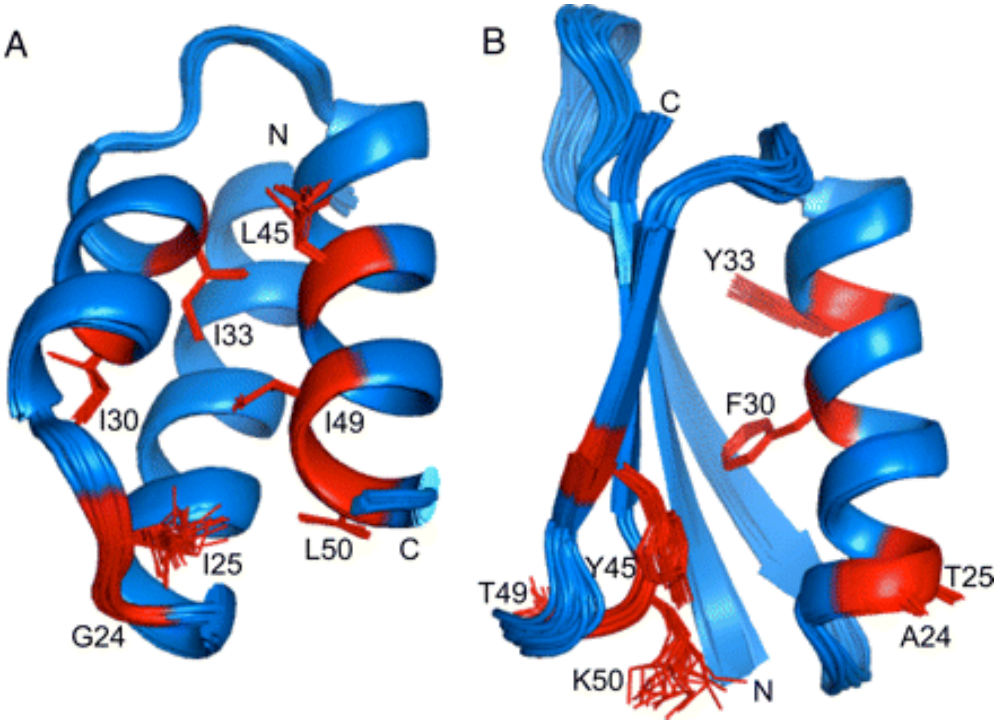


Christine Orengo, UCL

Sequence/Structure/Function relationships

Proteins with high sequence similarity can have very different structures

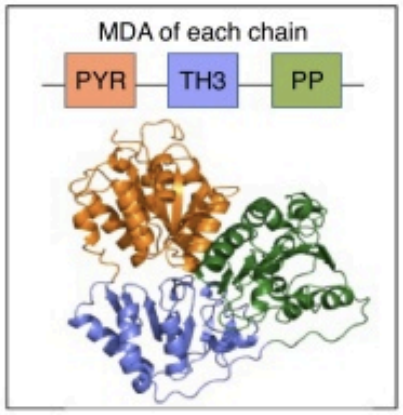
No absolute rules

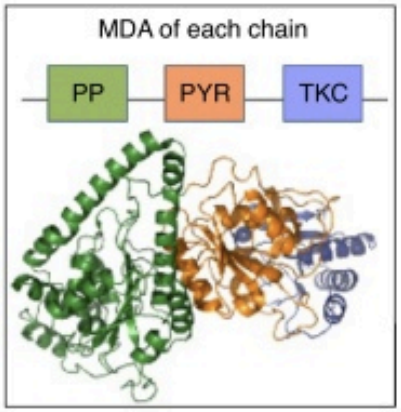


Two designed proteins with 7 amino acid differences (in red).
Change: all-alpha fold to an alpha/beta fold
88%sequence identity

He et al. (2008) PNAS, 105,14412

Sequence/structure/function relationships: domain context





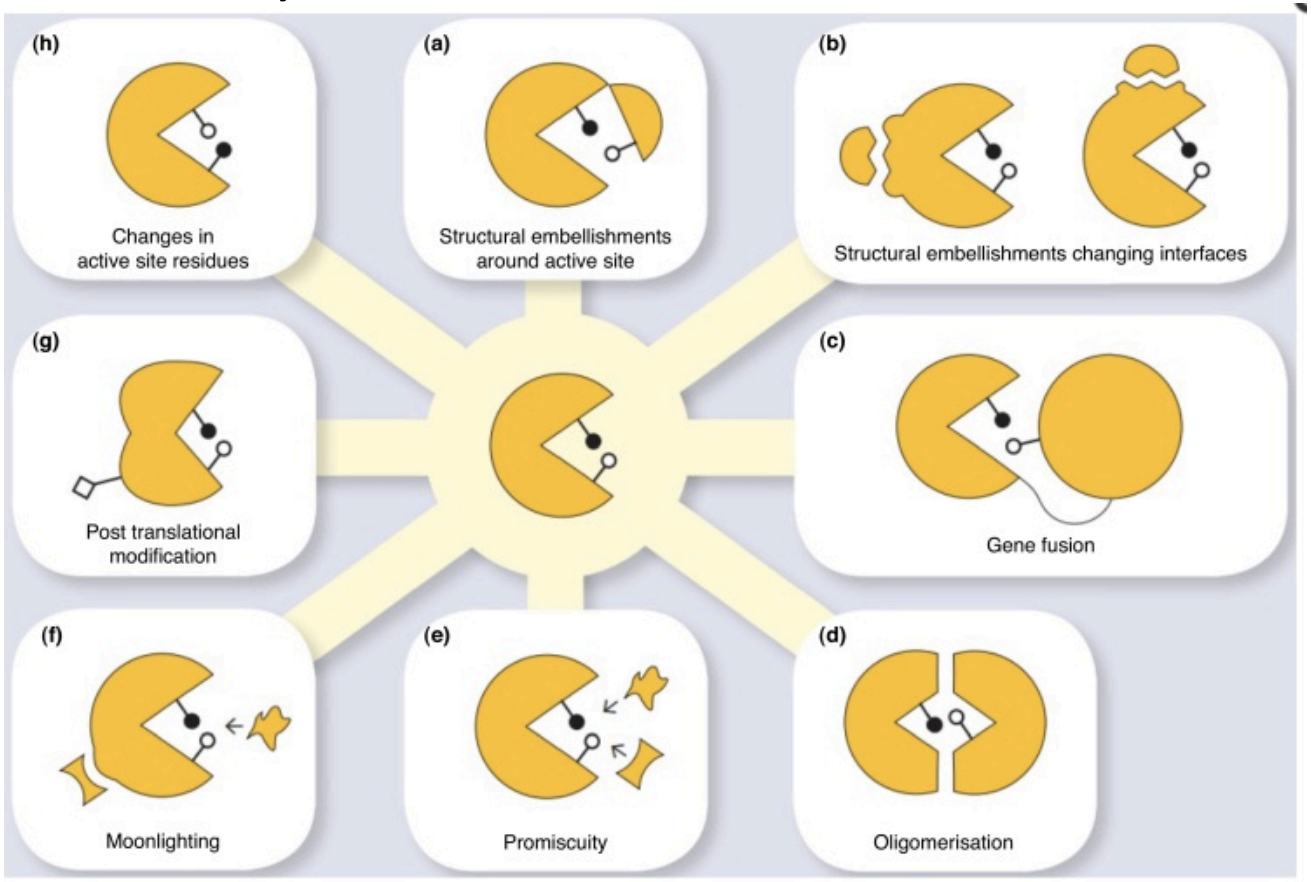
Pyruvate decarboxylase (PDC, EC 4.1.1.1)

Transketolase (TK, EC 2.2.1.1)

Das et al., (2015) Curr Opin Genet Dev: 35:40

Sequence/structure/function: functional diversity

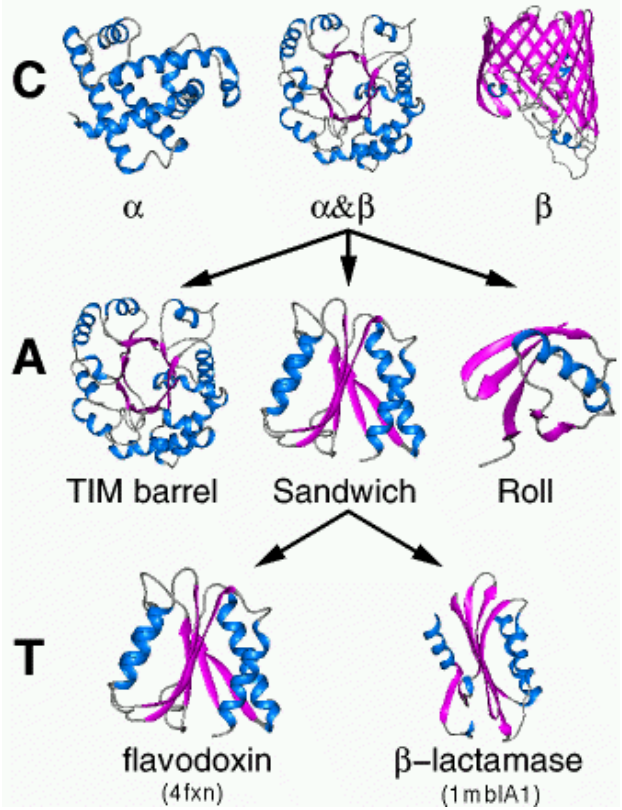
Functional diversity can arise due to one or more mechanisms



Das et al., (2015) Curr Opin Genet Dev: 35:40

Protein structure classification

- Families arise through speciation (orthologous relatives) and duplication (paralogous relatives)
- Explore how relatives diverge and mechanisms for functional change
- Hierarchical classification : CATH and SCOP
- As more structures solved: where does one fold end and another begin?: Fold continuum
- *www.cathdb.info/*

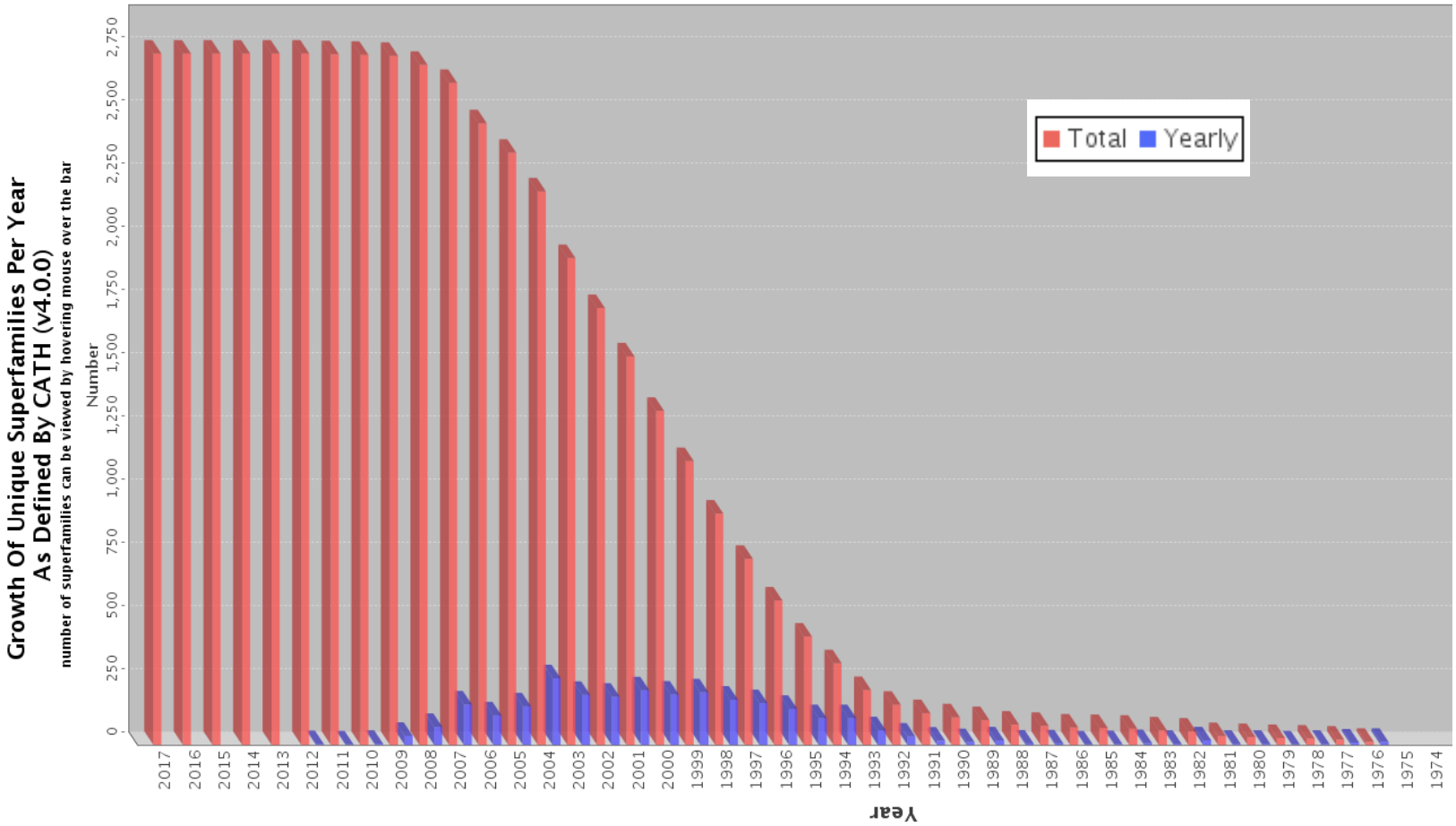


CATH: comprehensive structural and functional annotations for genome sequences.

Sillitoe I, Lewis, TE, Cuff AL, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees J, Lehtinen S, Studer R, Thornton JM, Orengo CA

Nucleic Acids Res. 2015 Jan doi: 10.1093/nar/gku947

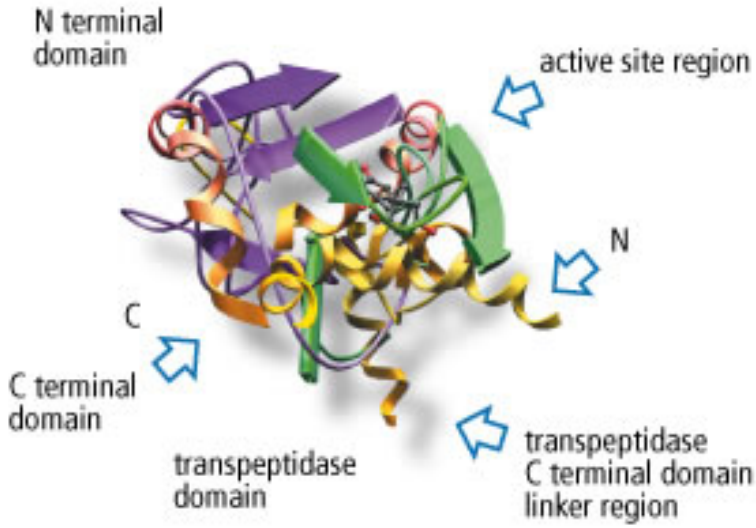
Limited number of folds?



http://www.rcsb.org/pdb/statistics

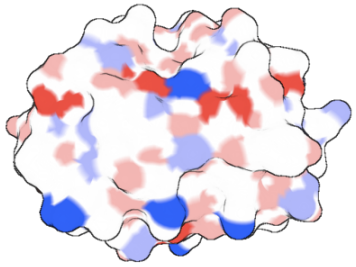
What can a protein structure be used for?

- Understand the function of a protein
- Understand/explain observations from experiment (e.g. the relationship between stability and mutation of a specific residue)
- Understand the binding mode of a ligand
- Protein design: engineered binding sites, engineer greater stability, engineer 'new' functions



Practical 04-03
Practical 04-04

Protein structure visualisation using Jmol
Functional regions in the WT PM lipase (4GW3)
Mutation sites and their effects in Dieselzyme4 (4HS9)
Superimposing WT and mutant proteins to see differences in structure



Section 4

Protein sequence-structure gap

Predicting protein structure from sequence

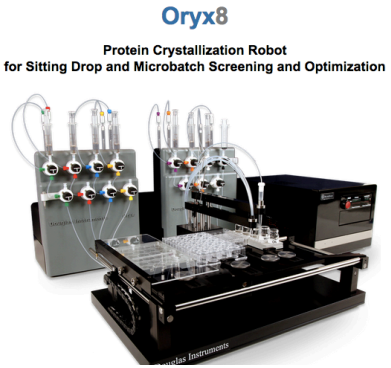
Sequence Structure gap

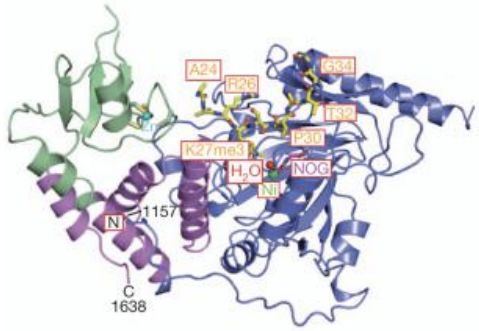
- 77 million protein sequences / 123,000 protein structures
- Sequencing is easy: experimentally determine structure: difficult
- As sequencing gets easier and cheaper the gap will grow
- Structural genomics consortia: targeting 'important' proteins
- Important = human = drug targets
- Robotics for crystallization processes: using 96 well plates to test out conditions

SGC

UNIVERSITY OF TORONTO · UNIVERSITY OF OXFORD · UNICAMP · KAROLINSKA INSTITUTET · GOETHE UNIVERSITÄT FRANKFURT AM MAIN

http://www.thesgc.org/





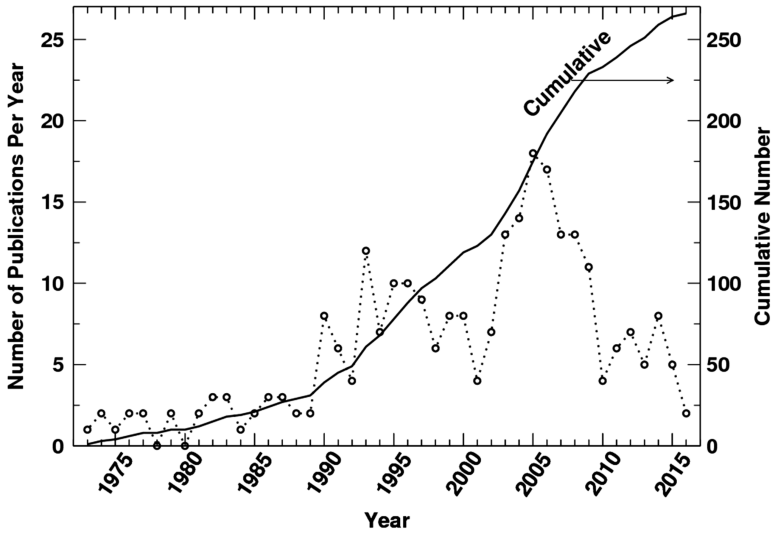
- Helped GSK to identify the potential of histone demethylase JMJD3 as a drug discovery target in inflammation

Secondary structure prediction

Sixty-five years of the long march in protein secondary structure prediction: the final stretch?

Yuedong Yang; Jianzhao Gao; Jihua Wang; Rhys Heffernan; Jack Hanson; Kuldip Paliwal; Yaoqi Zhou

Brief Bioinform bbw129.
DOI: https://doi.org/10.1093/bib/bbw129



Neural networks are a common feature of SS prediction methods

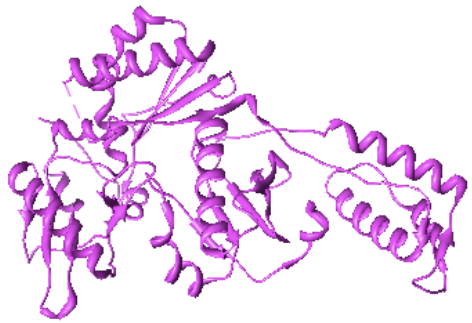
sequence identity from each other. The reported three-state accuracy of secondary structure prediction has gradually risen from 69.7% by PHD in 1993 [54], 76.5% by PSIPRED [55] in 1999, 80% by Structural Property prediction with Integrated Neural nEtwork (SPINE) [56] in 2007, 82% by Structural Property prediction with Integrated DEep neuRal network 2 (SPIDER2) [57] in 2015, to 84% for several test data sets by Deep Convolution Neural Field network (DeepCNF) [58] in 2016. Although accuracies reported by

Predicting tertiary structure from sequence : options

MLRLVVGALLLVLAFAGGYAVAACKTVTLTVD GTAMRVTT MKS RVIDI
VEENGFSVDDRDDLYPAAGVQVHDADTIVLRRS RP LQISLD GHDAK QV
WTTASTVDEALAQLAMTDTAPAAASRASRVPLSGMALPVVSAKTVQL
NDGGLVRTVHLPAPNVAGLLSAAGVP LLQSD HVVPAATAPIVEGM QI Q

No significant sequence similarity to protein with known structure: no homologues

ab initio modelling

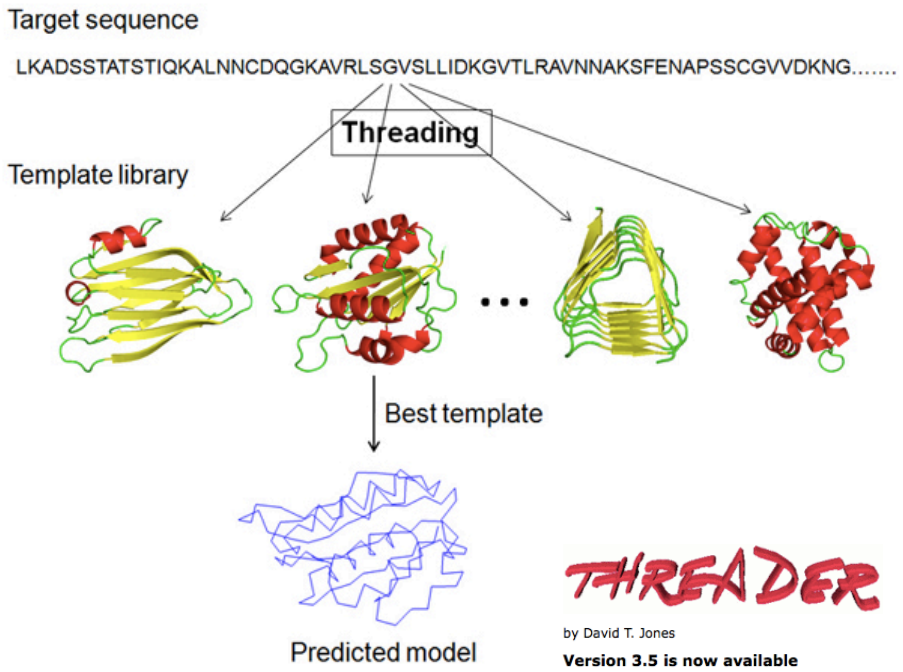


Significant sequence similarity to a protein with known structure : homologues

homology modelling ‘comparative modelling)

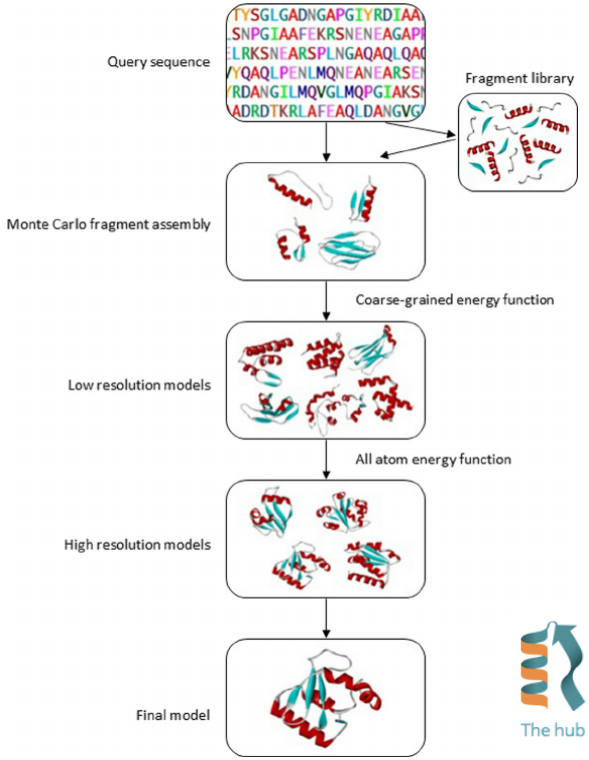
Ab initio modelling

- Threading



- Jones et al., 1992. Nature. 358, 86-89.

- Fragment assembly

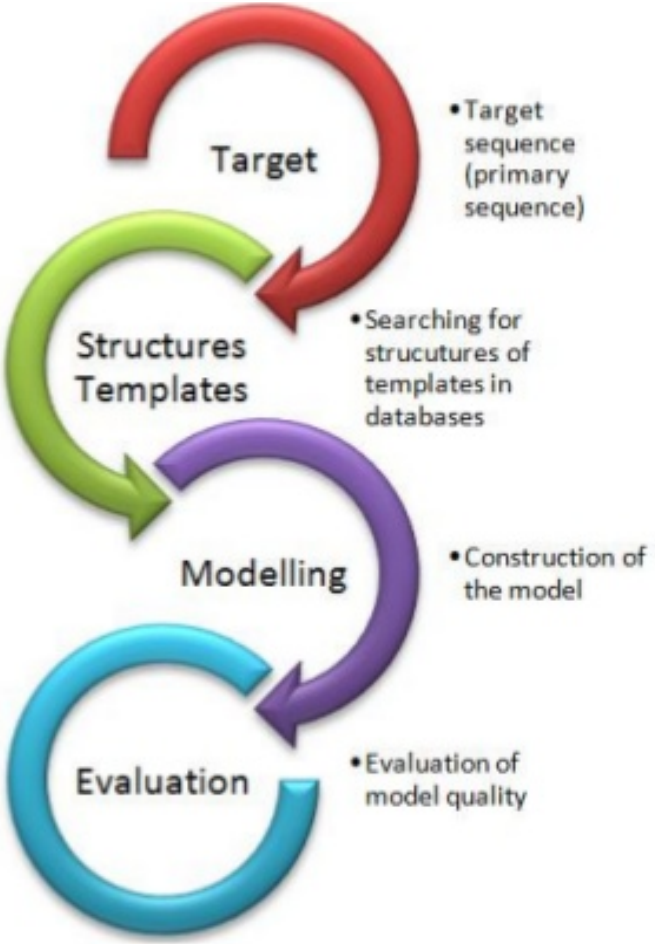


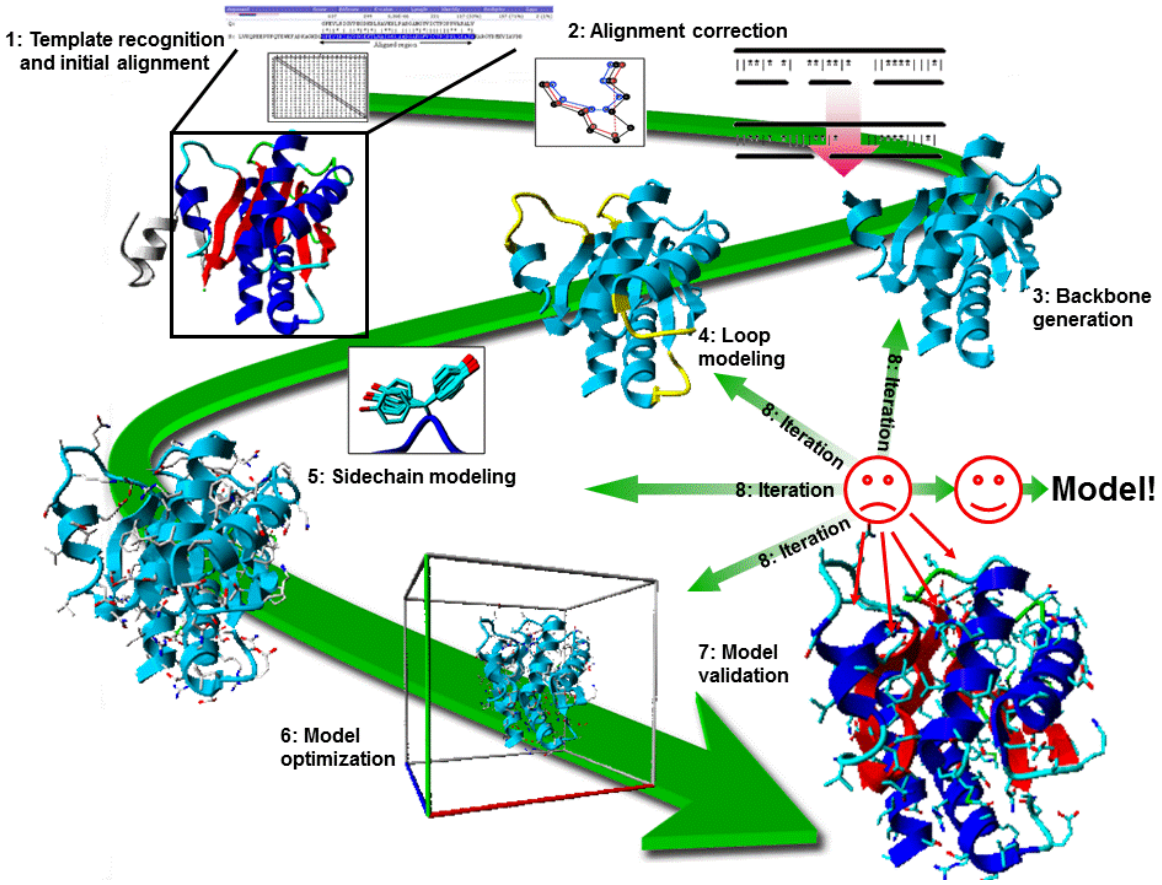
Rosetta Commons
The hub for Rosetta modeling software

- Simons et al., JMB 1997: 268:209

Homology modelling

A multi-step process, possibly with several iterations





http://swift.cmbi.ru.nl/teach/B4/drgdes_3.html

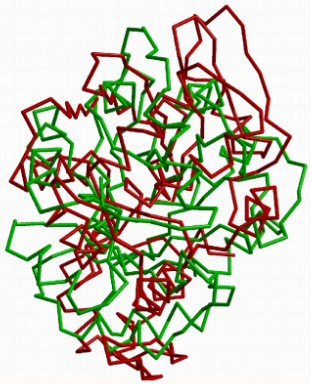
Typical errors in comparative modelling

MODEL

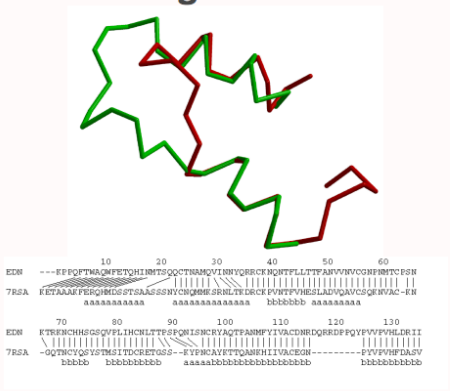
X-RAY

TEMPLATE

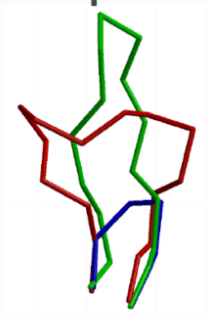
Incorrect template



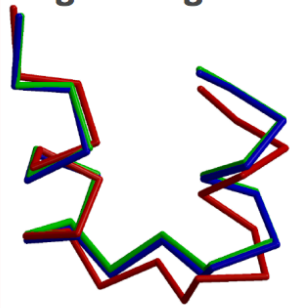
Misalignment



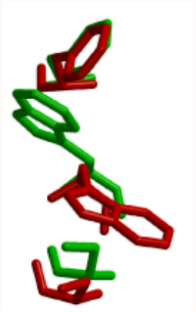
Region without a template



Distortion/shifts in aligned regions



Sidechain packing

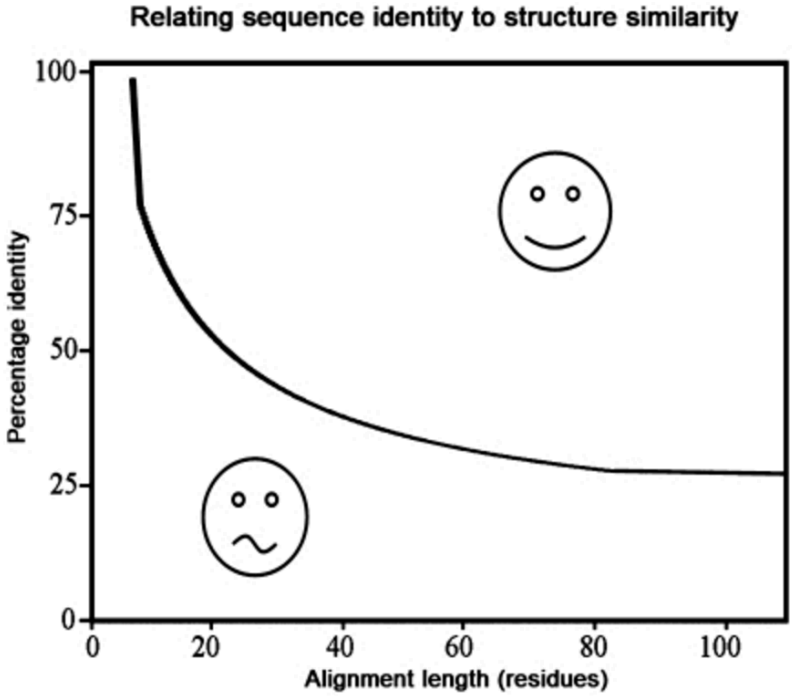


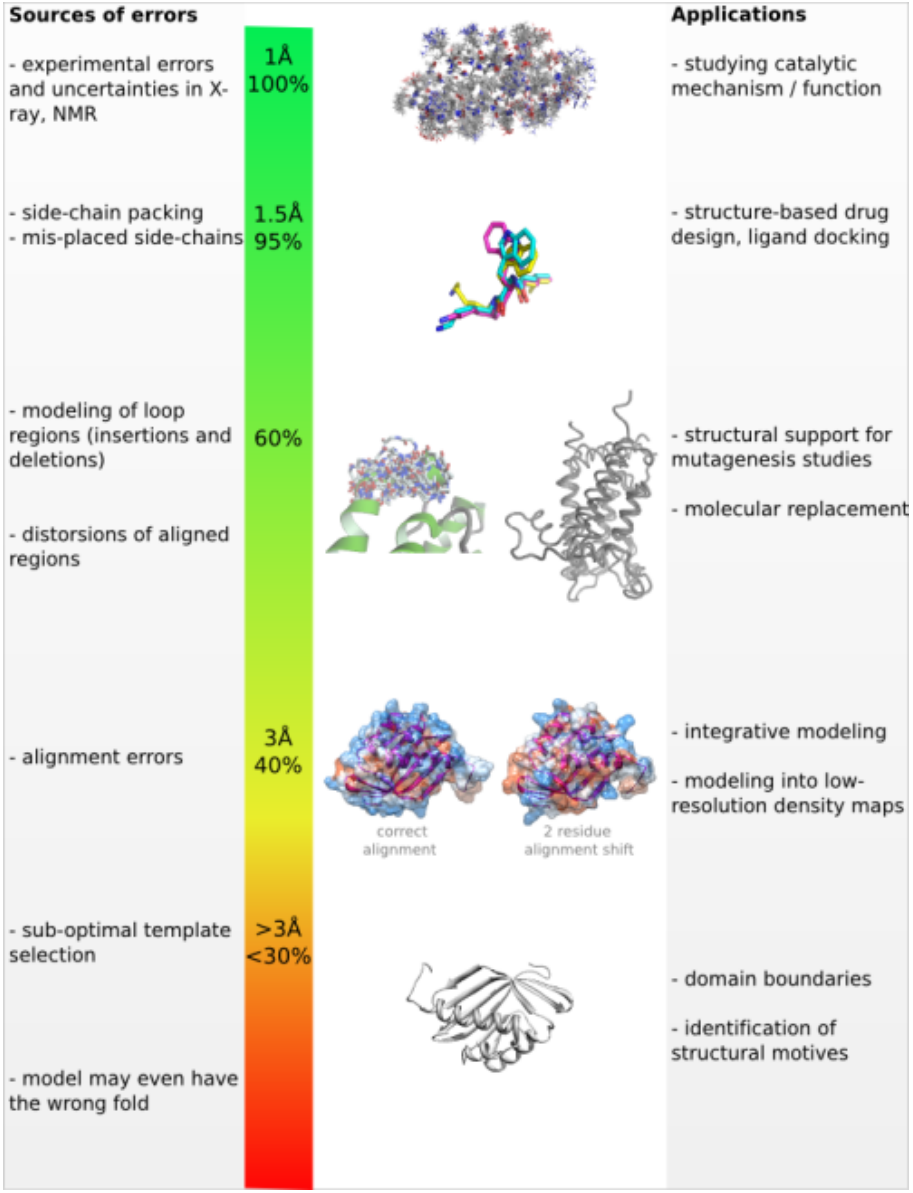
Marti-Renom *et al.* Annu.Rev.Biophys.Biomol.Struct. **29**, 291-325, 2000.

M.Topf

Model evaluation

- All models wrong: question is by how much.
- Evaluate:

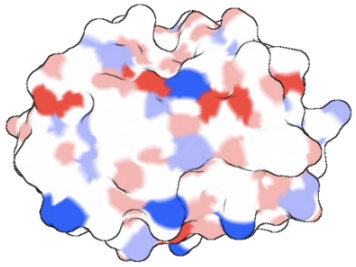




Practical 04-05

Predicting secondary structure from sequence

Predicting tertiary structure from sequence using homology modelling

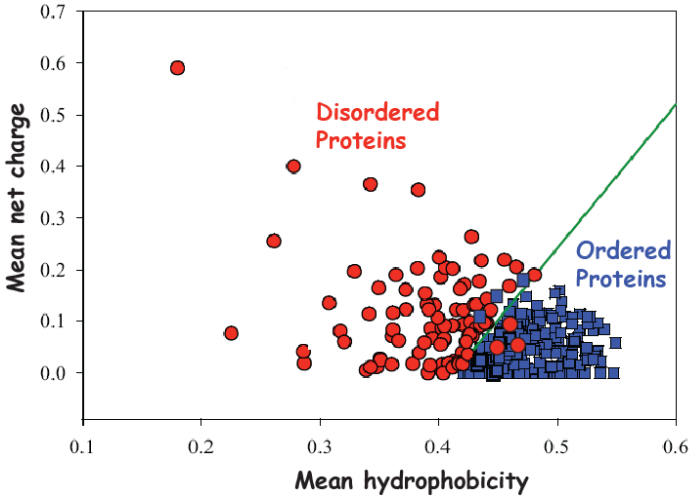


Section 5

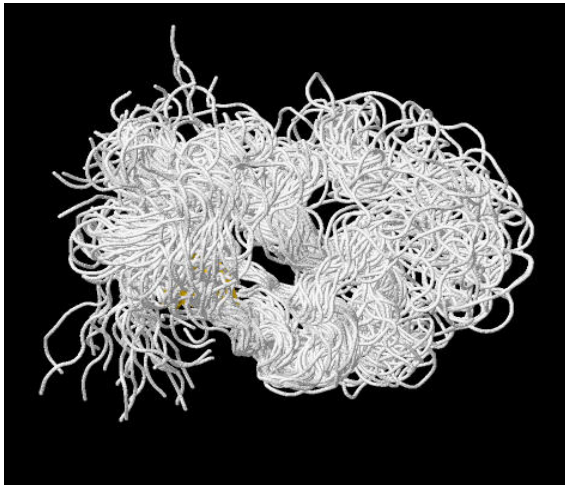
It is not all about structure

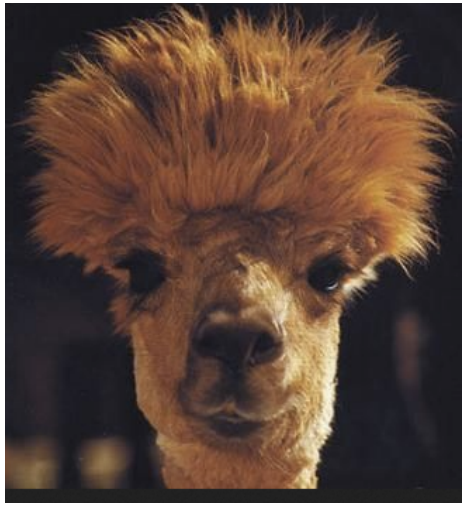
It is not all about structure: unstructured proteins

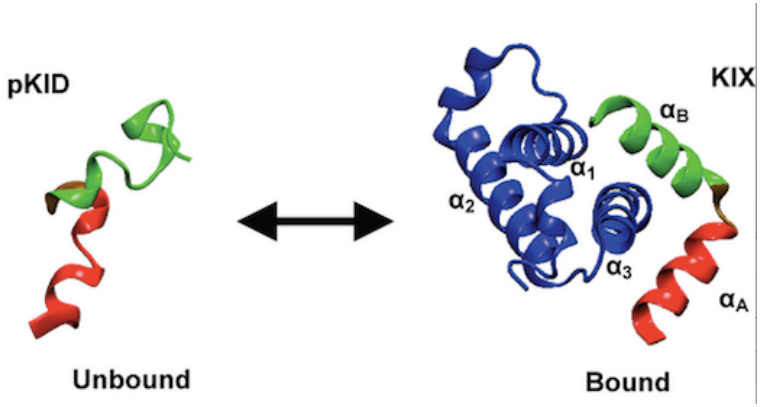
- Intrinsically unstructured protein common in proteomes (IUPs)
- Coupled folding and binding (fly-casting)



Uversky et al, 2000, *Proteins* 41:415-4



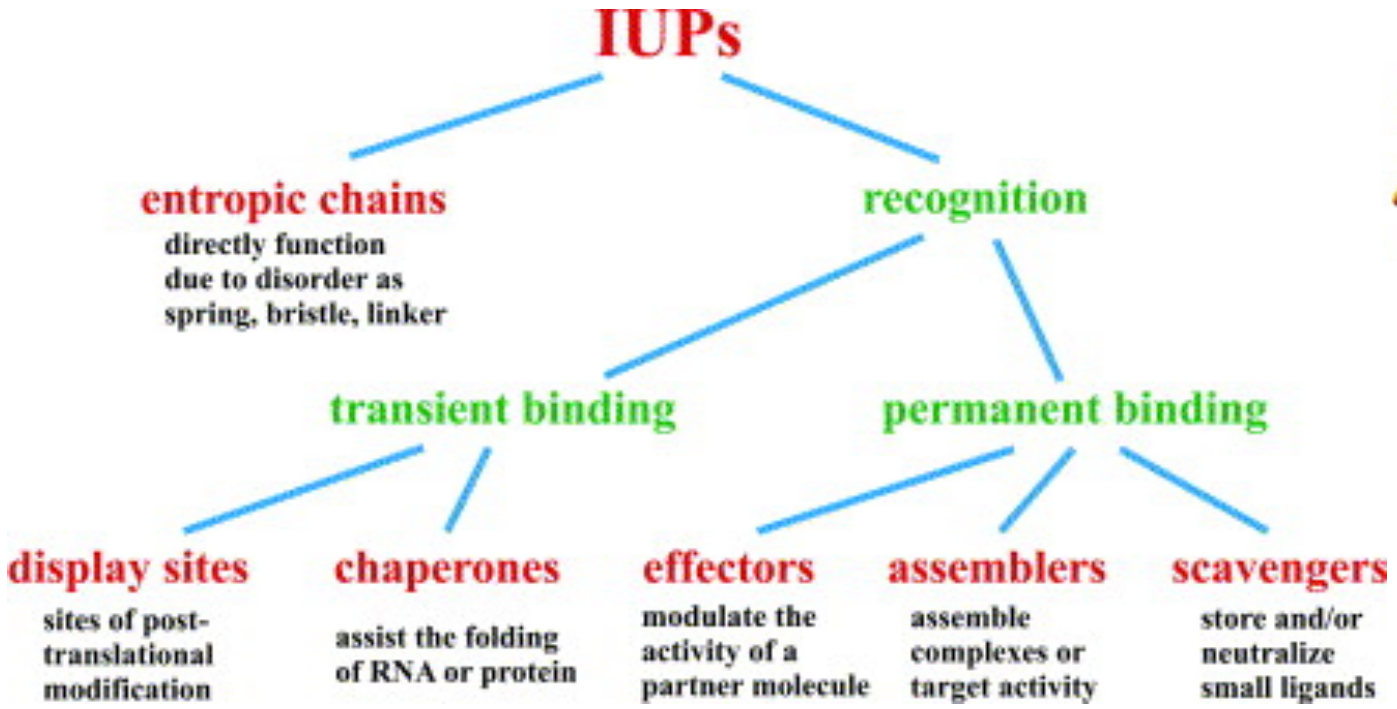


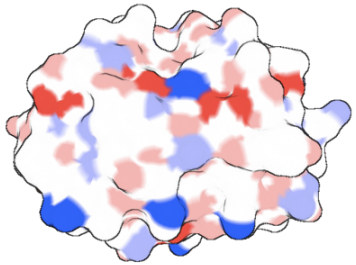


Citation: Turjanski AG, Gutkind JS, Best RB, Hummer G (2008) Binding-Induced Folding of a Natively Unstructured Transcription Factor. PLoS Comput Biol 4(4): e1000060. doi:10.1371/journal.pcbi.1000060

It is not all about structure: unstructured proteins

- Function directly linked to structural disorder
- Binding to multiple partners



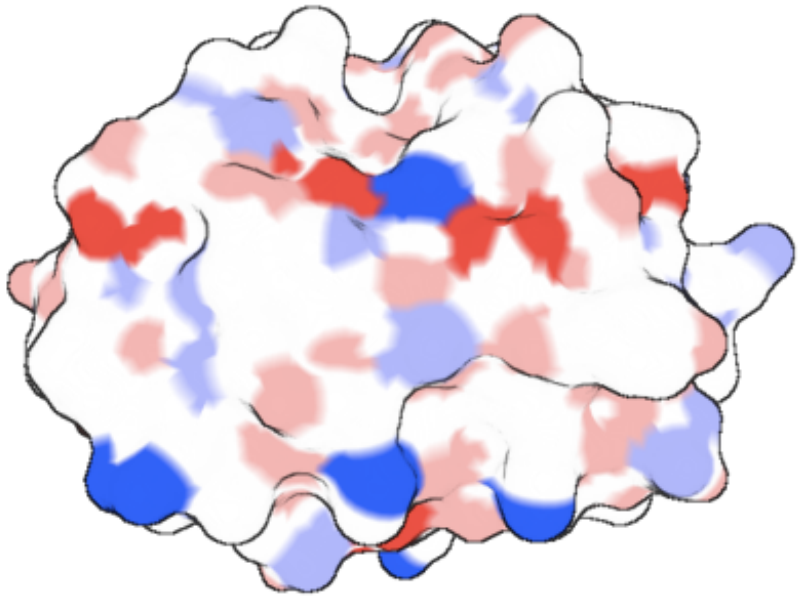


Section 6

Summary and conclusions
Acknowledgments

Summary

- Structures solved by crystallography, NMR and cryo-EM
- Data stored in RCSB
- Secondary, tertiary and quaternary structure
- Sequence/structure/function relationships are complex
- To fully understand function you need to know the structure



Acknowledgements

- Many slides provided by
  - Dr Irilenia Nobeli (Birkbeck College, London)
- Additional slides
  - Professor Christine Orengo (UCL, London)
  - Professor Maya Topf (Birkbeck College, London)
- Anyone who ever posted protein images or slides on the internet!

Seminar BINGO!

To play, simply print out this bingo sheet and attend a departmental seminar.

Mark over each square that occurs throughout the course of the lecture.

The first one to form a straight line (or all four corners) must yell out BINGO!! to win!

SEMINAR BINGO

Speaker bashes previous work	Repeated use of "um…"	Speaker sucks up to host professor	Host Professor falls asleep	Speaker wastes 5 minutes explaining outline
Laptop malfunction	Work ties in to Cancer/HIV or War on Terror	"…et al."	You're the only one in your lab that bothered to show up	Blatant typo
Entire slide filled with equations	"The data *clearly* shows…"	FREE Speaker runs out of time	Use of Powerpoint template with blue background	References Advisor (past or present)
There's a Grad Student wearing same clothes as yesterday	Bitter Post-doc asks question	"That's an interesting question"	"Beyond the scope of this work"	Master's student bobs head fighting sleep
Speaker forgets to thank collaborators	Cell phone goes off	You've no idea what's going on	"Future work will…"	Results conveniently show improvement