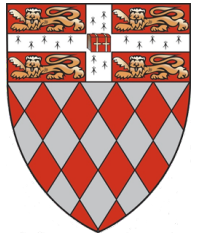
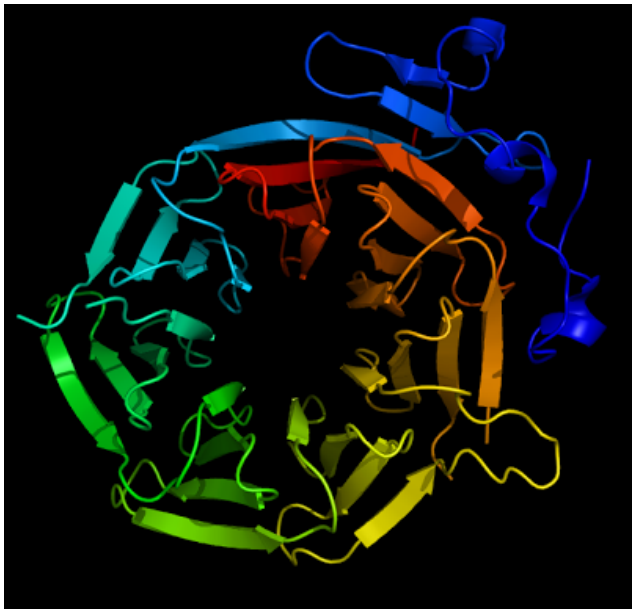
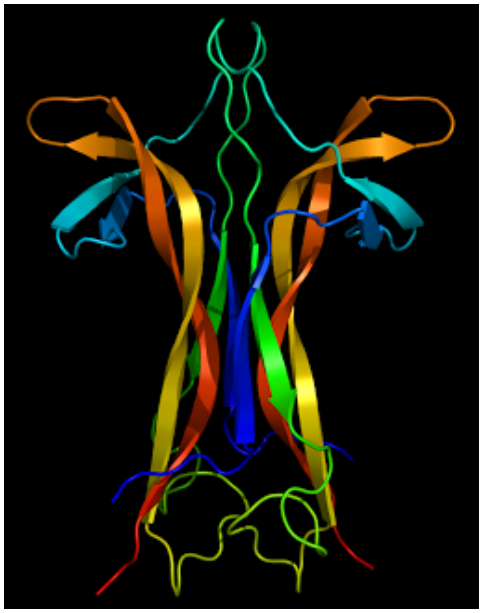


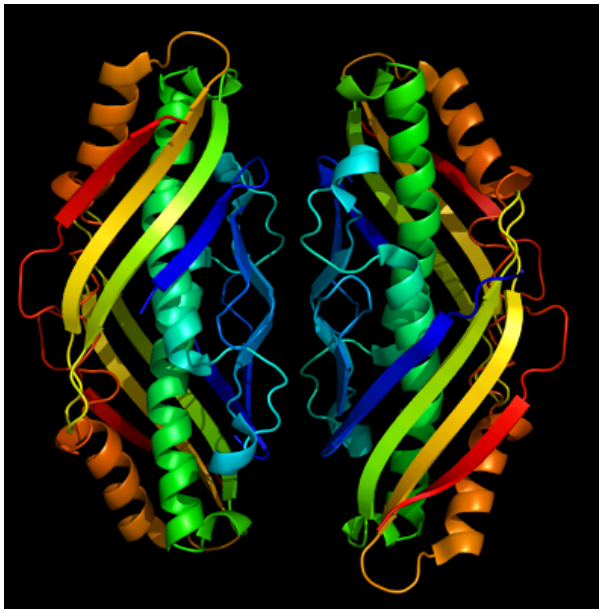
Structural Bioinformatics

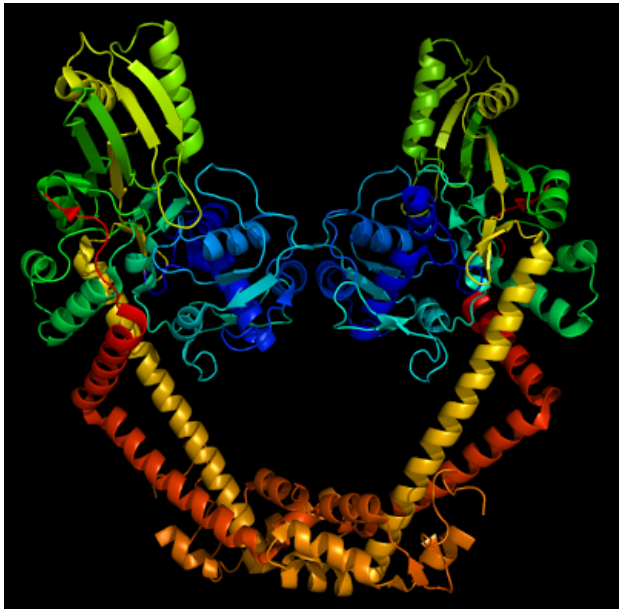
Alexey Morgunov

Fitzwilliam College & MRC-LMB, Cambridge

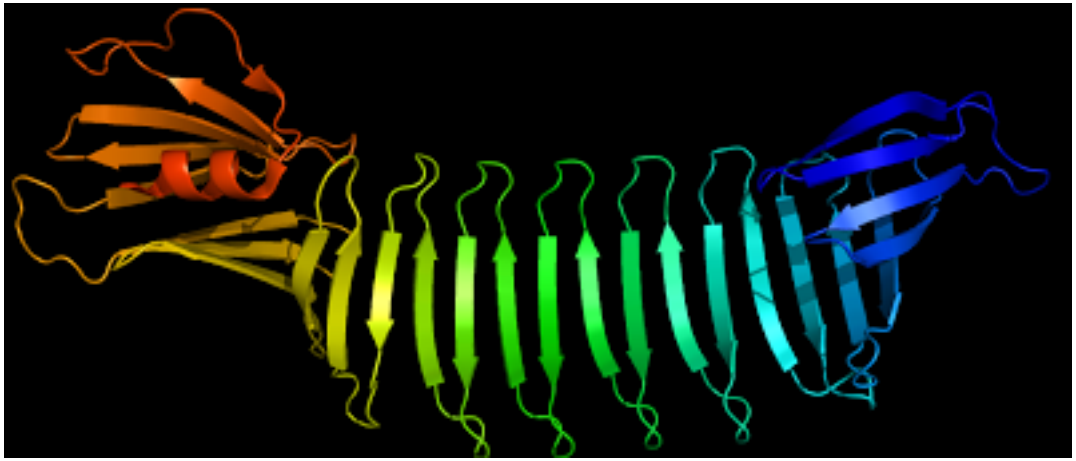








http://beautifulproteins.blogspot.co.uk/



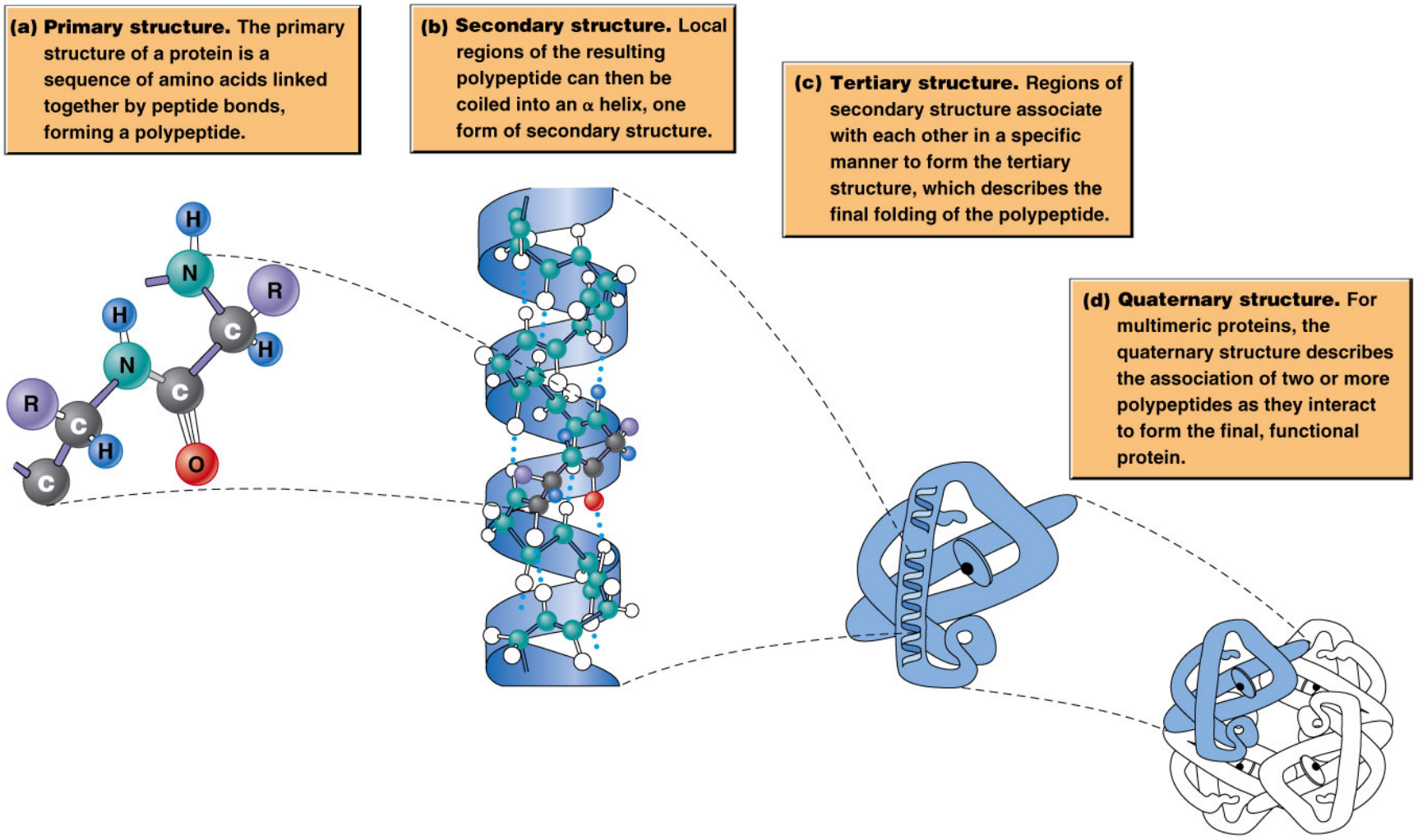
Aims of this lecture

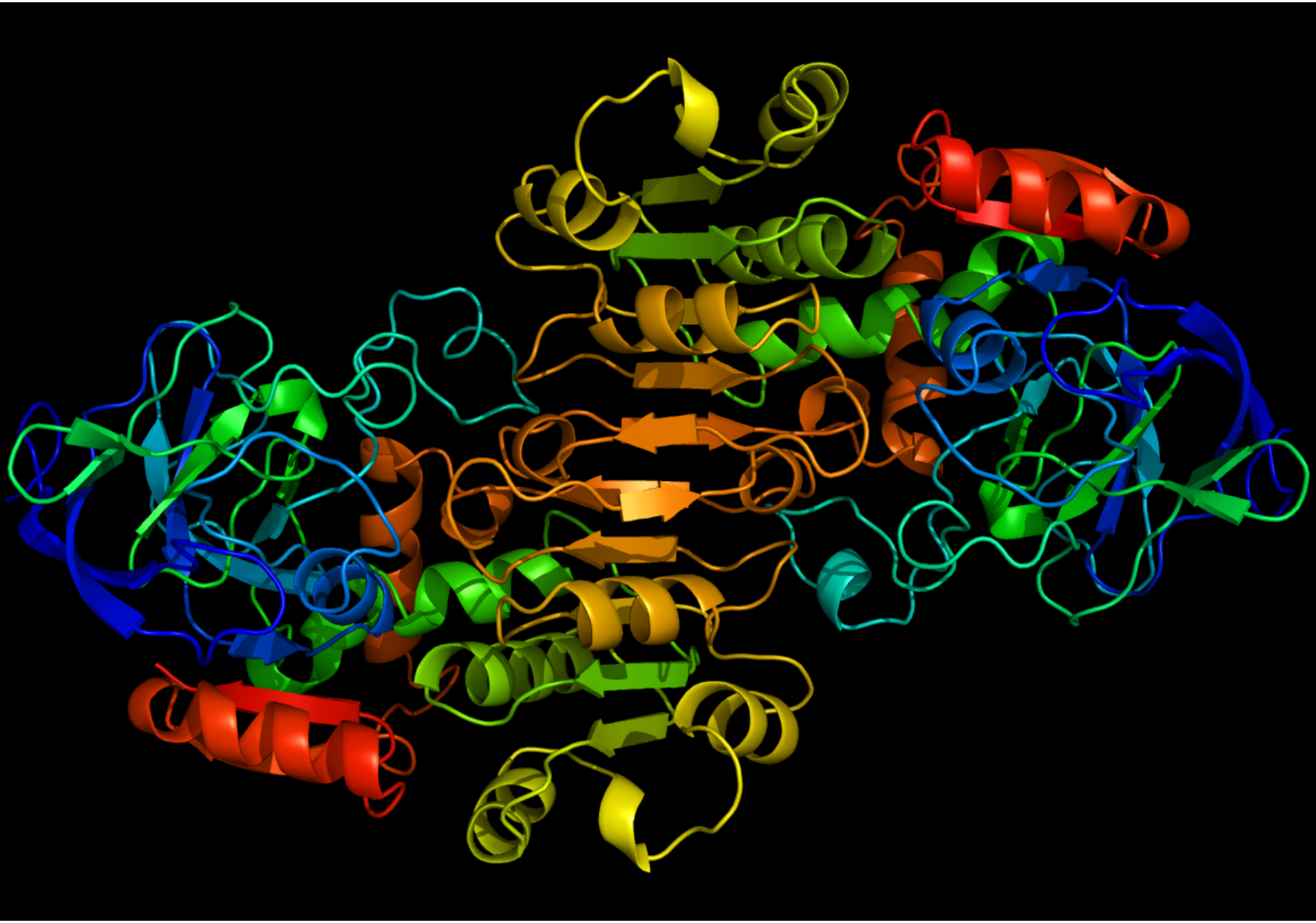
- A superficial introduction to a large and diverse field
- Describing the landscape of resources available
- Focus on what knowledge can be gained by analysing protein structures
- Introducing some future directions and applications

What is structural bioinformatics?

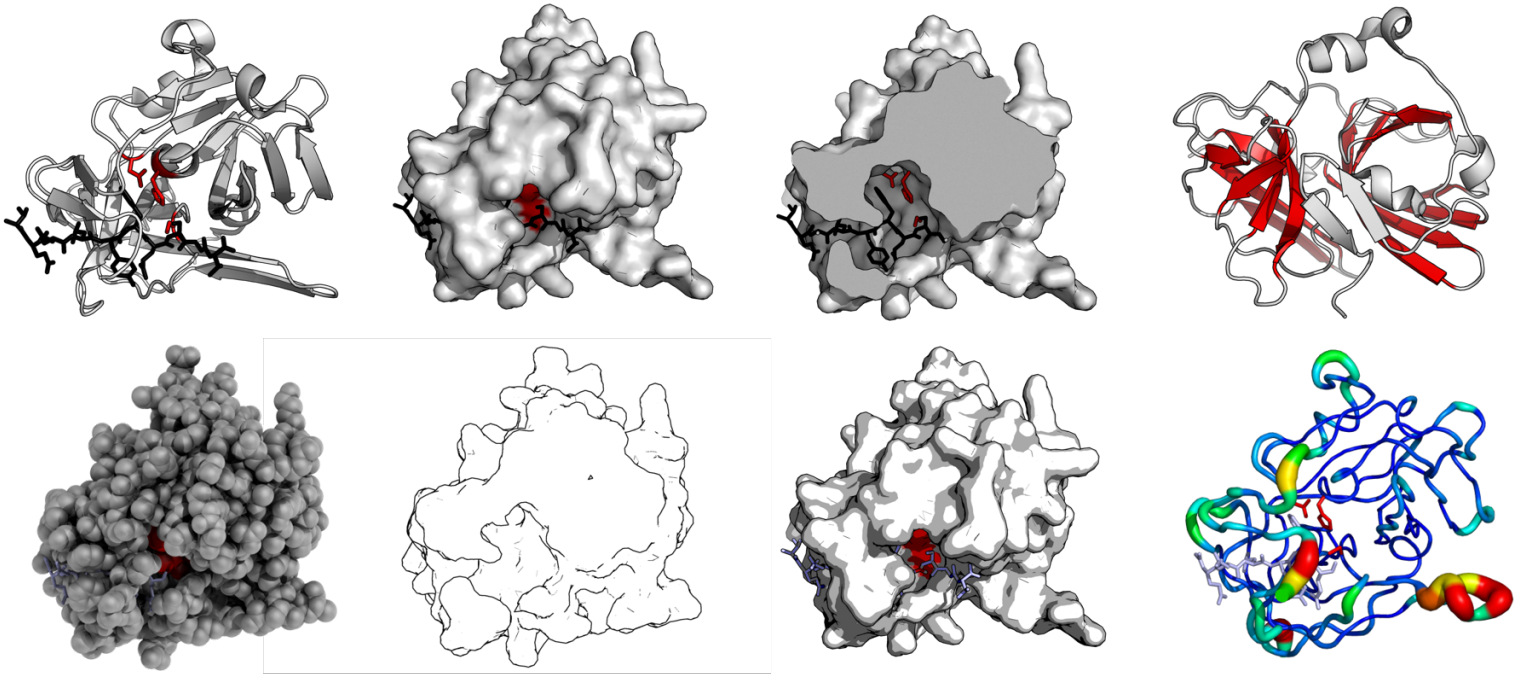
- Categorisation and descriptive analysis
- Comparative methods and techniques
- Theoretical foundations
- Predictive and design capability

Protein structure refresher



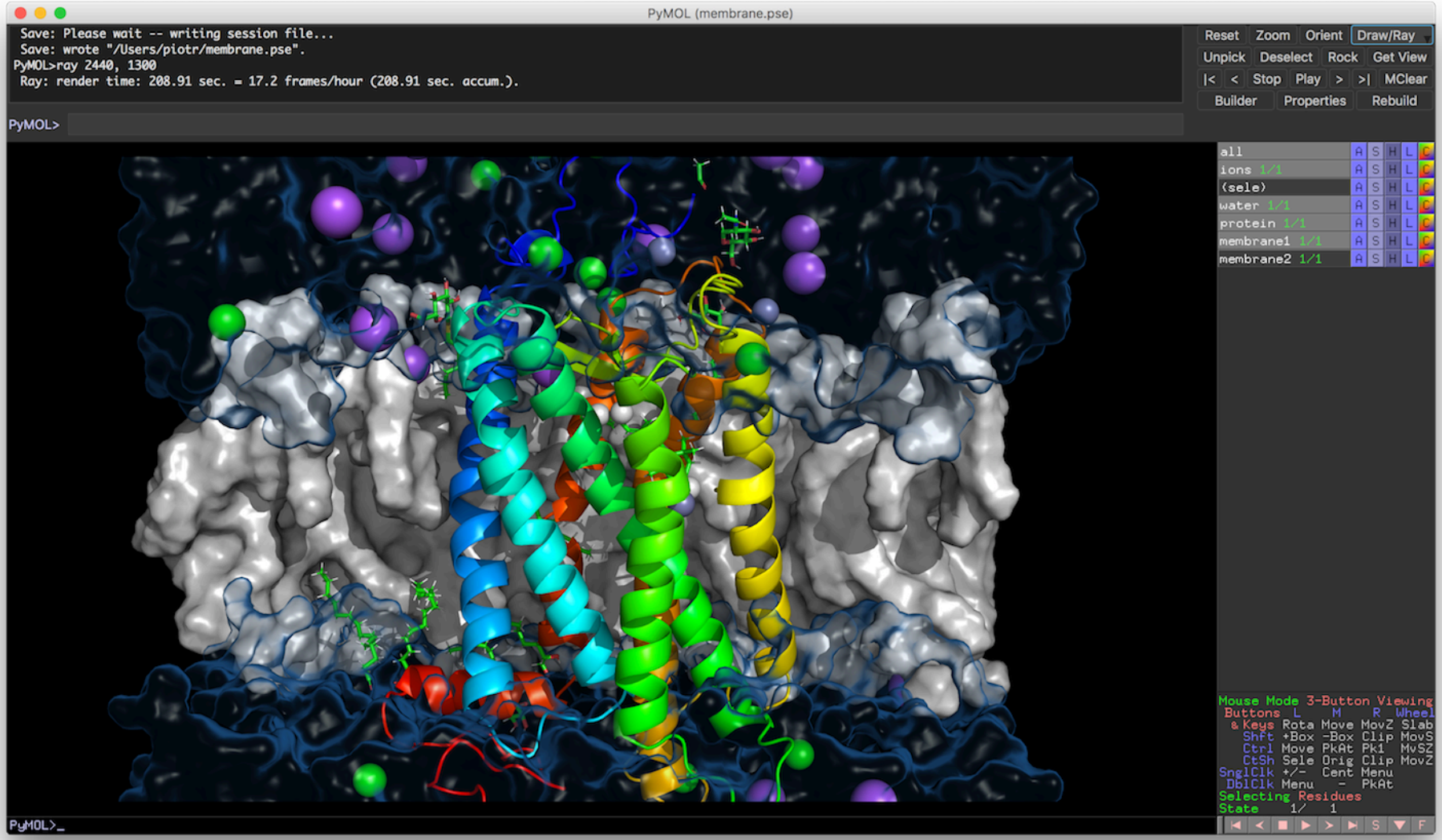


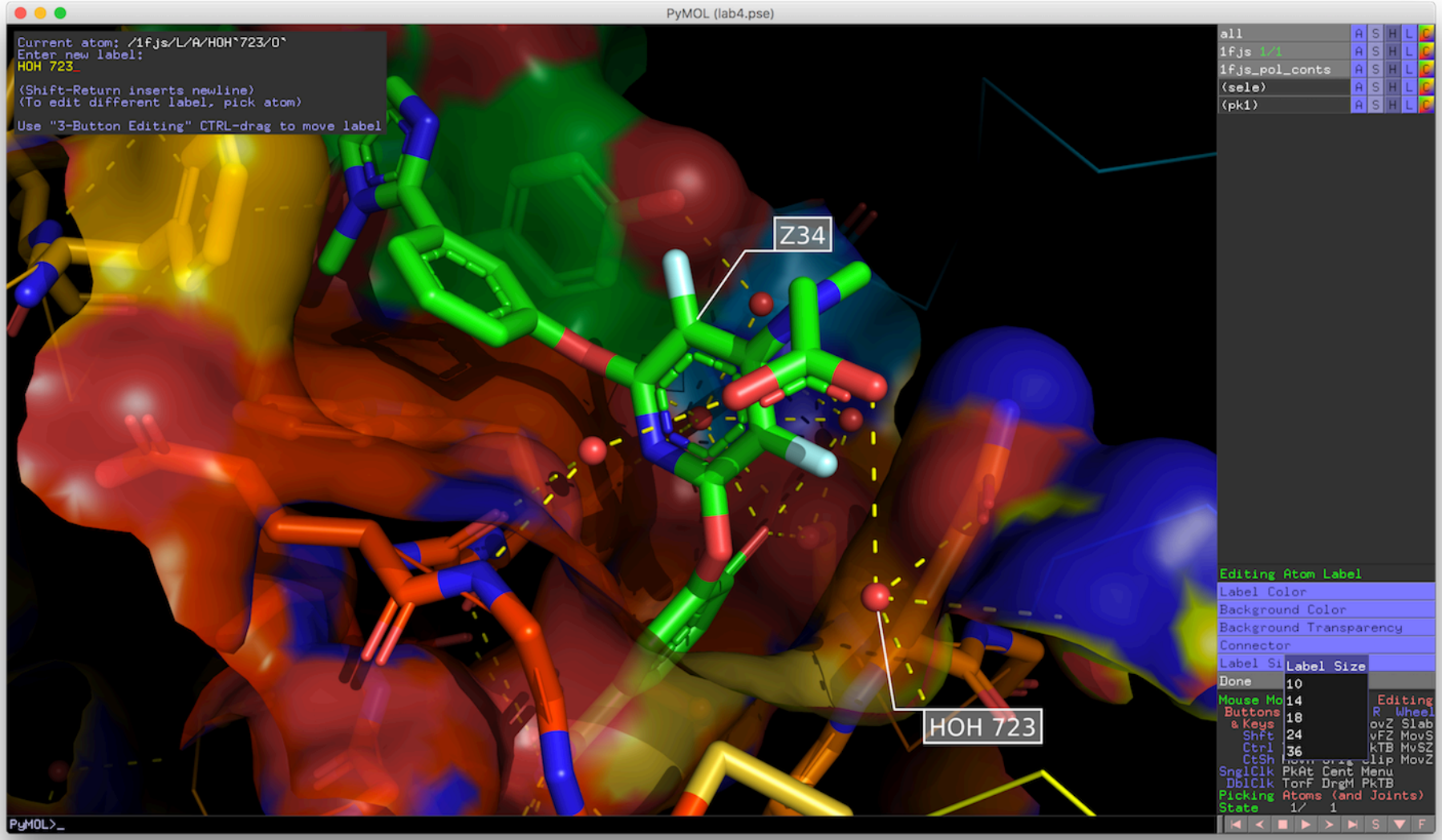
Same protein, different views



Molecular graphics systems

- Jmol & JSmol - http://jmol.sourceforge.net/
- PyMOL - http://www.pymol.org/
- RasMol - http://www.rasmol.org/
- UCSF Chimera - http://www.rbvi.ucsf.edu/chimera/
- VMD - http://www.ks.uiuc.edu/Research/vmd
- 3Dmol.js - http://3dmol.csb.pitt.edu/

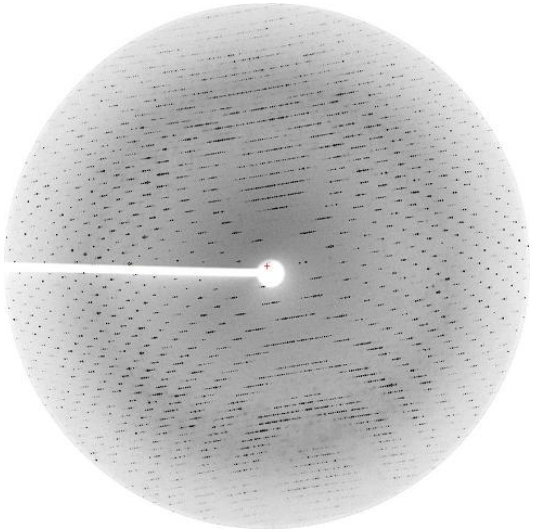

PyMOL (membrane.pse)
Save: Please wait -- writing session file...
Save: wrote "/Users/piotr/membrane.pse".
PyMOL>ray 2440, 1300
Ray: render time: 208.91 sec. = 17.2 frames/hour (208.91 sec. accum.).
Reset
Zoom
Orient
Draw/Ray
Unpick
Deselect
Rock
Get View
|<
<
Stop
Play
>
>|
MClear
Builder
Properties
Rebuild
PyMOL>
all
ions 1/1
(sele)
water 1/1
protein 1/1
membrane1 1/1
membrane2 1/1
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl Move PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State 1/ 1
PyMOL>_

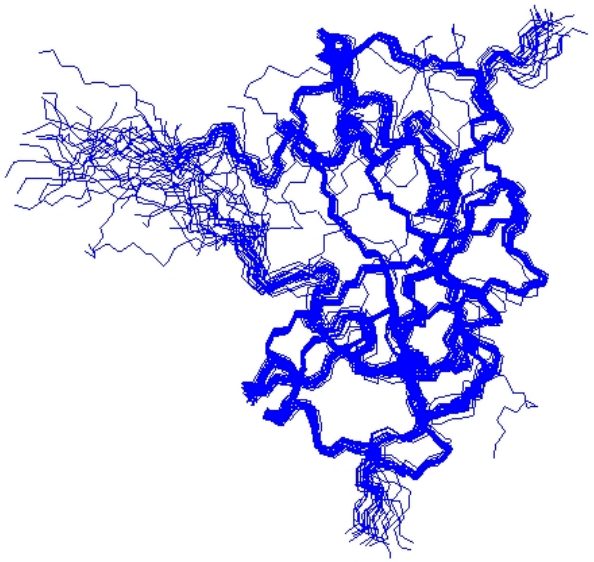

PyMOL (lab4.pse)
Current atom: /1fjs/L/A/HOH`723/O`
Enter new label:
HOH 723_
(Shift-Return inserts newline)
(To edit different label, pick atom)
Use "3-Button Editing" CTRL-drag to move label
Z34
HOH 723
all
1fjs 1/1
1fjs_pol_conts
(sele)
(pk1)
Editing Atom Label
Label Color
Background Color
Background Transparency
Connector
Label Size
Done
10
14
18
24
36
Editing
SnglClk PkAt Cent Menu
DblClk TorF DrgM PkTB
Picking Atoms (and Joints)
State 1/ 1
PyMOL>_

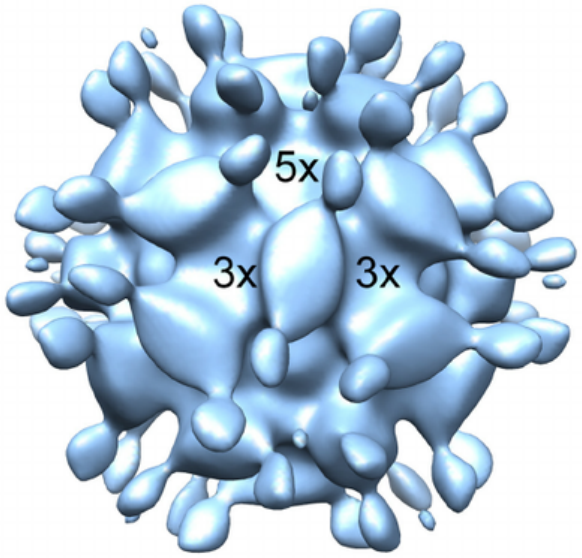
Experimental techniques for determining protein structure models

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR)
- Cryo-Electron Microscopy (Cryo-EM)

https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/methods-for-determining-structure

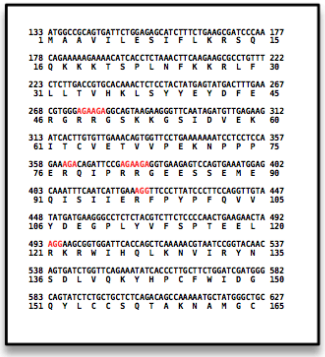




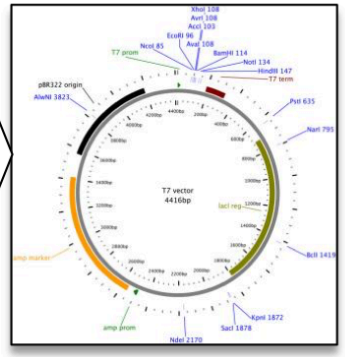


X-ray crystallography

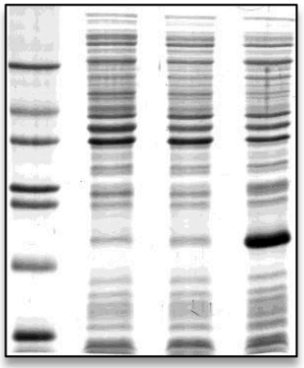
Sequence analysis



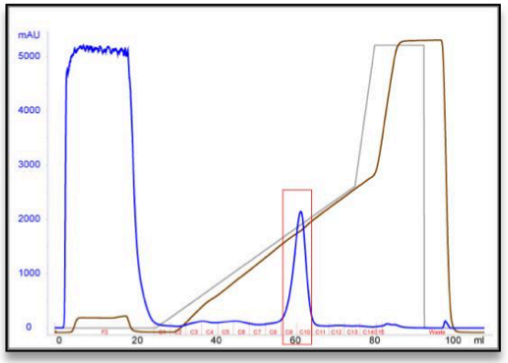
Expression construct



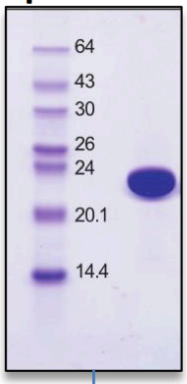
Recombinant expression



Protein purification

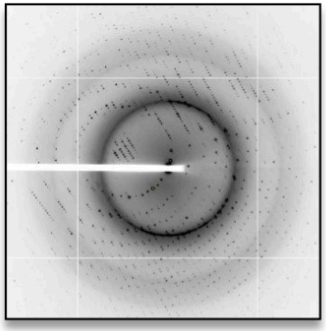


Homogenous protein

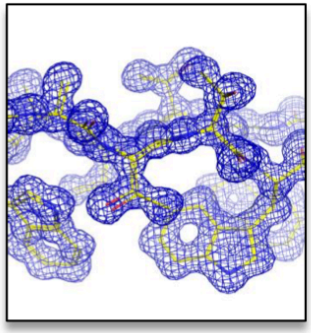


Crystallisation

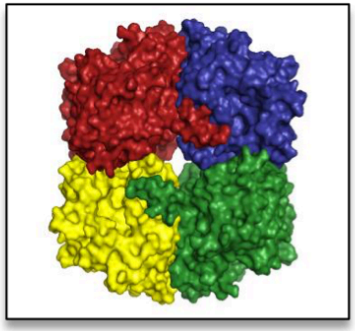
Diffraction data collection



Phase problem solution
Fit model to electron density



The final structure



Cryo-EM

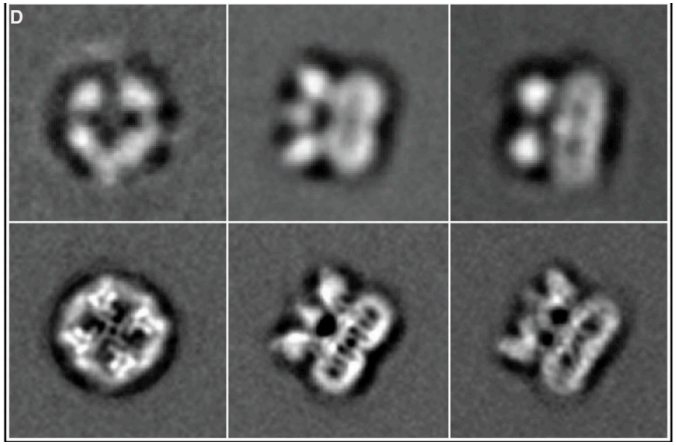
Can now do very high resolution

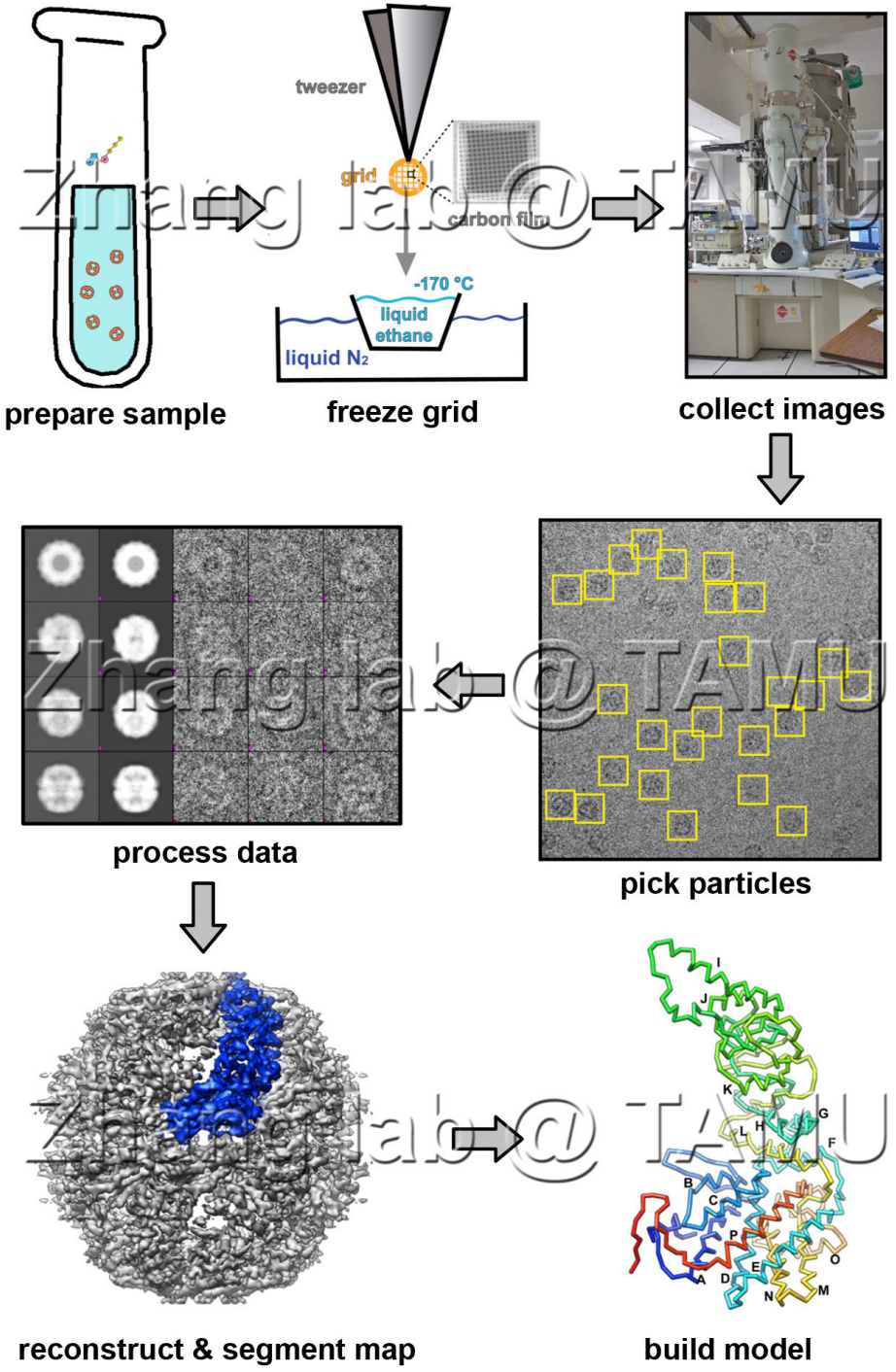
Less sample than X-ray

No need for crystals

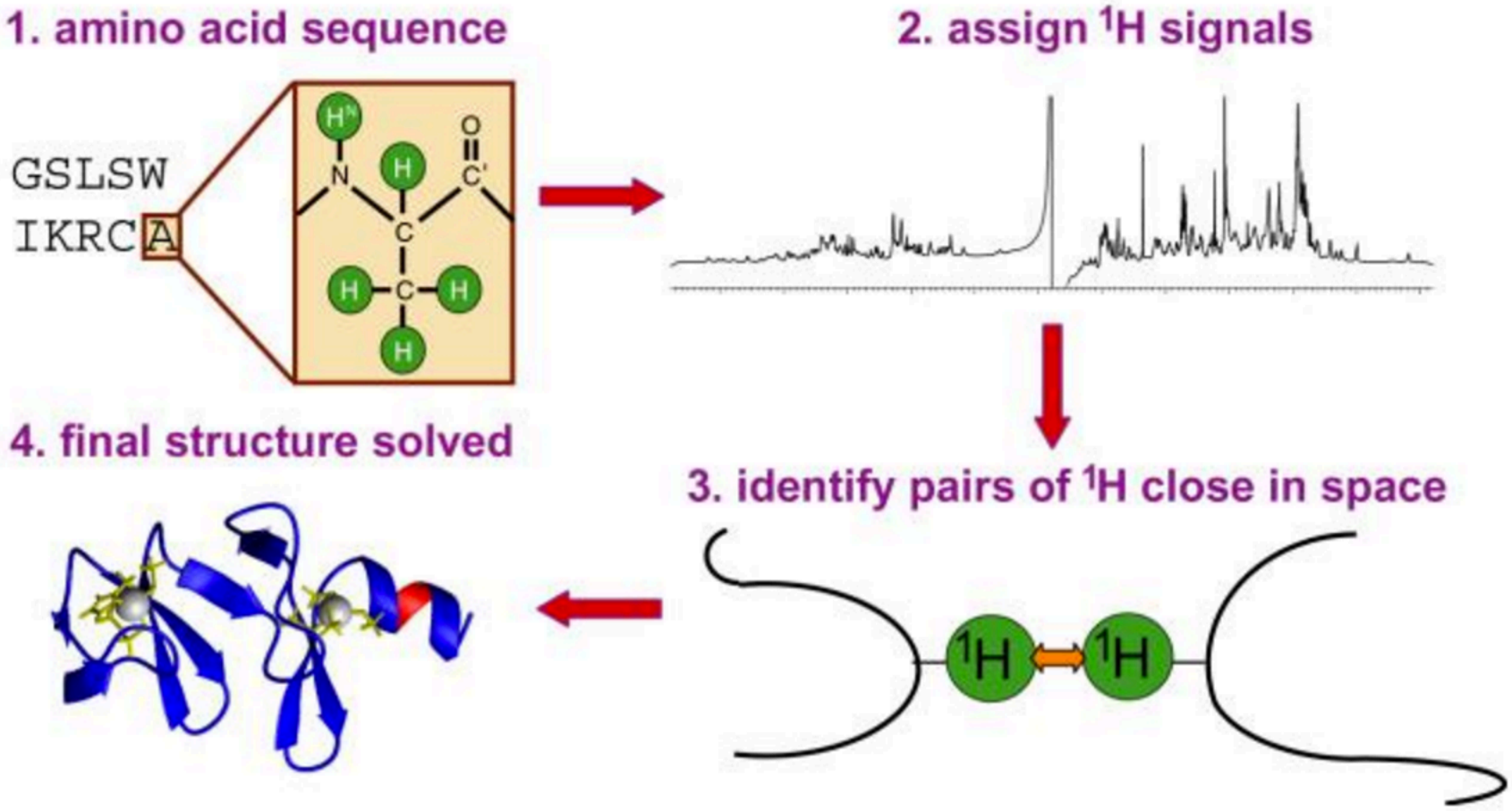
Much faster

https://doi.org/10.1016/j.tibs.2014.10.005





NMR


1. amino acid sequence
GSLSW
IKRCA
2. assign ^{1}H signals
3. identify pairs of ^{1}H close in space
^{1}H
^{1}H
4. final structure solved

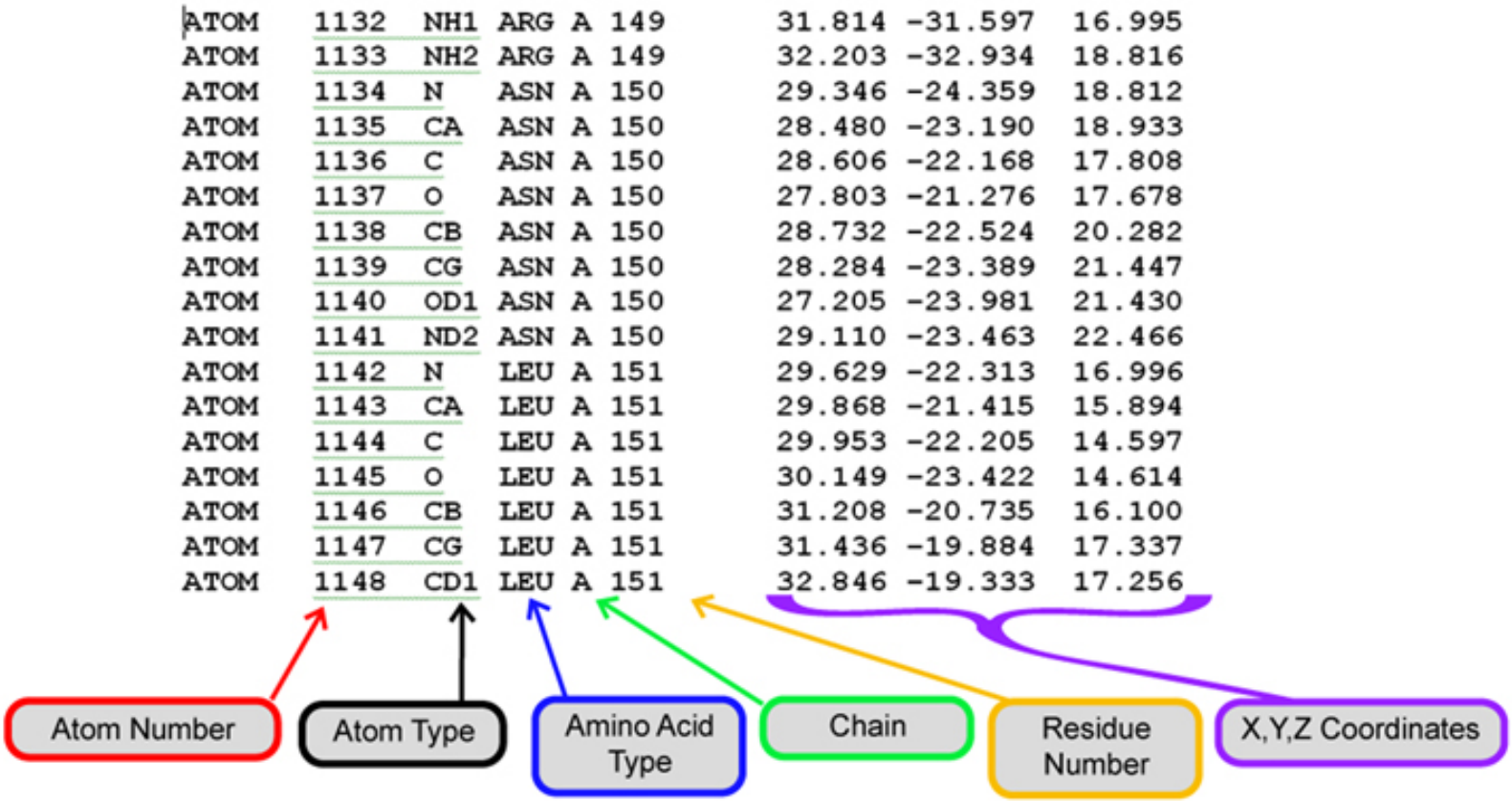
NMR

Ensemble of solutions to distance constraints

More rigid regions are more precisely determined

Useful in studying protein dynamics

How do we store a structural model? PDB file format



How do we store a structural model? PDB file format

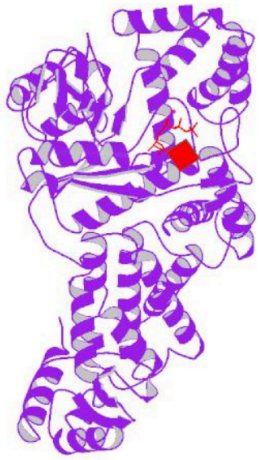
```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

Structural model databases

- Protein Data Bank (PDB) - http://www.rcsb.org/ or http://www.ebi.ac.uk/pdbe/
- Electron Microscopy Data Bank (EMDB) - https://www.ebi.ac.uk/pdbe/emdb/

PDBe
Protein Data Bank in Europe

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers
Protein Workshop | Ligand Explorer

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A

Biological assembly 1 assigned by authors.

Display Files | Download Files

2ZUE

Crystal structure of Pyrococcus horikoshii arginyl-tRNA synthetase complexed with tRNA(Arg) and an ATP analog (ANP)

DOI: 10.2210/pdb2ZUE/pdb **NDB:** PR0357

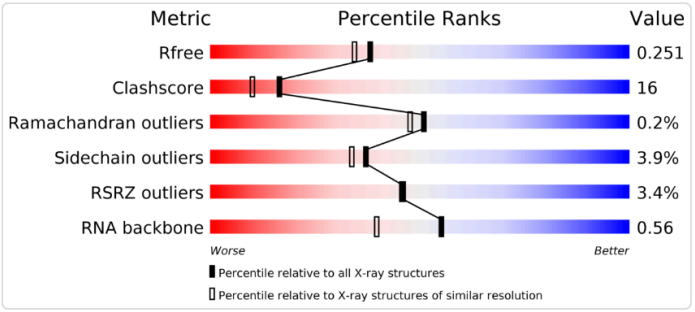
Classification: Ligase/RNA
Organism(s): Pyrococcus horikoshii (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)
Expression System: Escherichia coli BL21(DE3)

Deposited: 2008-10-16 **Released:** 2009-08-18
Deposition Author(s): Konno, M., Sumida, T., Uchikawa, E., Mori, Y., Yanagisawa, T., Sekine, S., Yokoyama, S.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2 Å
R-Value Free: 0.253
R-Value Work: 0.213

wwPDB Validation 3D Report | Full Report



This is version 1.1 of the entry. See complete history.

Macromolecules

Find similar proteins by: Sequence | Structure

Proteins 1 | Nucleic Acids / Hybrid 1

Entity ID: 1

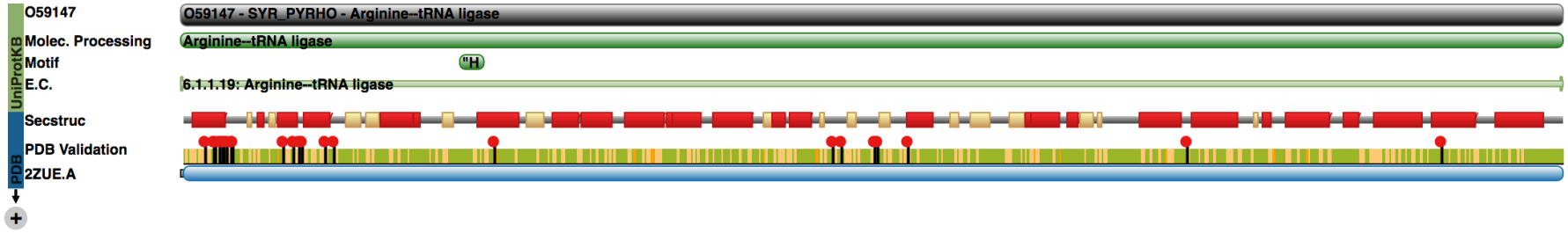
Molecule	Chains	Sequence Length	Organism	Details
Arginyl-tRNA synthetase	A	629	Pyrococcus horikoshii (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3)	**Gene Names**: argS **EC**: 6.1.1.19

Find proteins for O59147 *(Pyrococcus horikoshii (strain ATCC 700860 / DSM 12428 / JCM 9974 / NBRC 100139 / OT-3))*

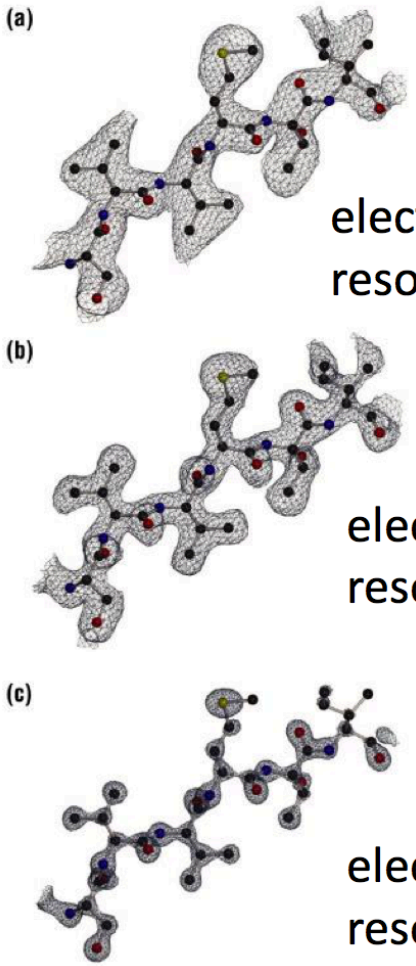
Go to UniProtKB: O59147

Protein Feature View

Full Protein Feature View for O59147



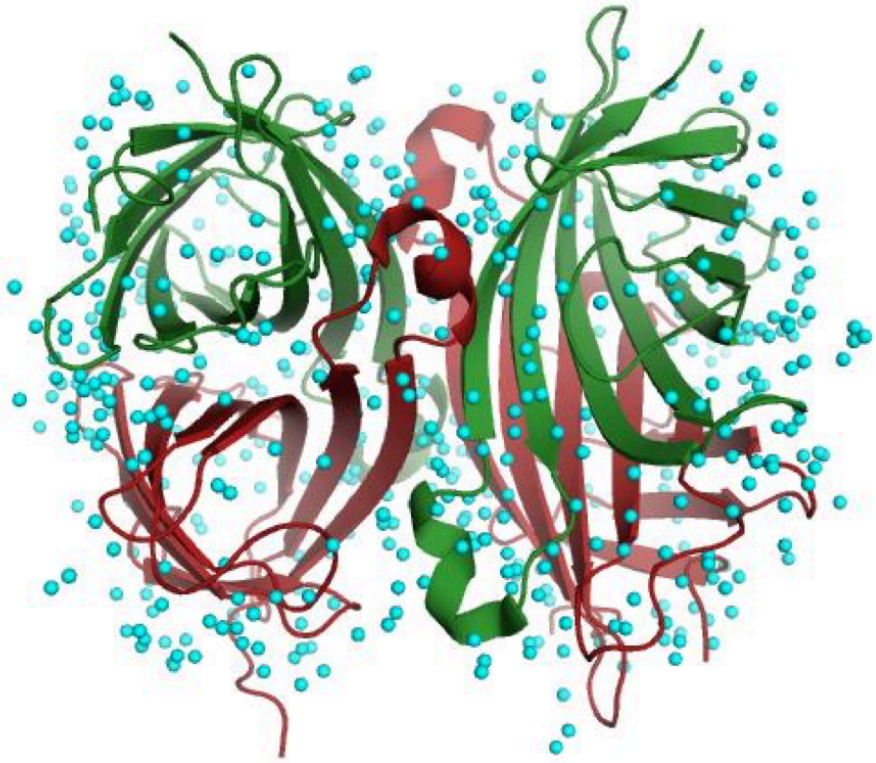
What is a good structural model?



electron density at 3Å resolution

electron density at 2Å resolution

electron density at 1Å resolution



Extensive water network seen in a high resolution structure

At high resolution, (1Å in this case) electron density of large numbers of water molecules surrounding the protein can be seen

What is a good structural model?

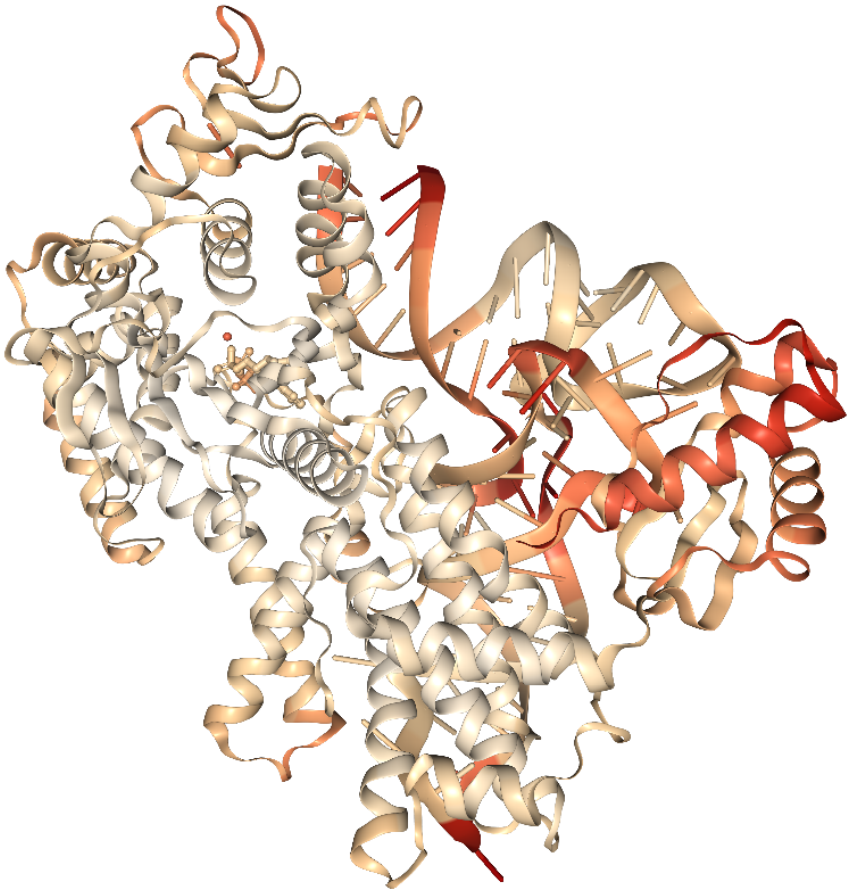
- *The quality of published molecular models is inversely related to the impacts of the journals in which they are published.* [ref]
- Missing residues and atoms − no electron density
- Geometric and conformation validation criteria
  - Bond lengths, angles and planes
  - Protein backbone conformation (Ramachandran plot)
  - Protein sidechain conformations (rotamers)
  - All-atom contacts and clashes
  - Underpacking (holes in core)

  Analyses available: PDBREPORTS, MolProbity, WHAT_IF

What is a good structural model?

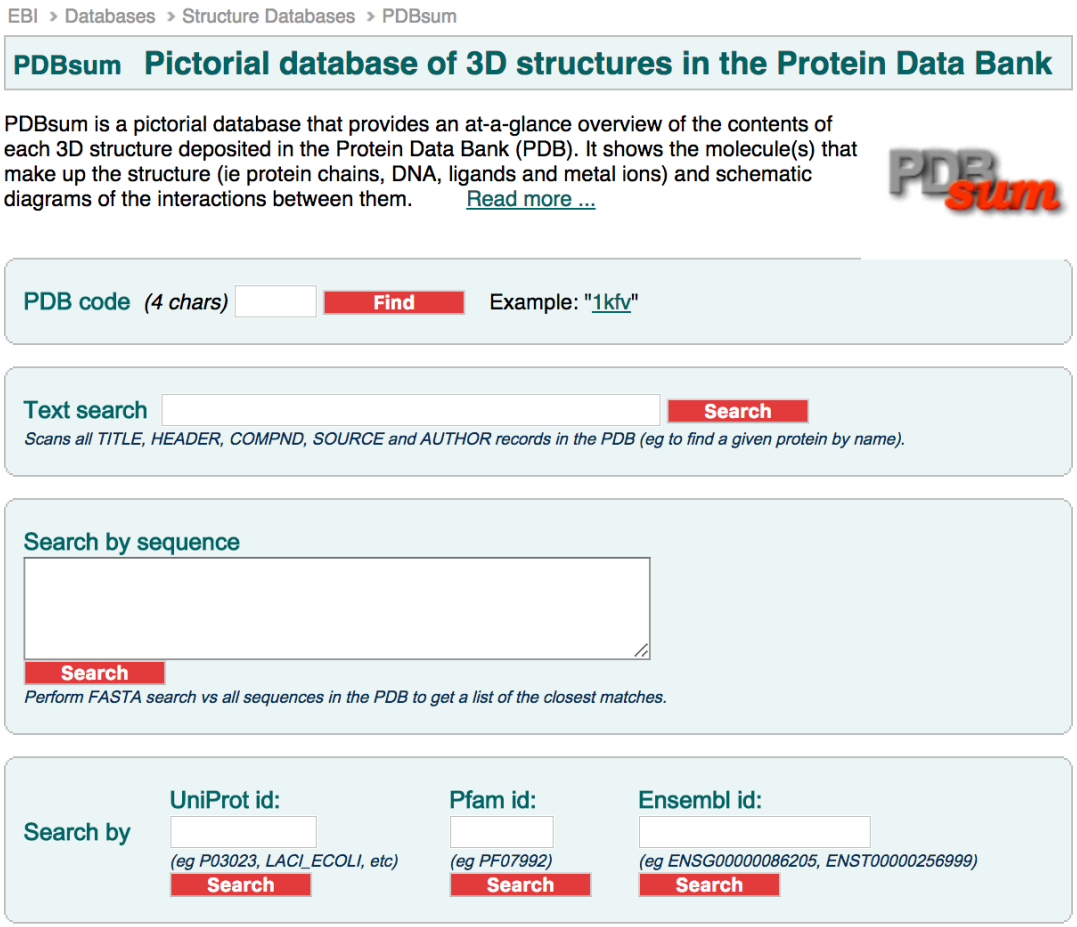
- Global quality − resolution
  - 2.5 Å is good enough
- Global quality − R value
  - a measure of error between the observed intensities from the diffraction pattern and the predicted intensities that are calculated from the model
  - a value of 0.20 or lower is considered good
  - not always reliable
- Global quality − free R value
  - calculated in the same manner as the R value, but from a subset of the data set aside for the calculation of free R, and not used in the refinement of the model
  - cross validation − free of refinement bias
  - should not exceed (resolution/10) by more than 0.05
- Local quality − temperature factor (B value)
  - uncertainty in the positions of atoms increases with disorder in the protein crystal
  - high temperature factor reflects a low empirical electron density for the atom
  - a value of less than 30 $Å^2$ signifies confidence in the position of a residue

2ZUE coloured by B factor

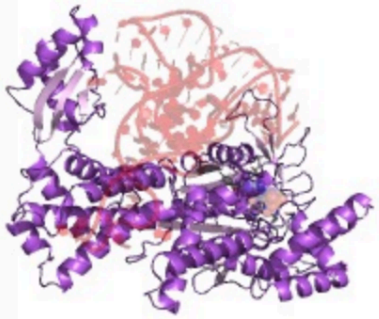


What can we learn from the structural model? Databases

- PDBsum - http://www.ebi.ac.uk/pdbsum/



Chain A (628 residues)



Protein chain A highlighted
(click to view)

Jmol Strap

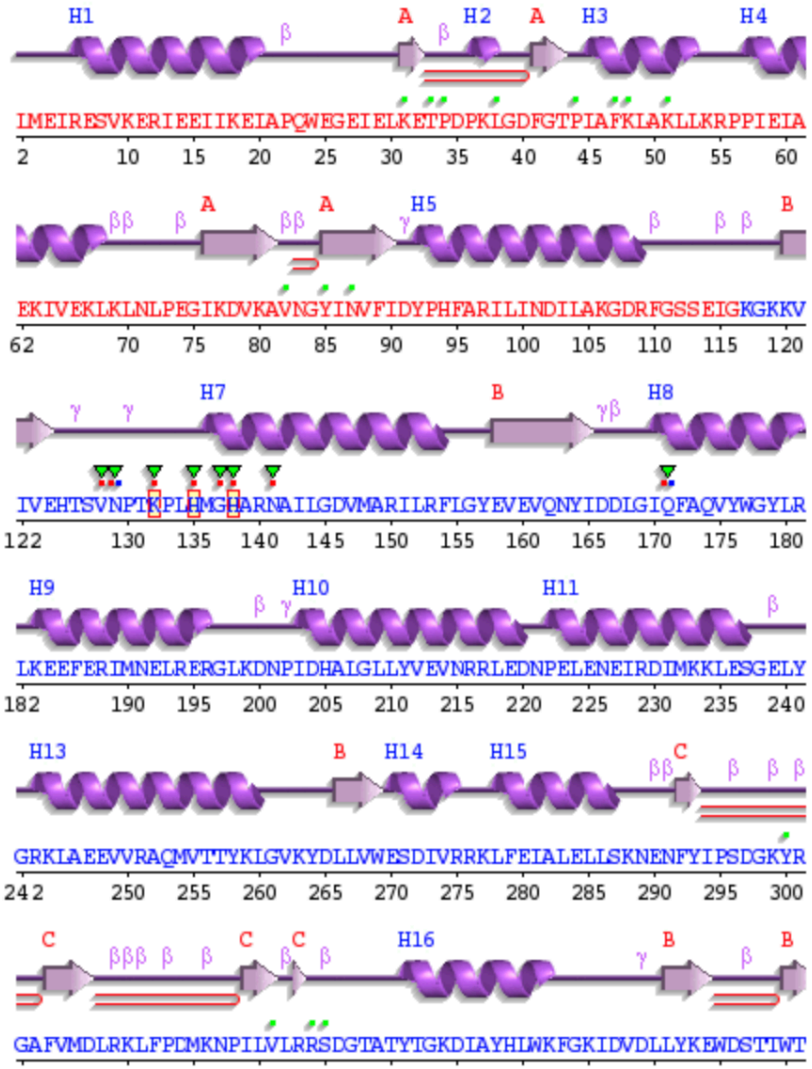
Motifs

- **Secondary structure**
  - **Wiring diagram**
  - Residue conservation
- **ProMotif**
  - 4 sheets
  - 4 beta alpha beta units
  - 5 beta hairpins
  - 3 beta bulges
  - 17 strands
  - 27 helices
  - 40 helix-helix interacs
  - 43 beta turns
  - 7 gamma turns
- **Catalytic residues**
  - H135-H138-K132

UniProt code: O59147 (SYR_PYRHO) [Pfam]

CATH **structural classification (3 domains) :**

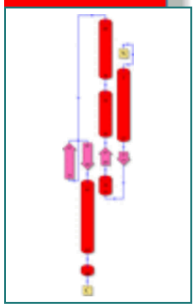
Domain	Links	CATH no.	Class	Architecture
1	CATH	3.30.1360.70	= *Alpha Beta*	*2-Layer Sandwich*
2	CATH	3.40.50.620	= *Alpha Beta*	*3-Layer(aba) Sandwich*
3	CATH	1.10.730.10	= *Mainly Alpha*	*Orthogonal Bundle*



Jmol

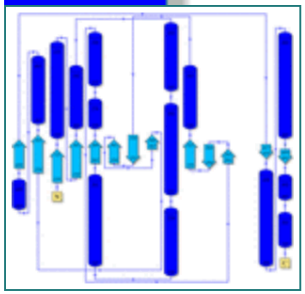
FASTA file

Domain 1



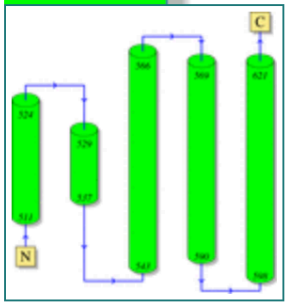
Hera

Domain 2



Hera

Domain 3



Hera

What can we learn from the structural model? Classification

- Structural Classification of Proteins (SCOP)
  - Original (unmaintained): http://scop.mrc-lmb.cam.ac.uk/scop
  - SCOP2 (in development): http://scop2.mrc-lmb.cam.ac.uk/
  - SCOPe (extension of the original): http://scop.berkeley.edu/
  - Hierarchical classification levels:
    - **Class**: type of fold (see next slide)
    - **Fold**: major structural similarity
    - **Superfamily**: probable common evolutionary origin
    - **Family**: clear evolutionary relationship
    - **Protein domain**: same protein unit
    - **Species**

Classes in SCOPe 2.06:

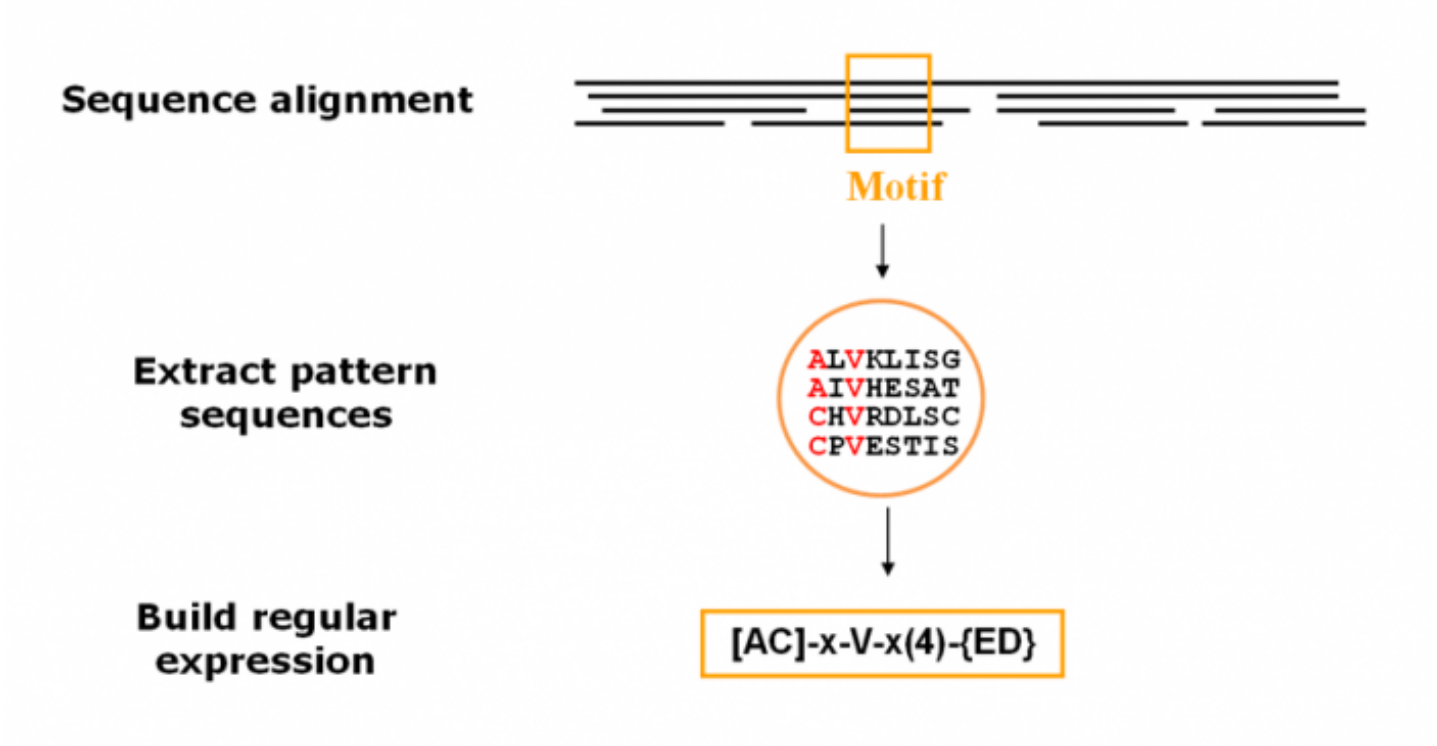
1.  a: All alpha proteins [46456] (289 folds)
2.  b: All beta proteins [48724] (177 folds)
3.  c: Alpha and beta proteins (a/b) [51349] (148 folds)
4.  d: Alpha and beta proteins (a+b) [53931] (385 folds)
5.  e: Multi-domain proteins (alpha and beta) [56572] (69 folds)
6.  f: Membrane and cell surface proteins and peptides [56835] (59 folds)
7.  g: Small proteins [56992] (94 folds)
8.  h: Coiled coil proteins [57942] (7 folds)
9.  i: Low resolution protein structures [58117] (25 folds)
10.  j: Peptides [58231] (133 folds)
11.  k: Designed proteins [58788] (44 folds)
12.  l: Artifacts [310555] (1 fold)

What can we learn from the structural model? Classification

- Structural Classification of Proteins (SCOP)
- Pfam - http://pfam.xfam.org/
  - large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)
- SUPERFAMILY - http://supfam.org/
  - database of structural and functional annotation for all proteins and genomes
- CATH - http://cathdb.info/
  - evolutionary relationships of protein domains

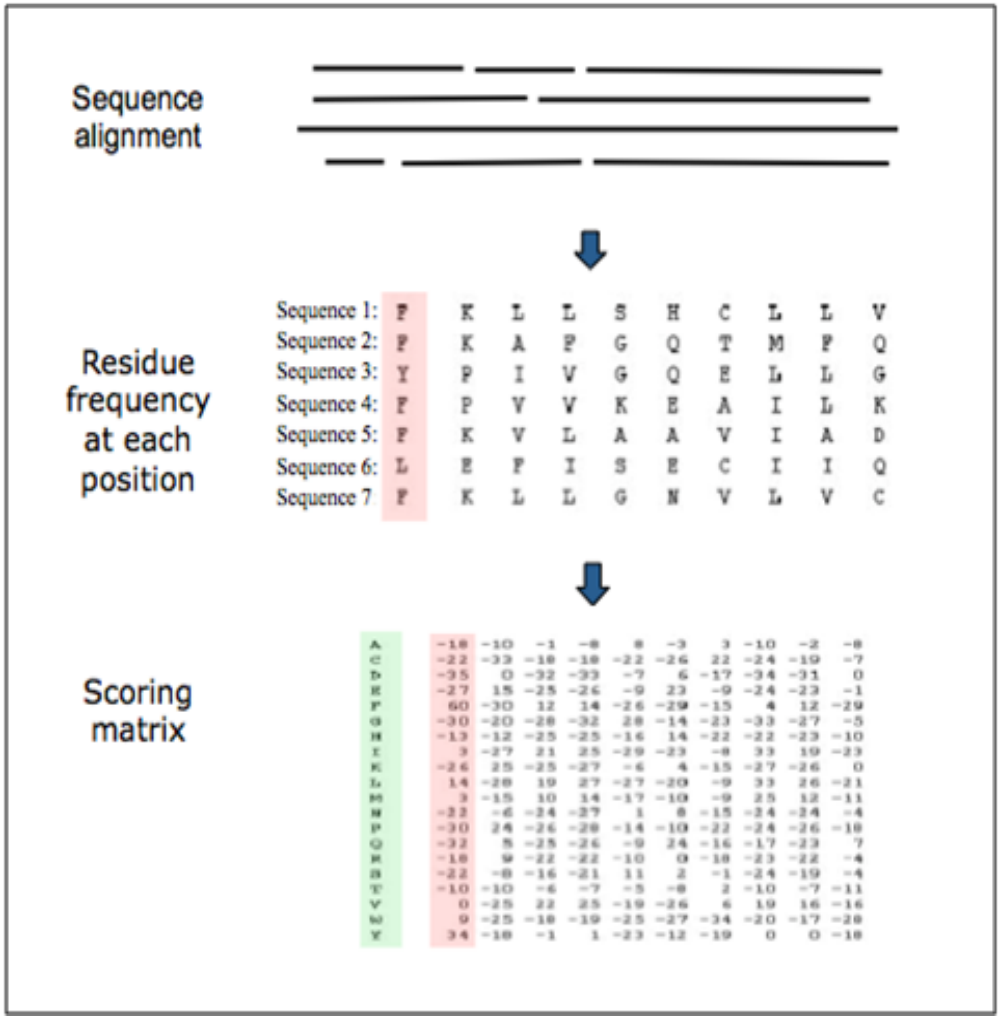
Classification is based on protein signatures

- **Patterns** - important sequence features, such as binding sites or the active sites of enzymes, consist of only a few amino acids that are essential for protein function (e.g. PROSITE)



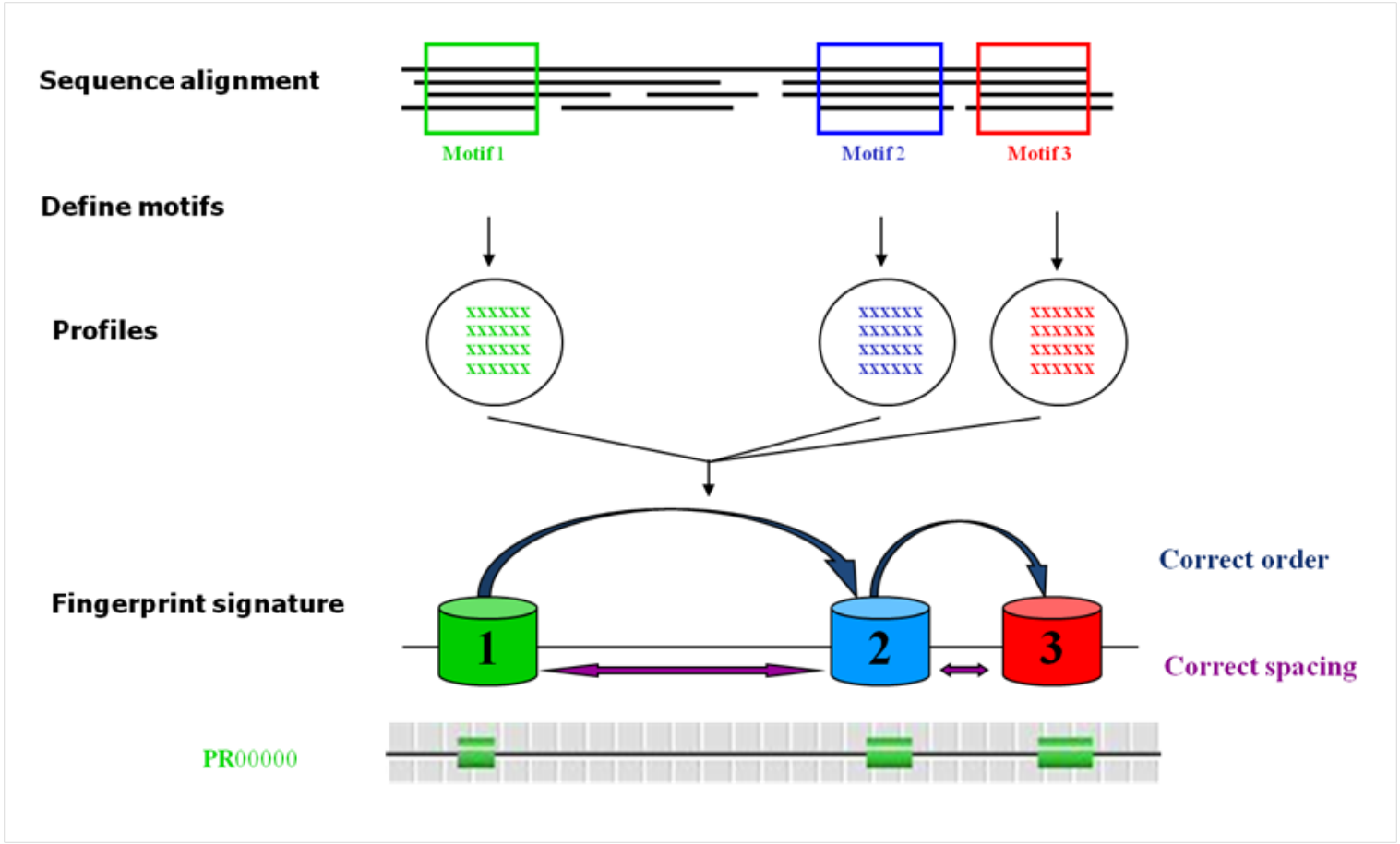
Classification is based on protein signatures

- **Profiles**
  - built by converting multiple sequence alignments into position-specific scoring systems (PSSMs)
  - amino acids at each position in the alignment are scored according to the frequency with which they occur
  - substitution matrices (e.g. BLOSUM) can be used to add evolutionary distance weighting these scores.
  - E.g. HAMAP, PROSITE



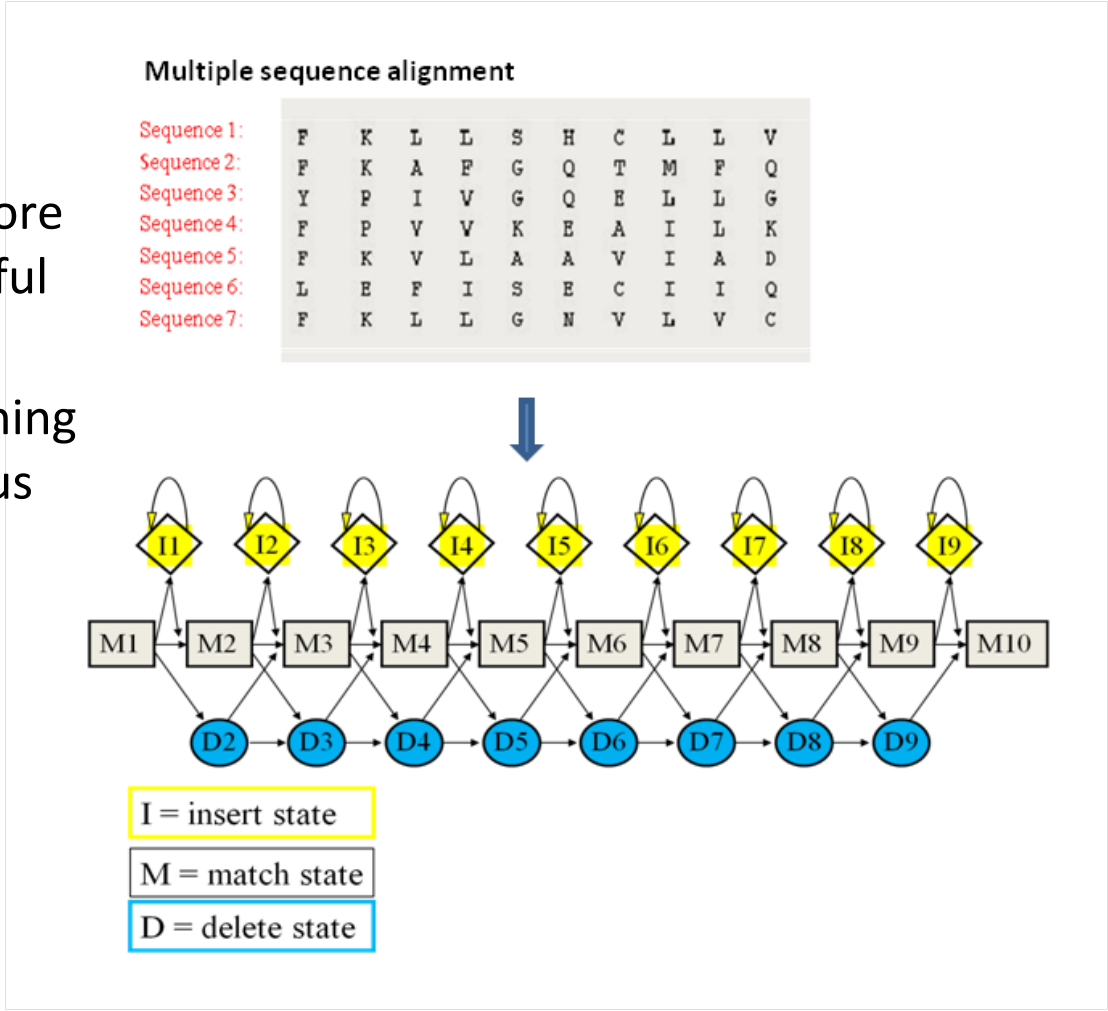
Classification is based on protein signatures

- **Fingerprints** – consist of multiple short conserved motifs, which are drawn from sequence alignments (e.g. PRINTS)



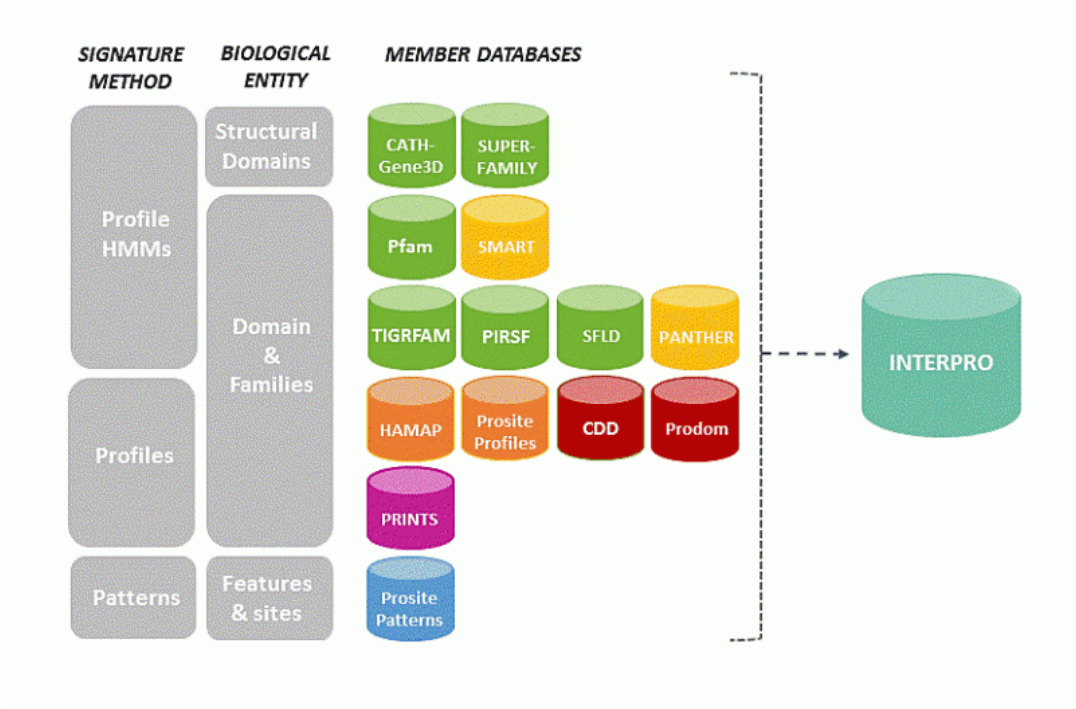
Classification is based on protein signatures

- **Hidden Markov Models** (HMMs)
  - similar to profiles, but more sophisticated and powerful statistical models
  - very well suited to searching databases for homologous sequences
  - e.g. Pfam, PANTHER



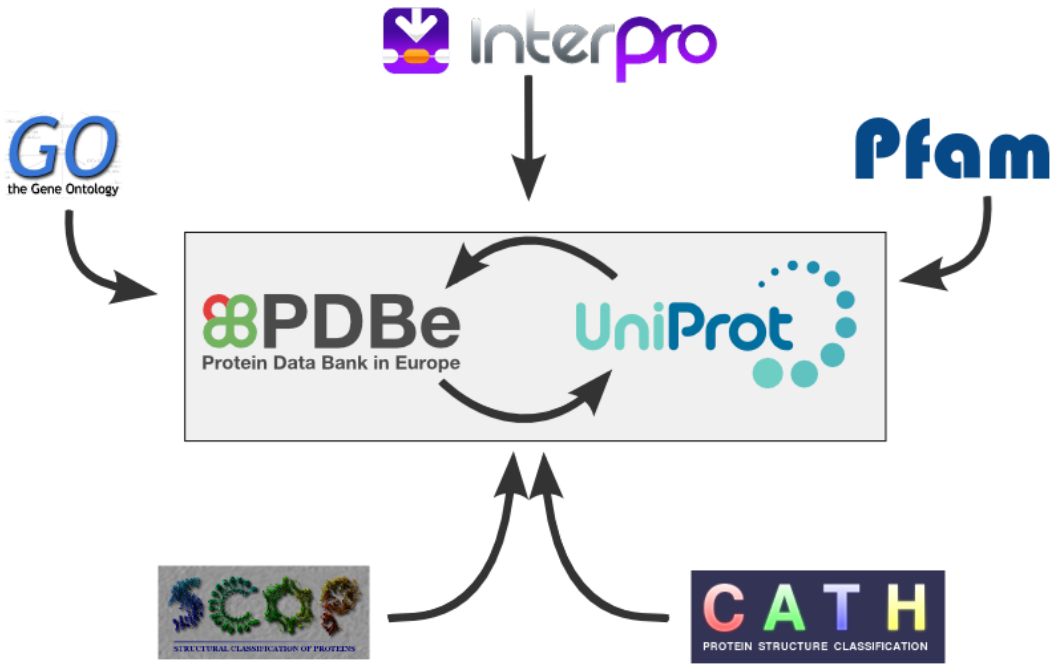
Integrating resources

- InterPro - https://www.ebi.ac.uk/interpro/
  - a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites
  - uses predictive models, known as signatures, provided by several different databases
  - list of consortium members: https://www.ebi.ac.uk/interpro/about.html

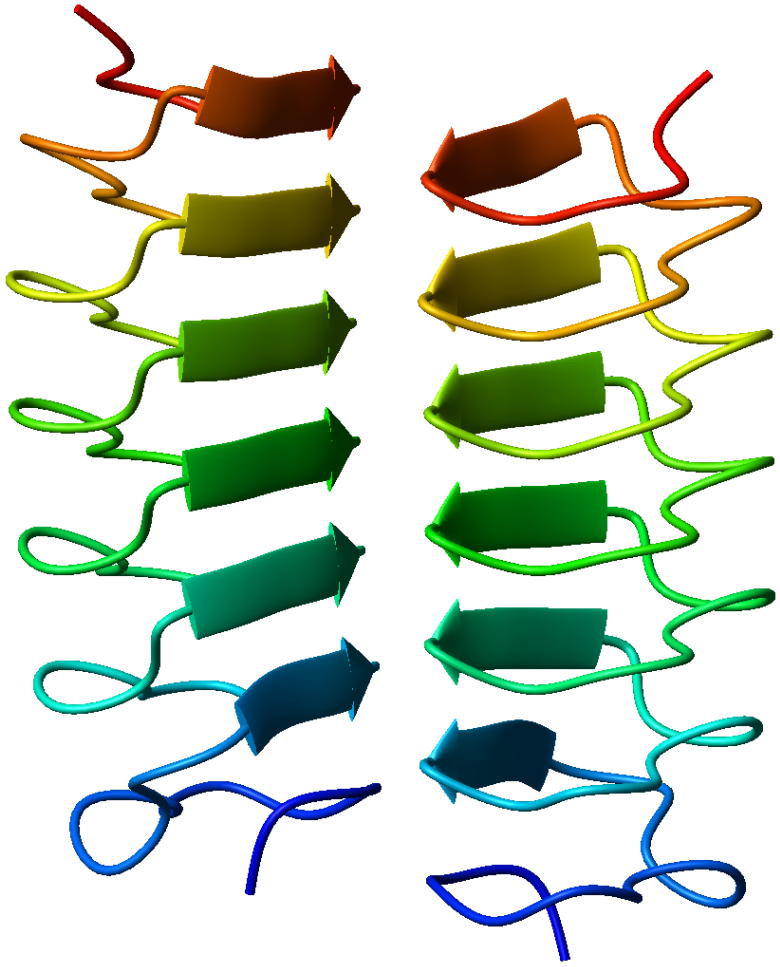


Integrating resources

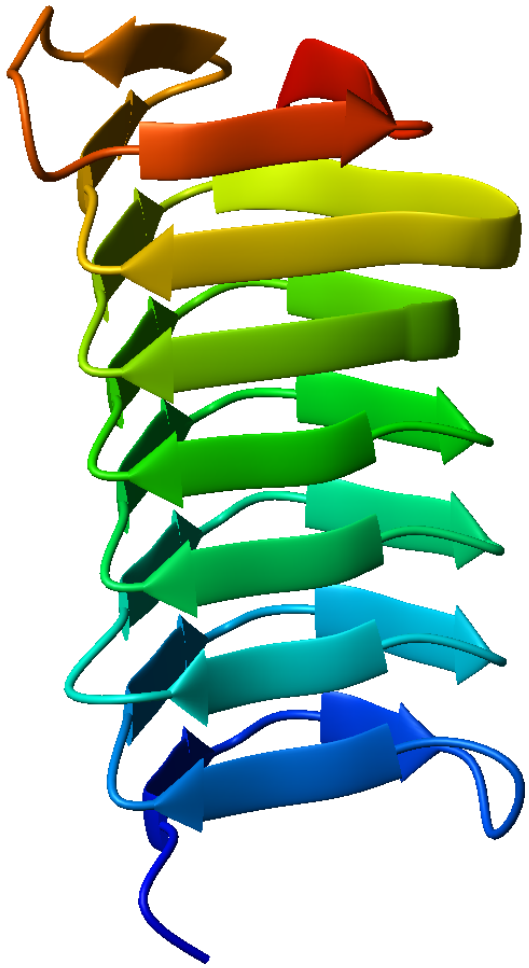
- Structure Integration with Function, Taxonomy and Sequence (SIFTS) - https://www.ebi.ac.uk/pdbe/docs/sifts/index.html



Intermission! BETA helices

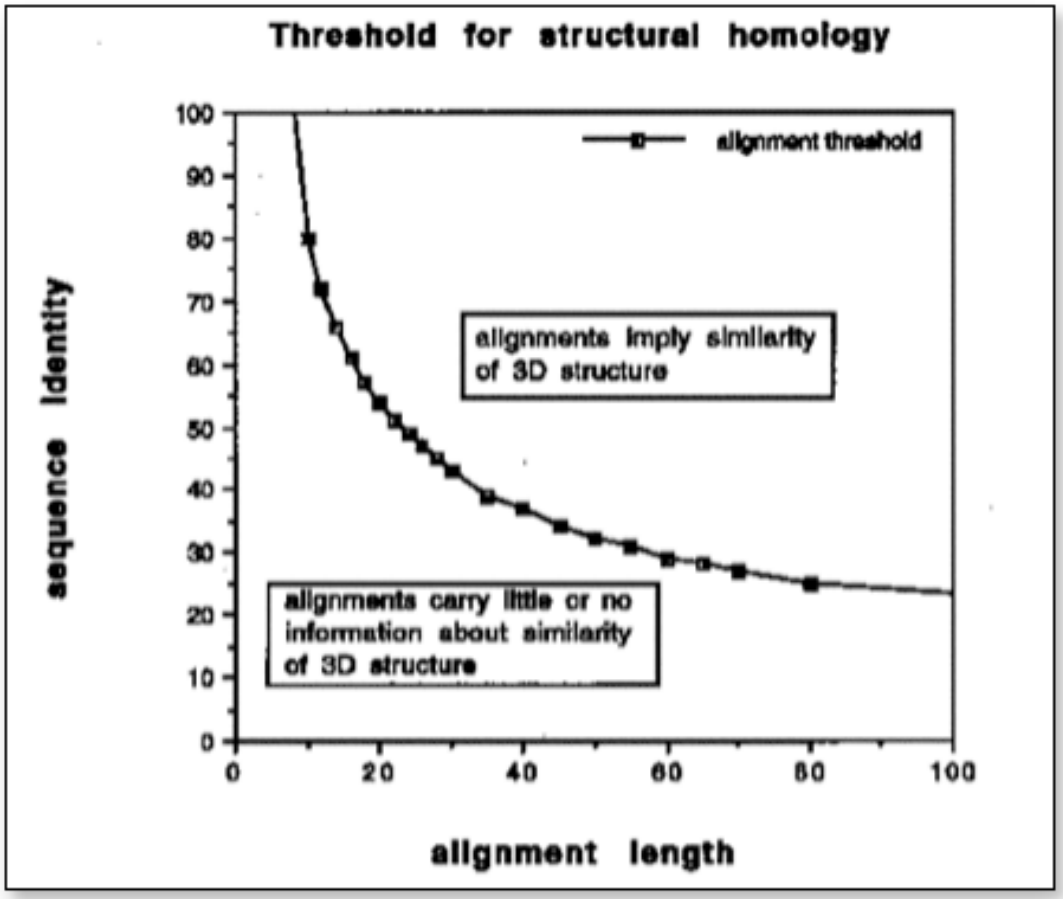


1EZG



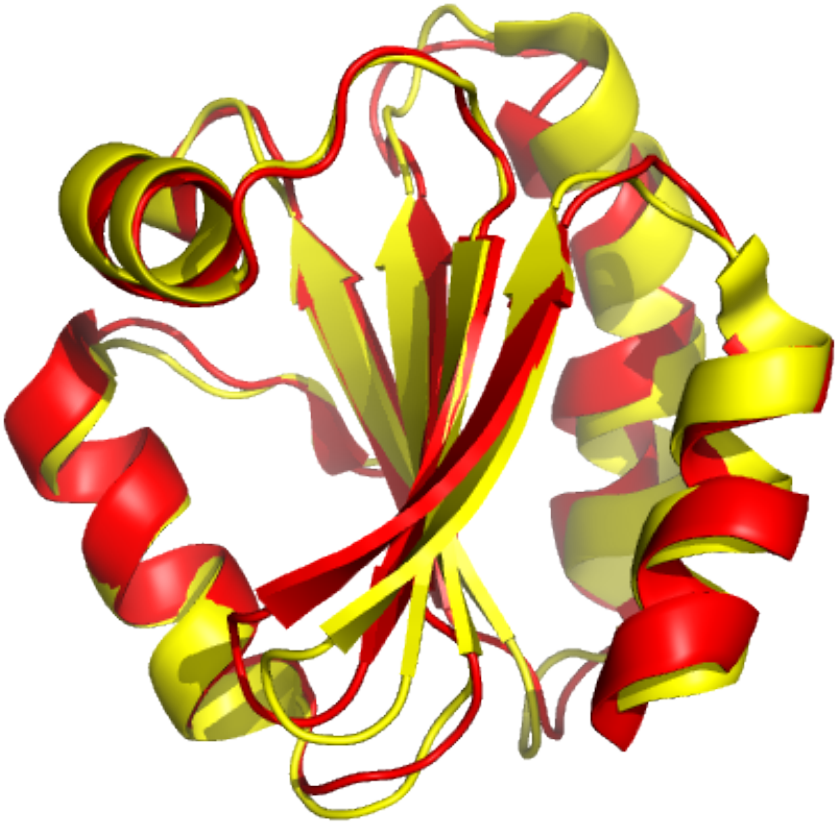
1M8N

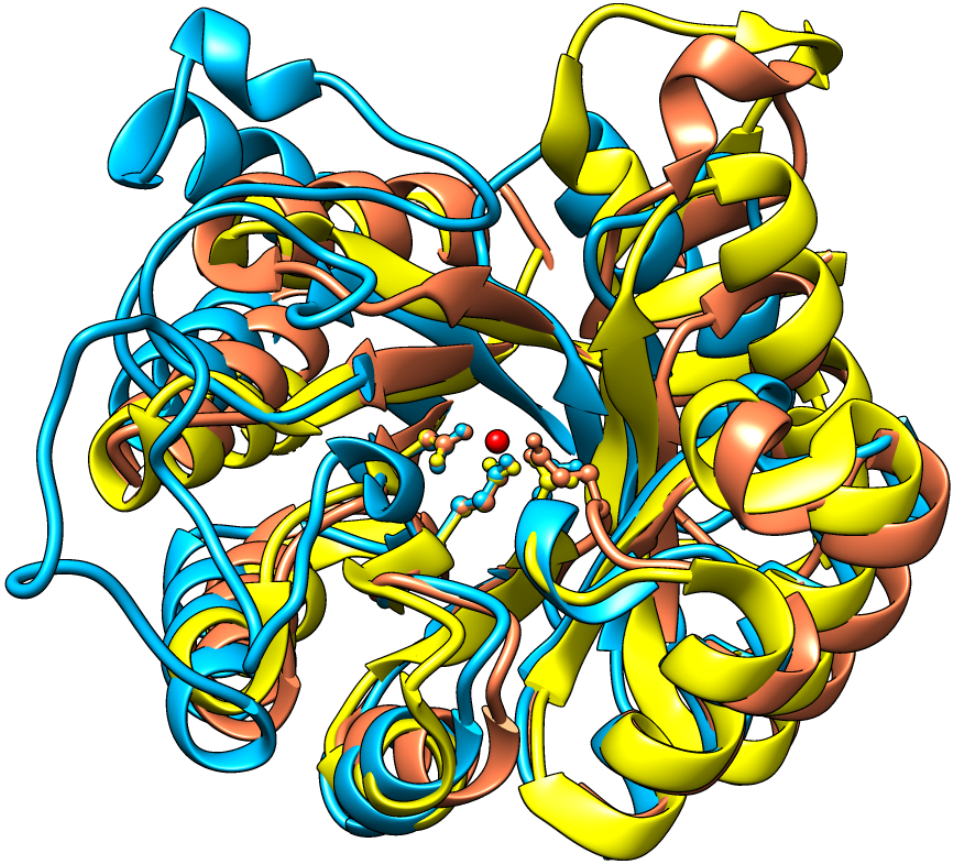
Structure is more conserved than sequence



Structural alignment

- a way to establish homology between proteins based on their structure
- 3D analogue of sequence alignment
- valuable tool for the comparison of proteins with low sequence similarity, where evolutionary relationships between proteins cannot be easily detected by standard sequence alignment techniques
- tools: MUSTANG, MATT, FUGUE, MAMMOTH
- databases: DALI, HOMSTRAD
- output: superposition of the atomic coordinate sets and a minimal root mean square deviation (RMSD) between the structures

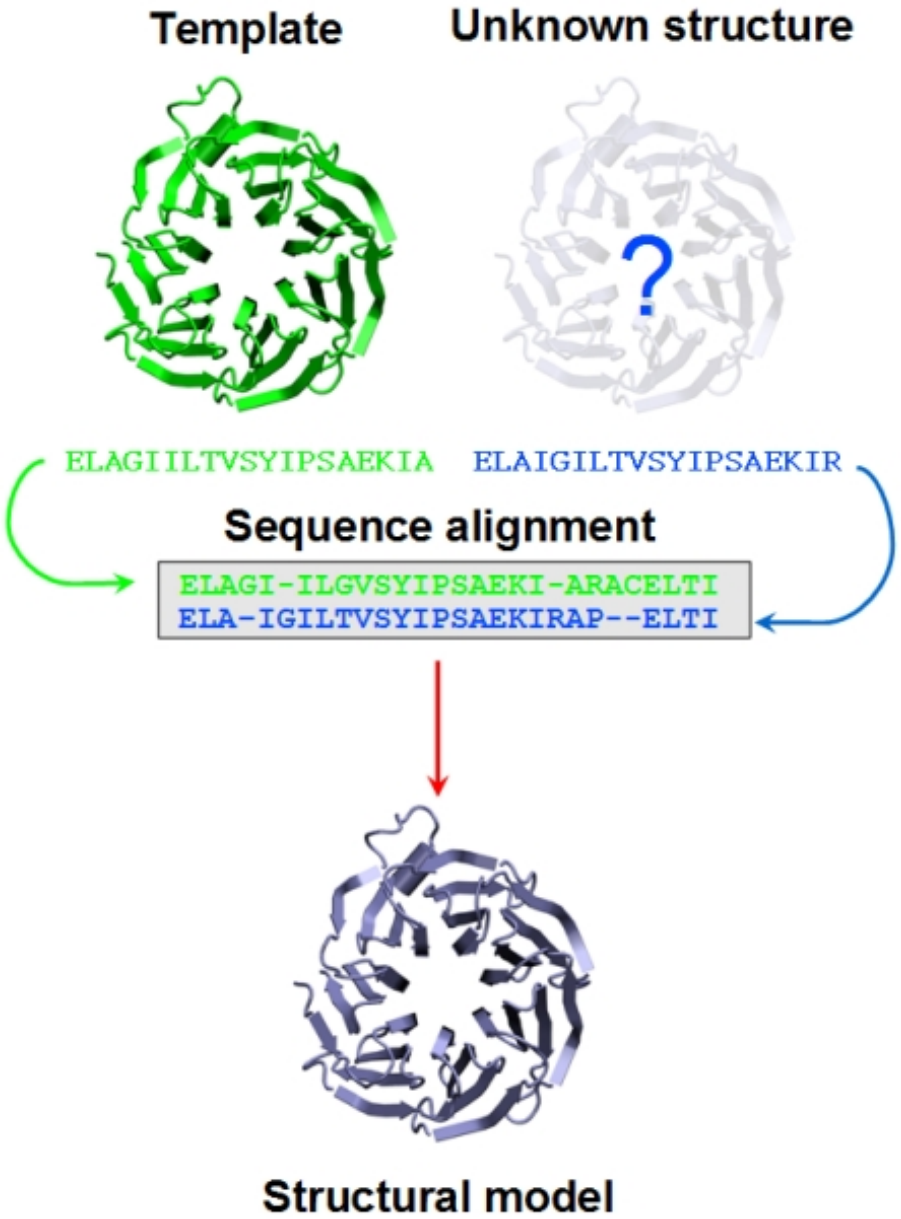




What if I don’t have a structure?

- Comparison to known proteins/domains based on sequence similarity and signatures
- Homology modelling

Homology modelling

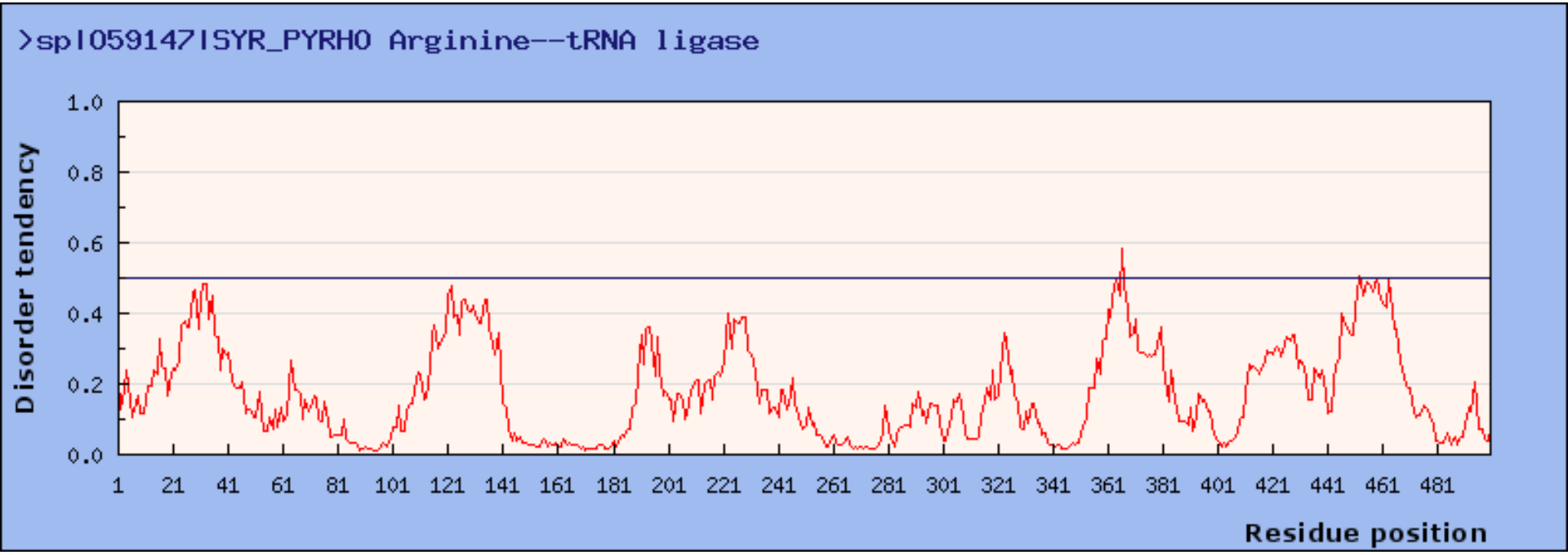


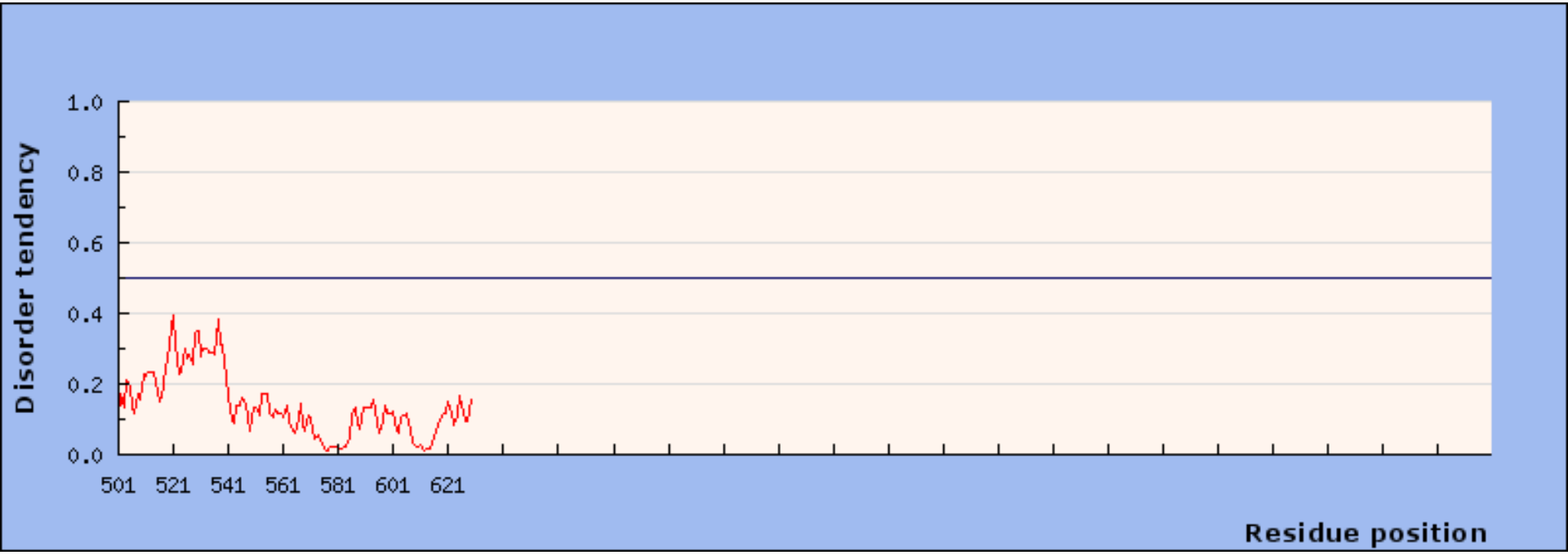
Homology modelling

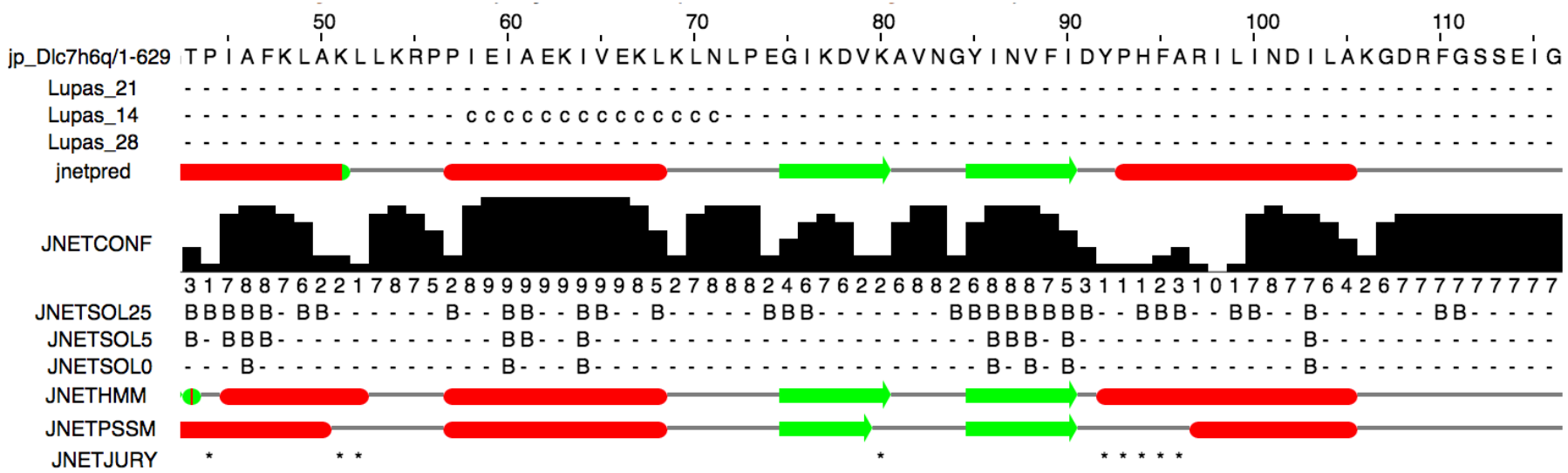
- Select a template structure – needs to share at least 30% sequence similarity with your protein, preferably 50%
- Align the sequences
- Generate geometric criteria and “fit” the new sequence into the structure
- Assess the model according to statistical potentials or physics-based energy calculations
- Accuracy highly dependent on sequence identity between target and template
- SWISS-MODEL - https://swissmodel.expasy.org/

What if I don’t have a structure?

- Predictions based on sequence
- Secondary structure
  - PSIPRED - http://bioinf.cs.ucl.ac.uk/psipred/
  - JPred - http://www.compbio.dundee.ac.uk/jpred/
- Disorder prediction
  - DISOPRED3 - http://bioinf.cs.ucl.ac.uk/psipred/
  - IUPRED - http://iupred.enzim.hu/


>sp|O59147|SYR_PYRHO Arginine--tRNA ligase
Disorder tendency
1.0
0.8
0.6
0.4
0.2
0.0
1 21 41 61 81 101 121 141 161 181 201 221 241 261 281 301 321 341 361 381 401 421 441 461 481
Residue position


Disorder tendency
1.0
0.8
0.6
0.4
0.2
0.0
501 521 541 561 581 601 621
Residue position


50
60
70
80
90
100
110
jp_Dlc7h6q/1-629 T P I A F K L A K L L K R P P I E I A E K I V E K L K L N L P E G I K D V K A V N G Y I N V F I D Y P H F A R I L I N D I L A K G D R F G S S E I G
Lupas_21
Lupas_14
Lupas_28
jnetpred
JNETCONF
3 1 7 8 8 7 6 2 2 1 7 8 7 5 2 8 9 9 9 9 9 9 9 8 5 2 7 8 8 8 2 4 6 7 6 2 2 6 8 8 2 6 8 8 8 7 5 3 1 1 1 2 3 1 0 1 7 8 7 7 6 4 2 6 7 7 7 7 7 7 7
JNETSOL25
JNETSOL5
JNETSOL0
JNETHMM
JNETPSSM
JNETJURY

What can I do with a structure?

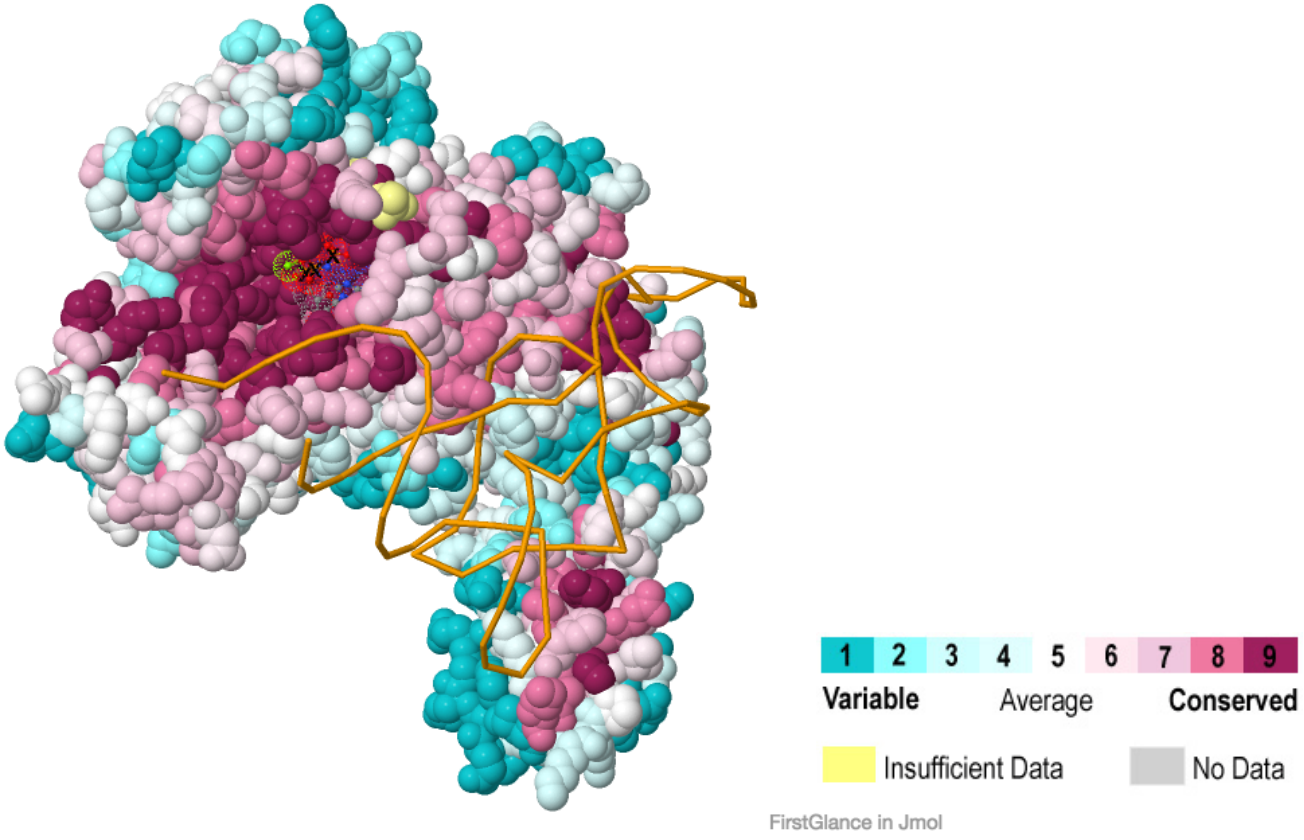
- Evolutionary analysis
- Homology modelling of related proteins
- Analysis and prediction of interactions
- Prediction of function
- Prediction of disease phenotypes
- Drug design
- Protein engineering
- Molecular dynamics simulations

What can I do with a structure? Interactions

- PISA (Proteins, Interfaces, Structures and Assemblies) - http://www.ebi.ac.uk/pdbe/pisa/
  - analyses interfaces occurring in asymmetric units and also predicts probable quaternary structures based on interactions occurring in the macromolecular crystal
- InterProSurf - http://curie.utmb.edu/prosurf.html
  - find interface residues in protein complex structures
- HADDOCK (High Ambiguity Driven protein-protein DOCKing) - http://milou.science.uu.nl/services/HADDOCK2.2/haddock.php
  - information-driven flexible docking approach for the modeling of biomolecular complexes

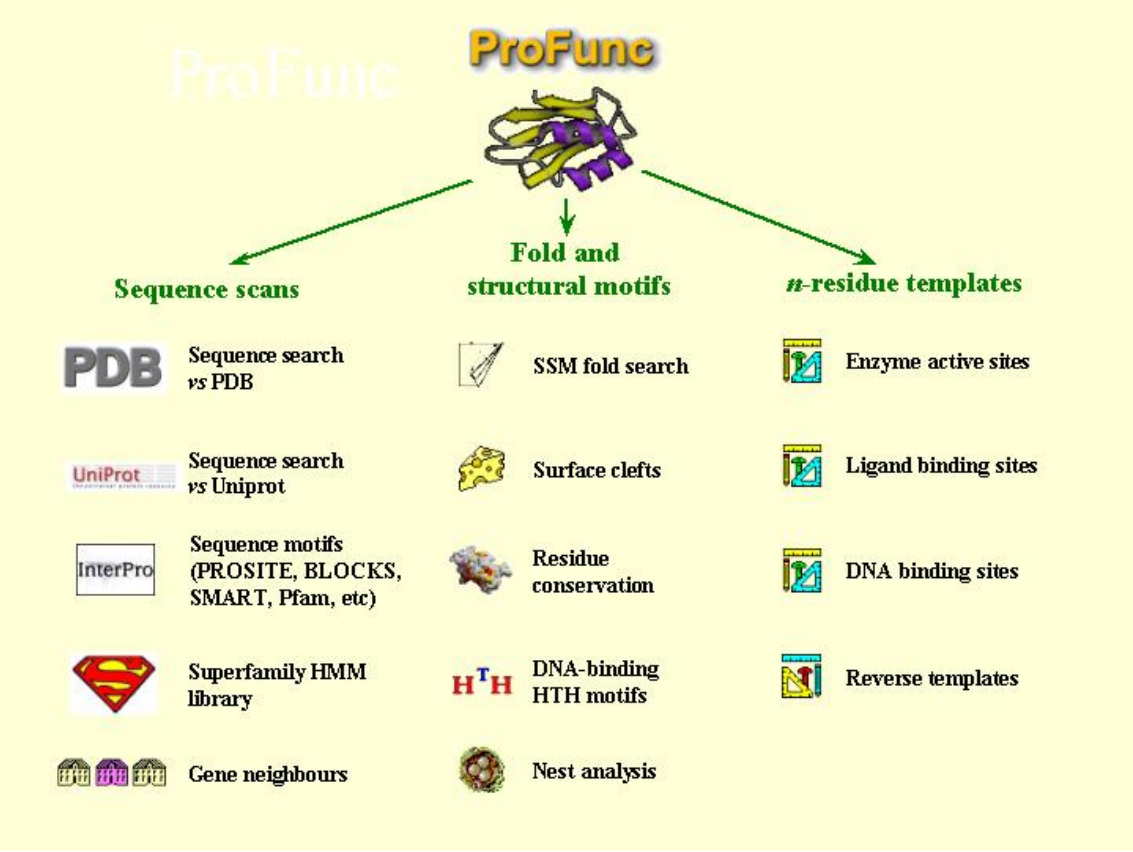
What can I do with a structure? Function

- ConSurf & ConSurf-DB
  - http://consurf.tau.ac.il/ and http://consurfdb.tau.ac.il/
  - Function based on evolutionary conservation - looking for evolutionarily conserved patches of amino acids in a 3D protein structure is a good way to locate functional sites



What can I do with a structure? Function

- ProFunc - https://www.ebi.ac.uk/thornton-srv/databases/profunc/
  - prediction of protein function from 3D structure
  - uses a series of methods, including fold matching, residue conservation, surface cleft analysis, and functional 3D templates, to identify both the protein's likely active site and possible homologues in the PDB

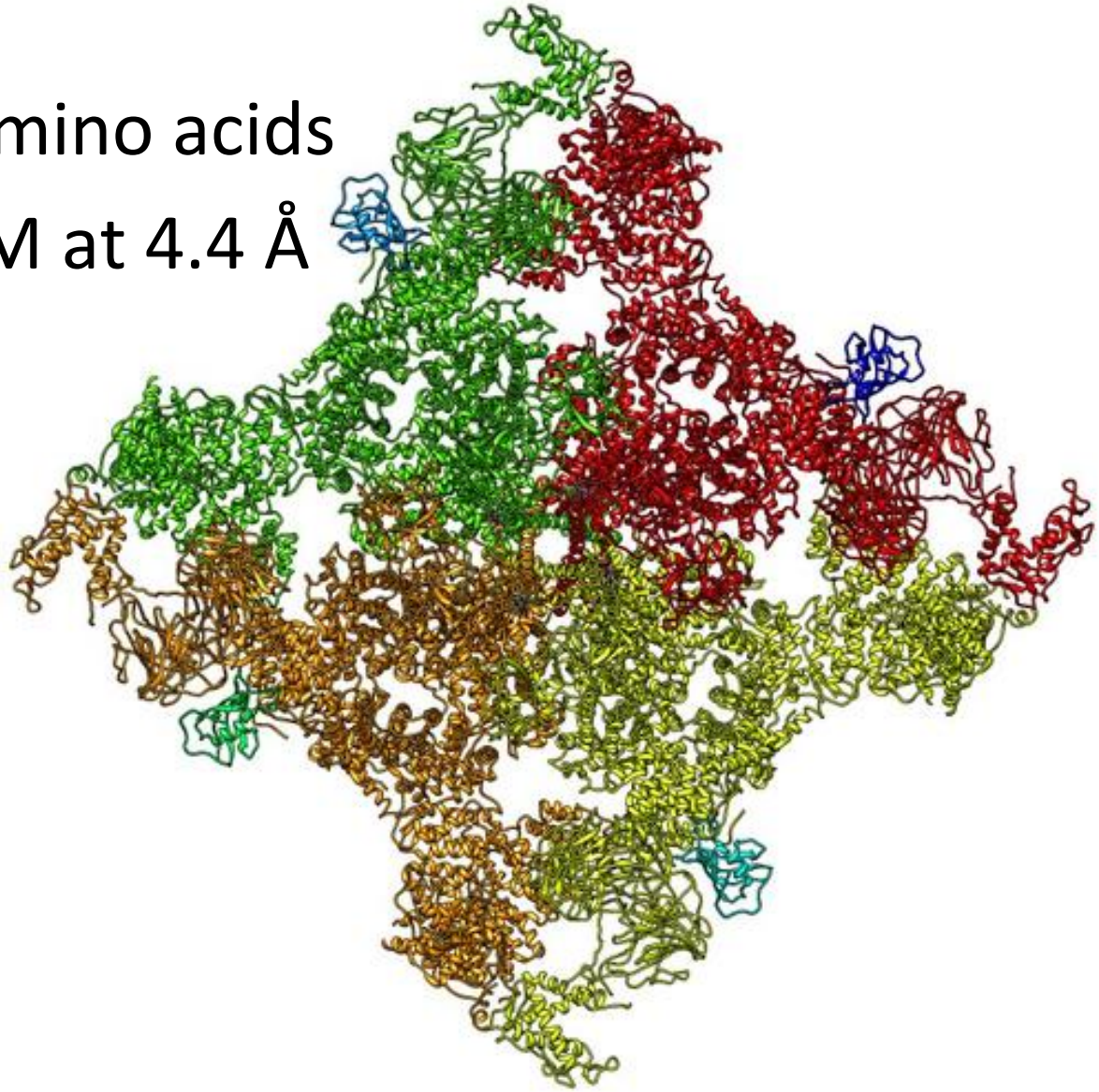


What can I do with a structure? Disease phenotypes

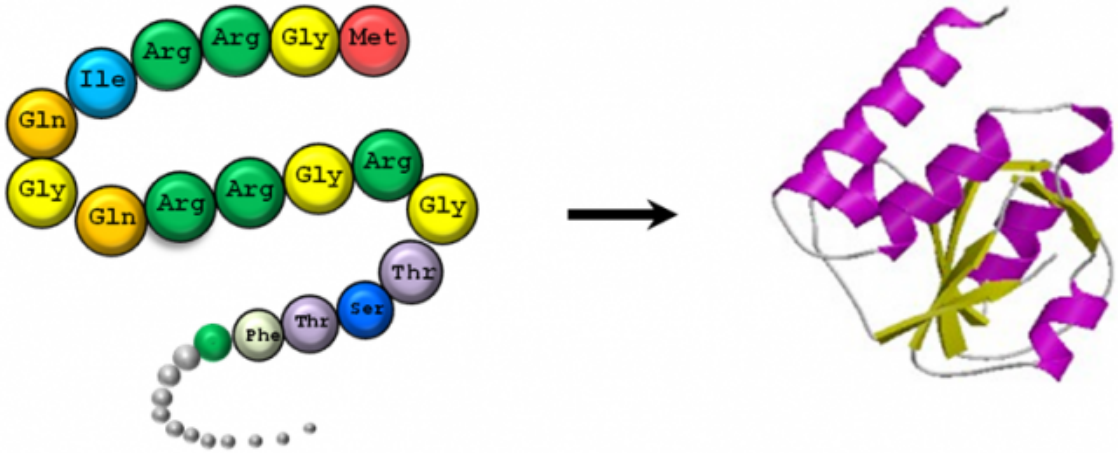
- SAAPdap - http://www.bioinf.org.uk/saap/dap/
  - likely structural effects of a mutation
- Ensembl Variant Effect Predictor (VEP) - https://www.ensembl.org/info/docs/tools/vep/index.html

Intermission! Largest structure

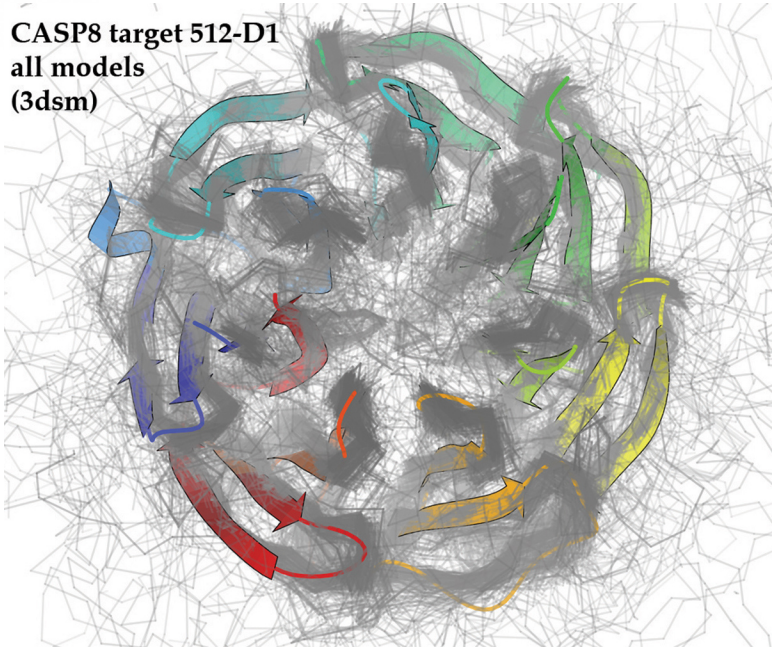
- 5TA3 - 39,852 amino acids
- RYR1 by Cryo-EM at 4.4 Å



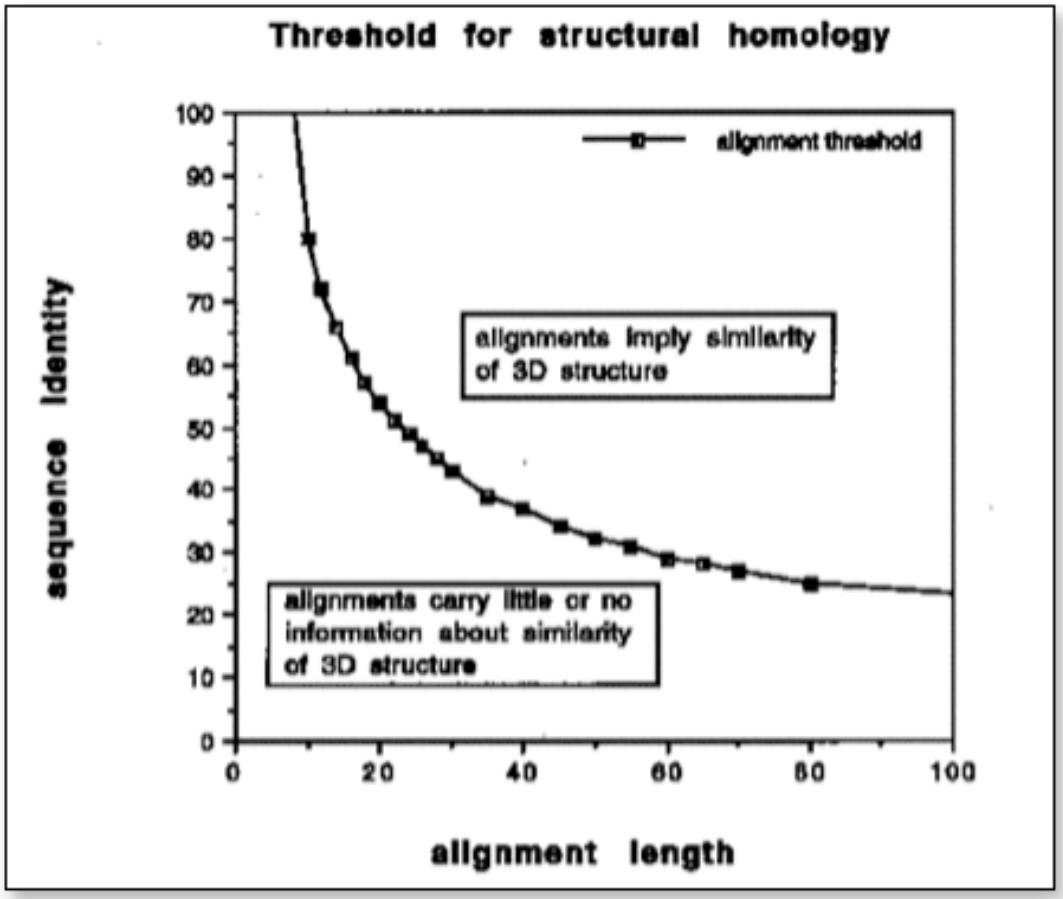
The Holy Grail of structural biology – structure prediction from sequence

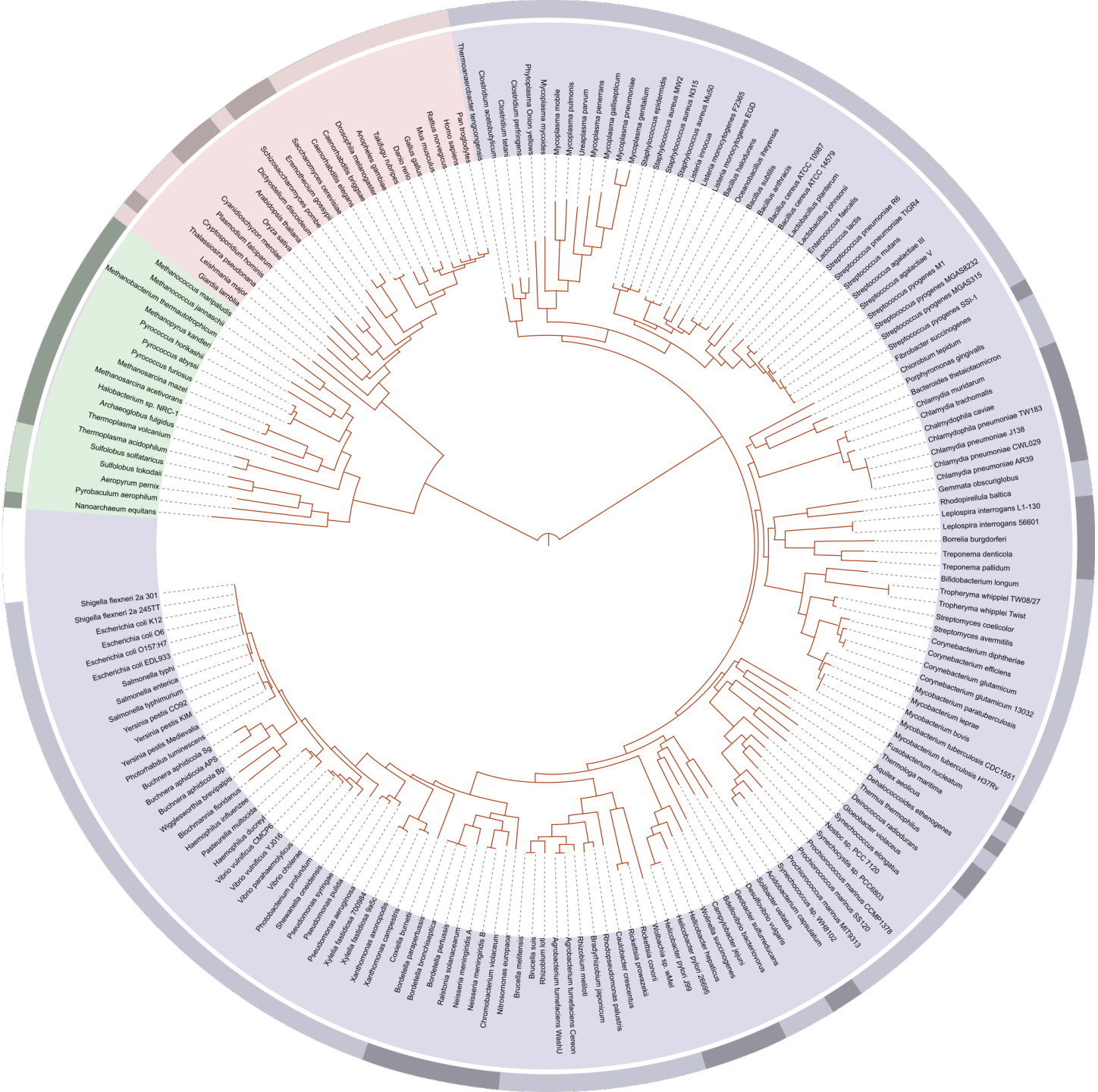


CASP – Critical Assessment of techniques for protein Structure Prediction

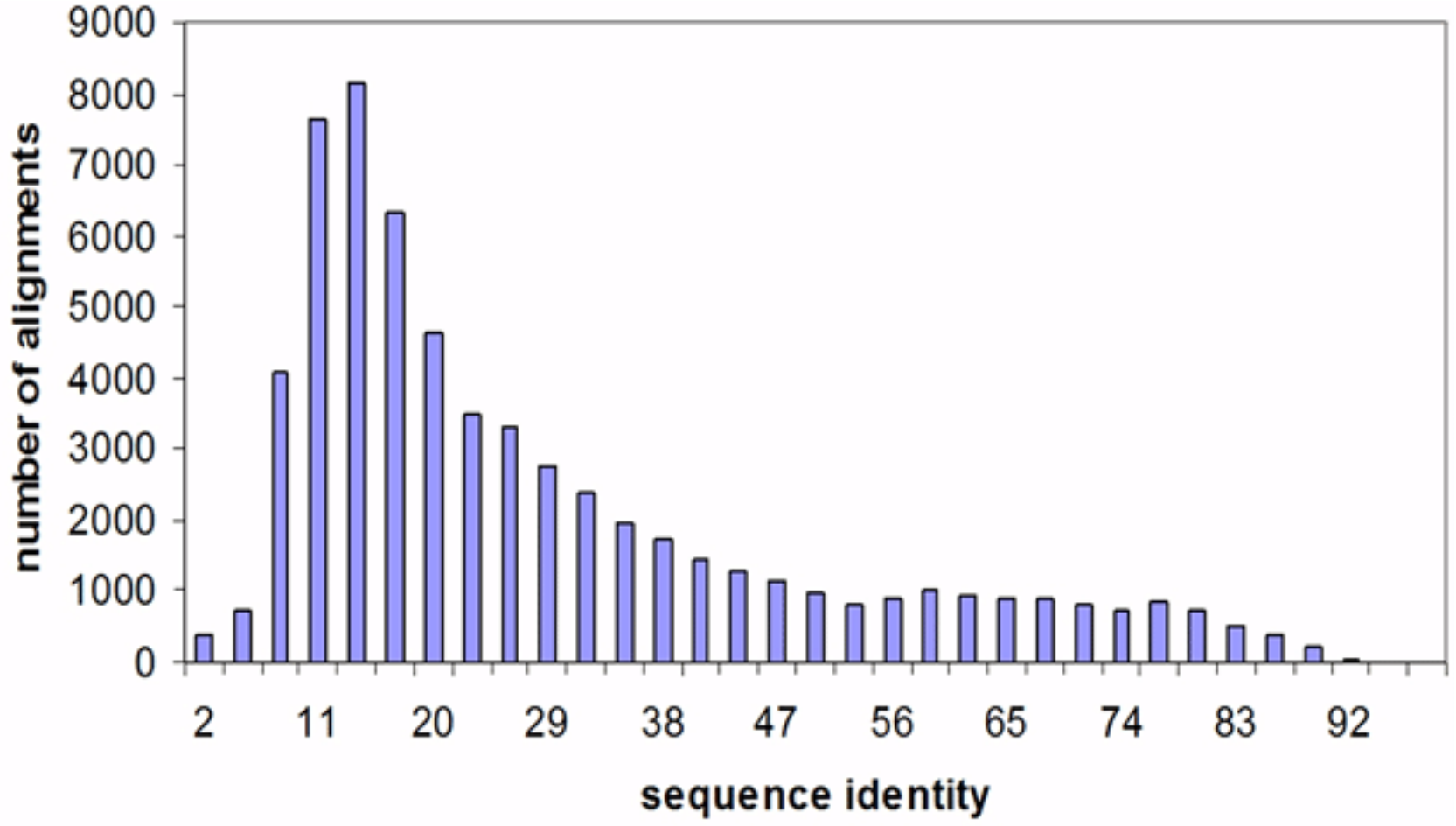


Structure is more conserved than sequence




Thermoanaerobacter tengcongensis
Clostridium acetobutylicum
Clostridium tetani
Clostridium perfringens
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma mobile
Mycoplasma pulmonis
Ureaplasma urealyticum
Mycoplasma penetrans
Mycoplasma gallisepticum
Mycoplasma pneumoniae
Staphylococcus genitalium
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus N315
Staphylococcus aureus Mu50
Listeria innocua
Listeria monocytogenes F2365
Listeria monocytogenes EGD
Bacillus halodurans
Oceanobacillus iheyensis
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Lactobacillus plantarum
Lactobacillus johnsonii
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus mutans
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus pyogenes M1
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SSI-1
Streptococcus sucnogenes
Fibrobacter sucnogenes
Chlorobium tepidum
Porphyromonas gingivalis
Bacteroides thetaiotaomicron
Chlamydia muridarum
Chlamydia trachomatis
Chlamydophila caviae
Chlamydophila pneumoniae TW183
Chlamydia pneumoniae J138
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae AR39
Gemmata obscuriglobus
Rhodopirellula baltica
Leptospira interrogans L1-130
Leptospira interrogans 56601
Borrelia burgdorferi
Treponema denticola
Treponema pallidum
Bifidobacterium longum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Streptomyces coelicolor
Streptomyces avermitilis
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum
Corynebacterium glutamicum 13032
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv
Fusobacterium nucleatum
Thermologa maritima
Aquifex aeolicus
Dehalococcoides ethenogenes
Thermus thermophilus
Deinococcus radiodurans
Gloeobacter violaceus
Synechococcus elongatus
Nostoc sp. PCC 7120
Synechocystis sp. PCC6803
Prochlorococcus marinus CCMP1378
Prochlorococcus marinus MIT9313
Synechococcus sp. WH8102
Prochlorococcus marinus SS120
Acidobacterium capsulatum
Solibacter usitatus
Desulfovibrio vulgaris
Geobacter sulfurreducens
Ralstonia solanacearum
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori 26695
Helicobacter pylori J99
Wolbachia sp. wMel
Rickettsia prowazekii
Rickettsia conorii
Caulobacter crescentus
Rhodopseudomonas palustris
Bradyrhizobium japonicum
Rhizobium meliloti
Agrobacterium tumefaciens Cereon
Agrobacterium tumefaciens WashU
Rhizobium loti
sins sipedon
Brucella melitensis
Brucella suis
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis A
Neisseria meningitidis B
Ralstonia solanacearum
Bordetella pertussis
Bordetella parapertussis
Bordetella bronchiseptica
Coxiella burnetii
Xanthomonas campestris
Xanthomonas axonopodis
Xylella fastidiosa 9a5c
Xylella fastidiosa 7000964
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus YJ016
Vibrio vulnificus CMCP6
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Blochmannia floridanus
Wigglesworthia brevipalpis
Buchnera aphidicola Bp
Buchnera aphidicola APS
Buchnera aphidicola Sg
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis KIM
Salmonella typhimurium
Salmonella typhi COS2
Salmonella enterica
Escherichia coli EDL933
Escherichia coli O157:H7
Escherichia coli O6
Escherichia coli K12
Shigella flexneri 2a 245TT
Shigella flexneri 2a 301
Nanoarchaeum equitans
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus tokodaii
Sulfolobus solfataricus
Thermoplasma acidophilum
Thermoplasma volcanium
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Methanosarcina acetivorans
Methanosarcina mazei
Pyrococcus abyssi
Pyrococcus furiosus
Pyrococcus horikoshii
Methanopyrus kandleri
Methanobacterium thermautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Giardia lamblia
Leishmania major
Trypanosoma brucei
Thalassiosira pseudonana
Cryptosporidium hominis
Plasmodium falciparum
Cyanidioschyzon merolae
Oryza sativa
Arabidopsis thaliana
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Eremothecium gossypii
Caenorhabditis briggsae
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Takifugu rubripes
Danio rerio
Gallus gallus
Mus musculus
Rattus norvegicus
Homo sapiens
Pan troglodytes

Structure is more conserved than sequence

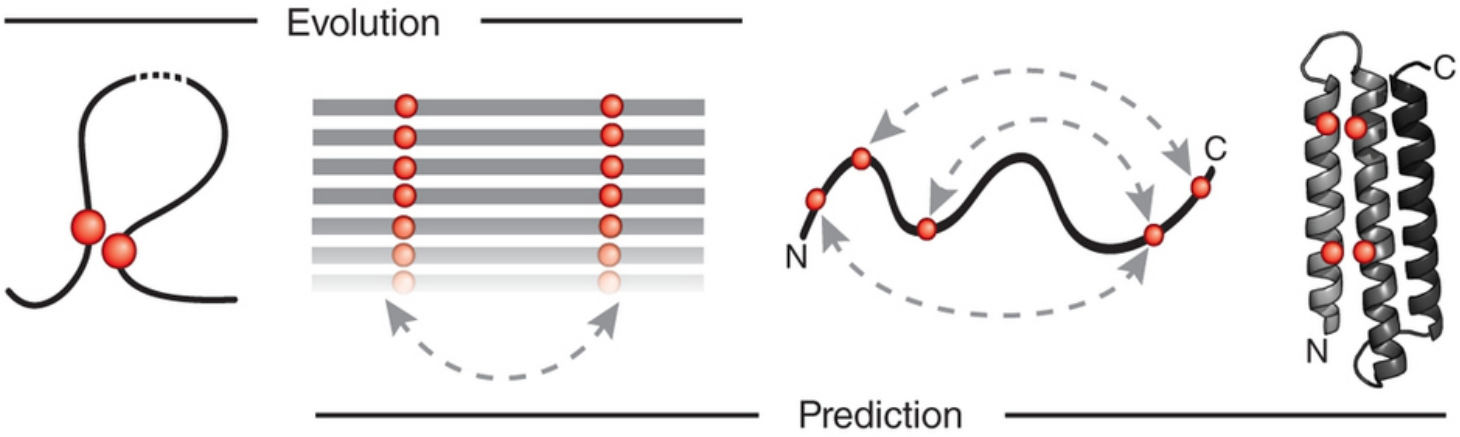


Hypothesis

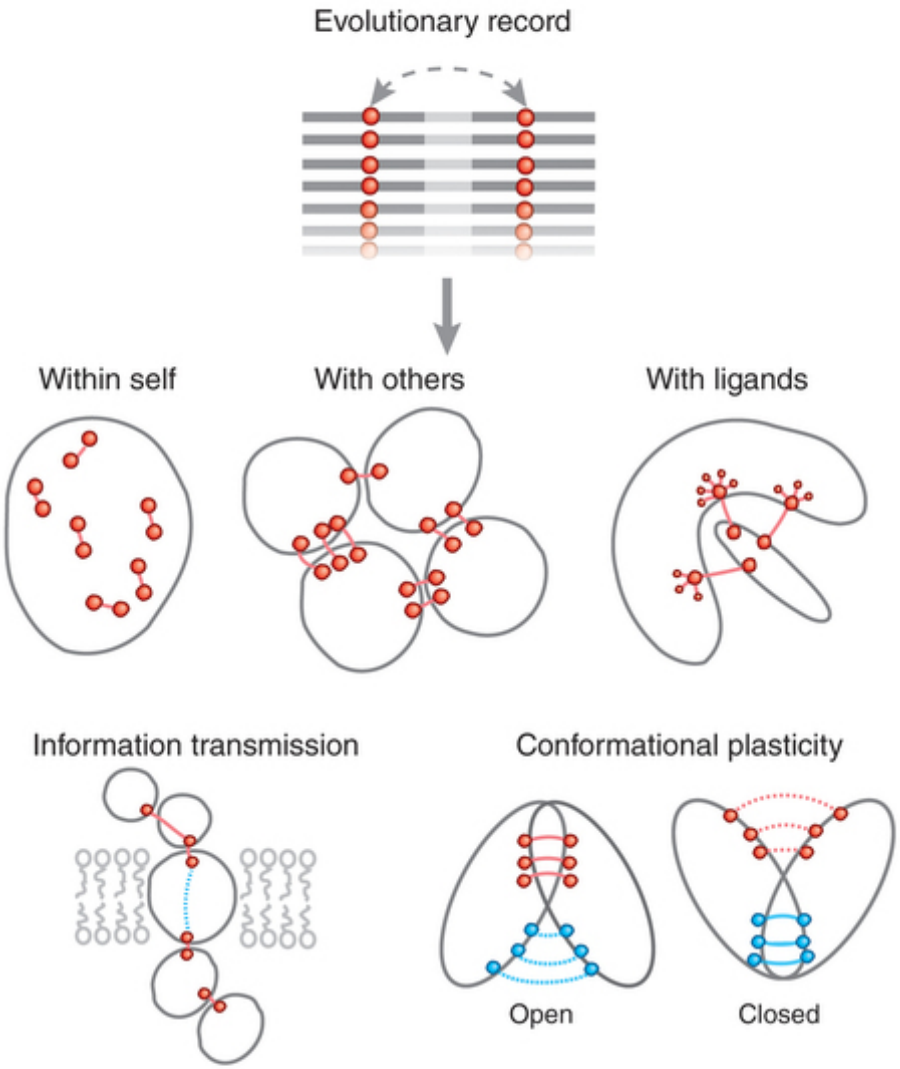
- Conserved structure and function can encompass an "*ensemble*" of related sequences
- Can we *predict* structure and function from a sequence ensemble?
- What *principles* govern conservation beyond sequence?
- What is the *minimum sufficient information* to uniquely describe a protein fold or function?

Coevolution

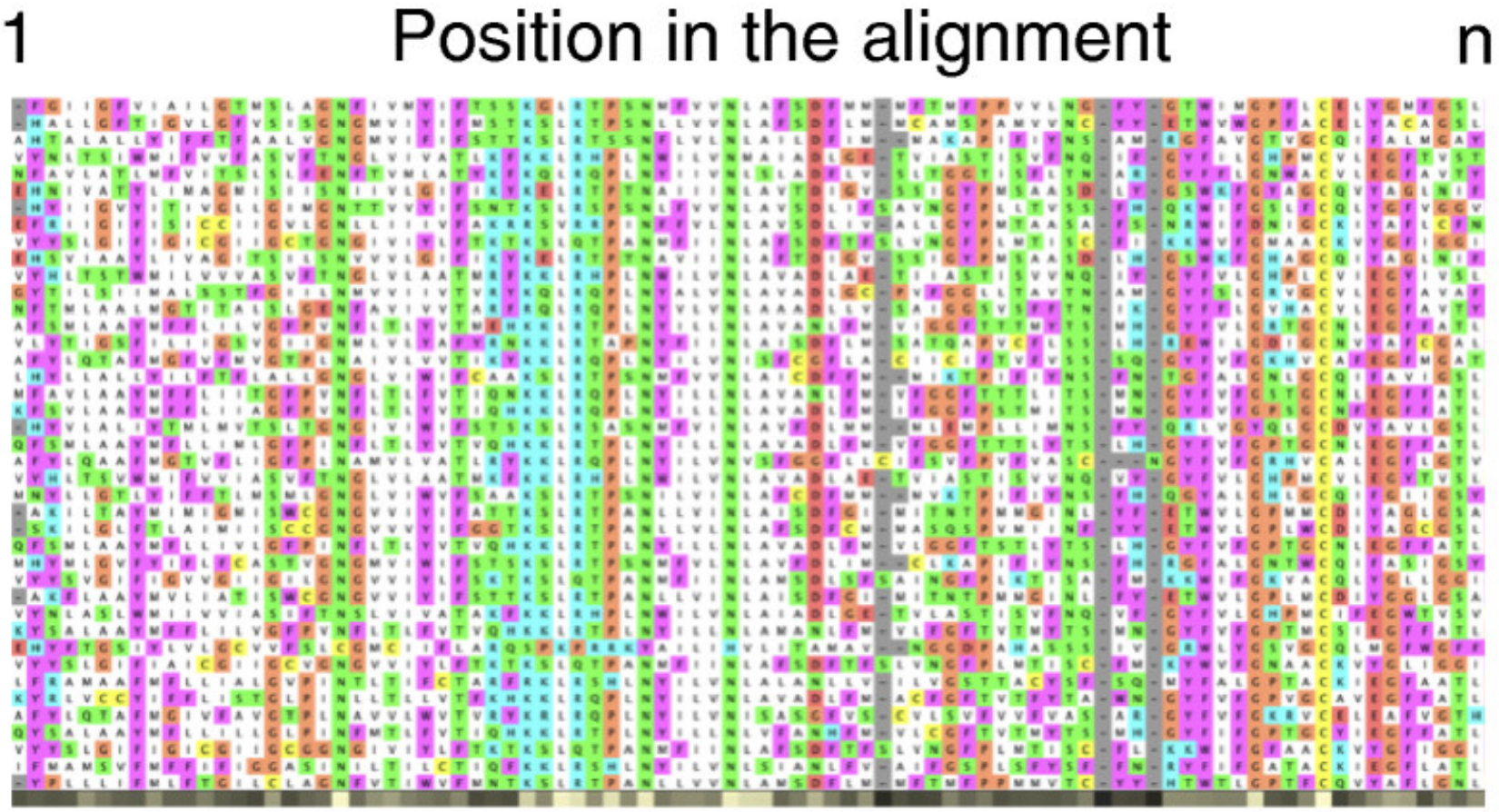
Co-evolutionary signal carries information?



Co-evolutionary signal carries information?

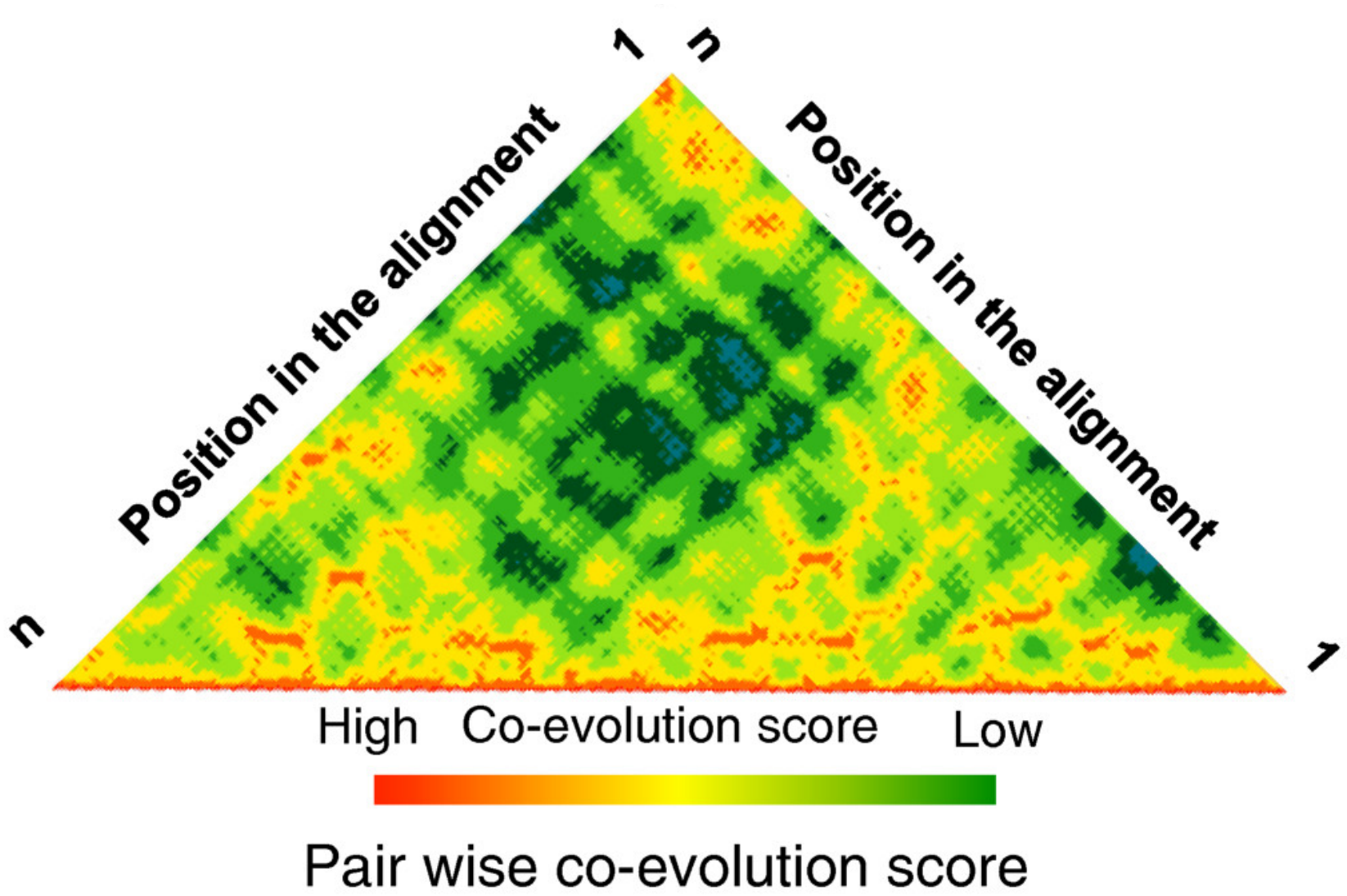


Thanks to cheap DNA sequencing!

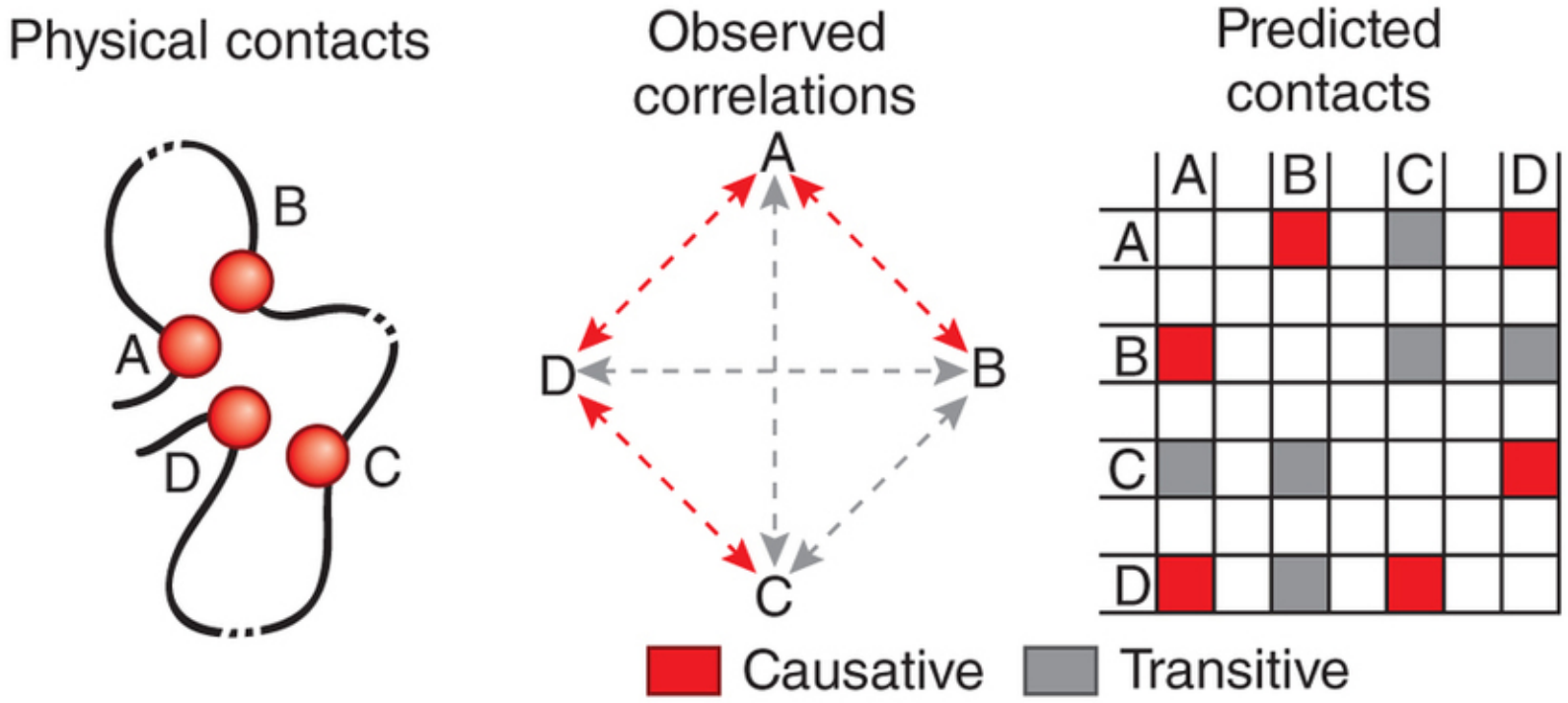


Multiple sequence alignment

How to score co-evolution?



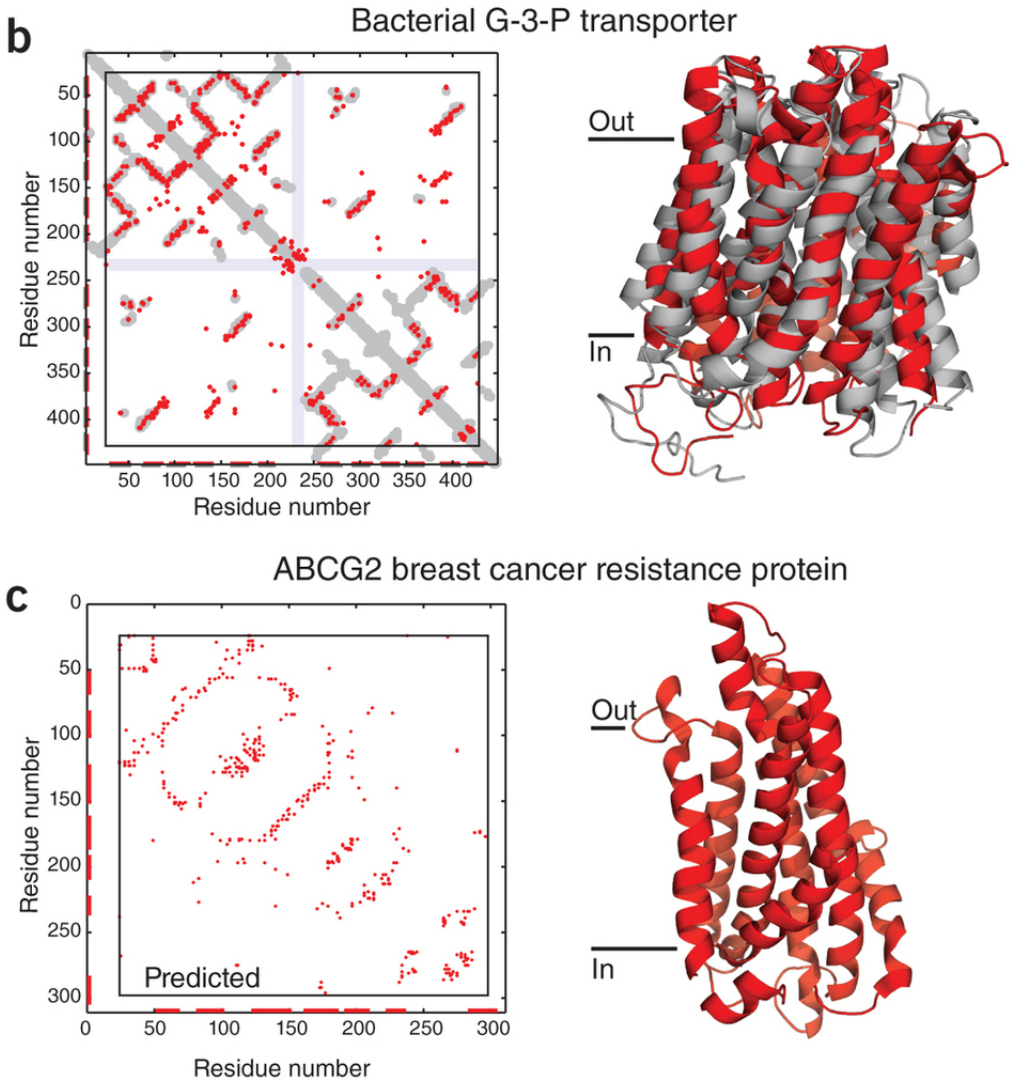
Causative vs transitive correlations (inverse Ising problem)

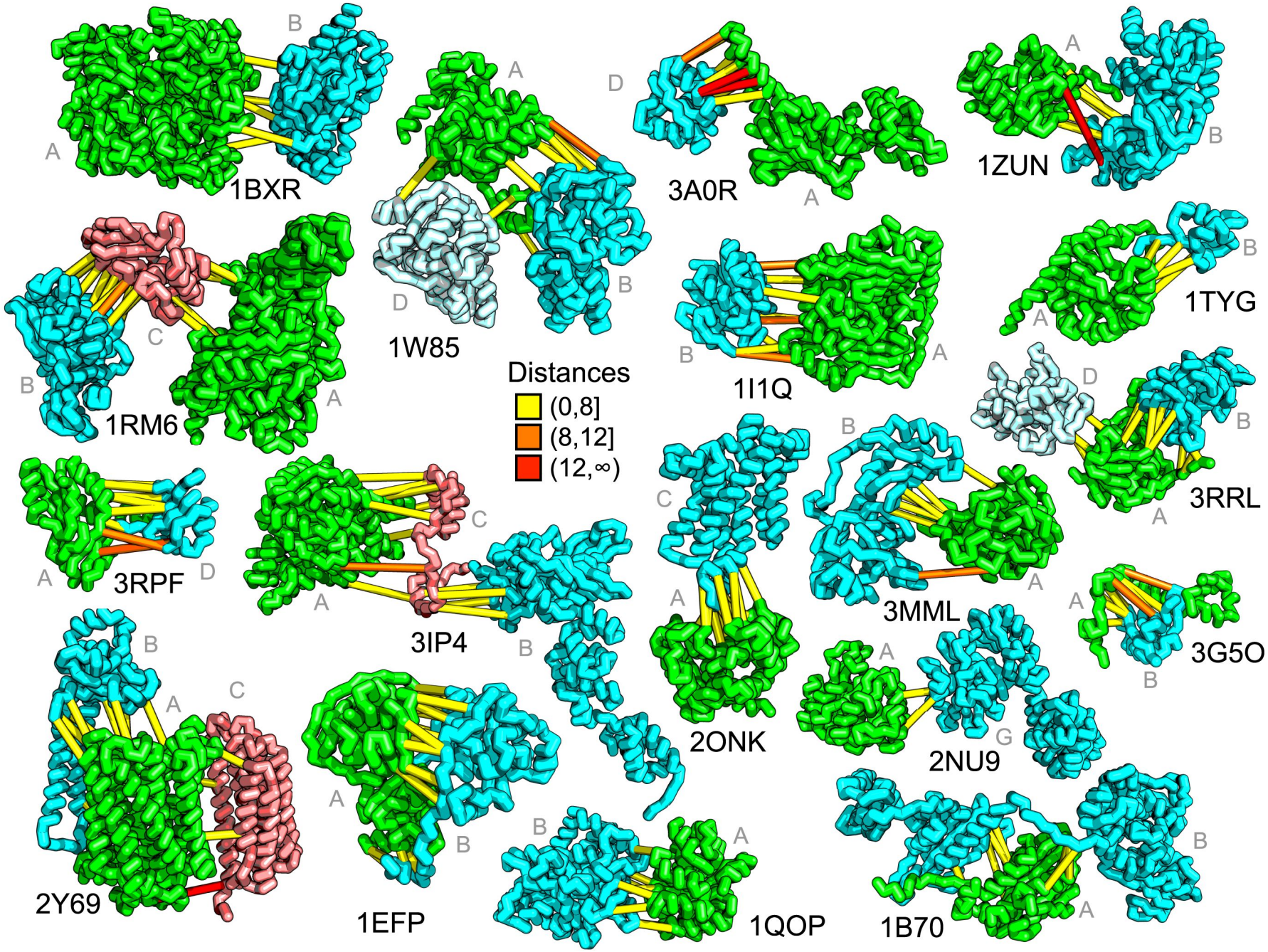


Various methods

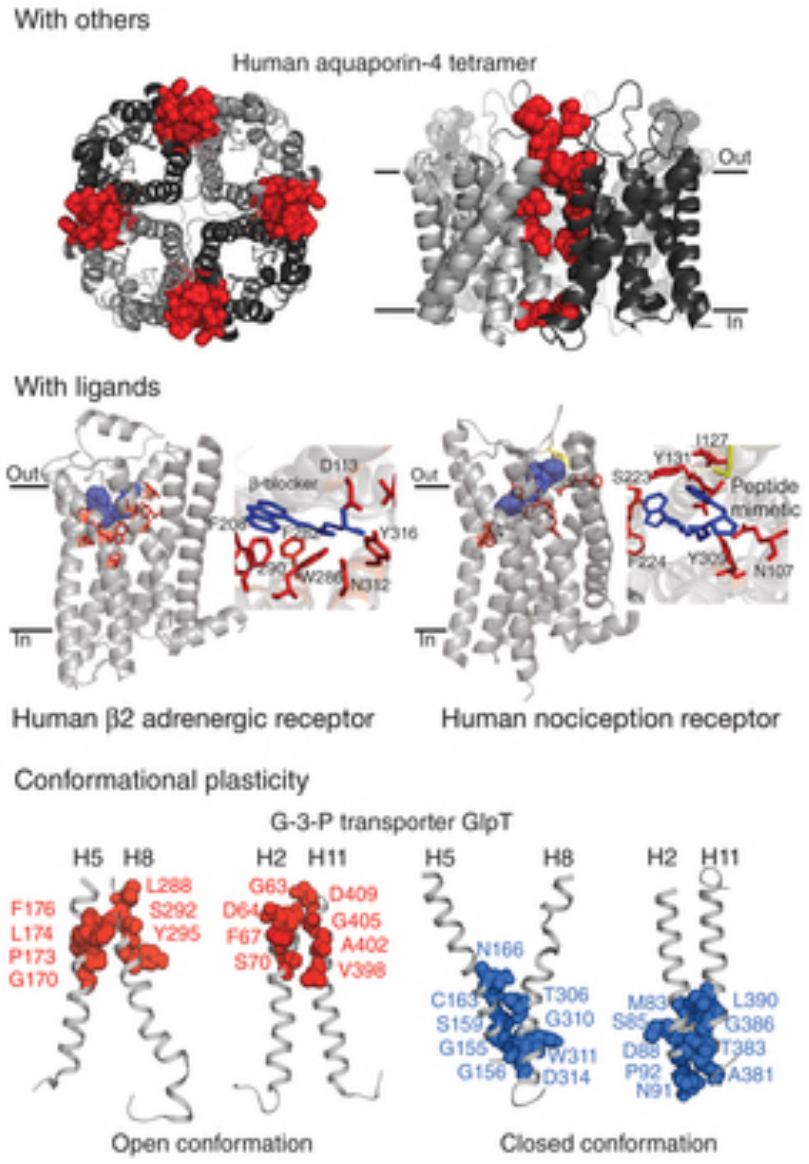
- Correlation coefficients (e.g. Pearson's)
- Difference between observed and expected patterns of data distribution (OMES)
- Mutual information based (various)
- Alignment perturbation (e.g. statistical coupling analysis – SCA)
- New generation of global methods:
  - Protein sparse inverse covariance (PSICOV)
  - Direct coupling analysis / direct information (DCA/DI)
    - Mean-field approximation (mfDCA)
    - Pseudo-likelihood maximisation (plmDCA, GREMLIN)
- Machine Learning based approaches
  - CNNs, ResNets etc.

Success story: membrane protein structure prediction




1BXR
1W85
3A0R
1ZUN
1RM6
1I1Q
1TYG
Distances
(0,8]
(8,12]
(12,∞)
3RRL
3RPF
3IP4
3MML
3G5O
2ONK
2NU9
2Y69
1EFP
1QOP
1B70

Evolutionary couplings reveal function



REPORT

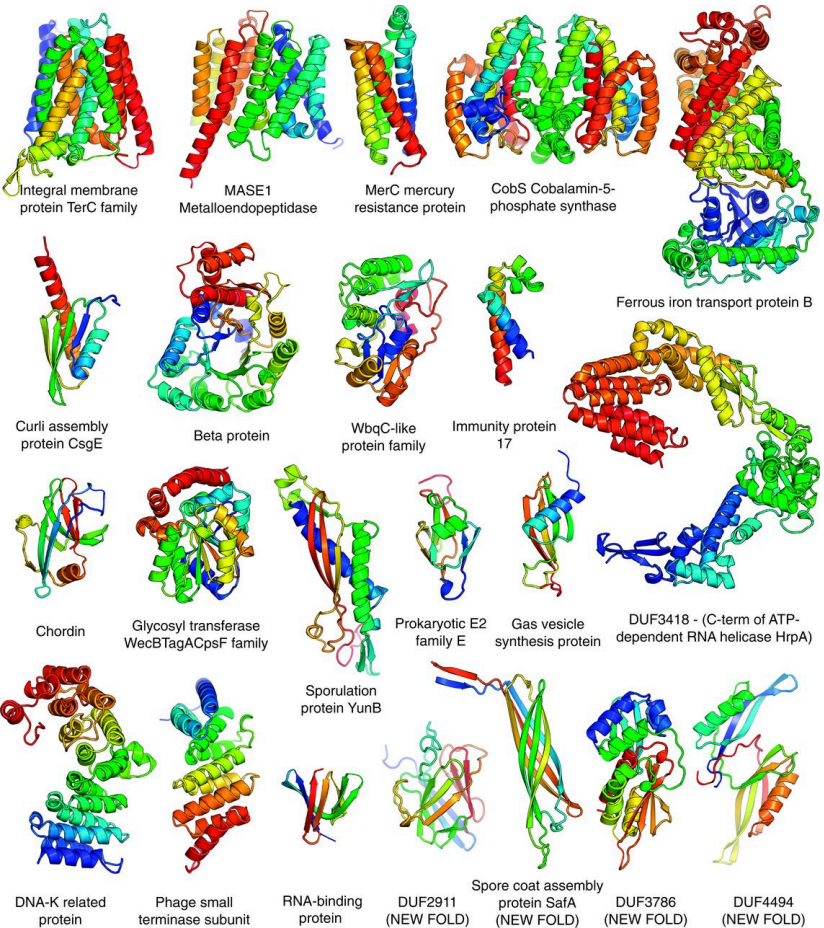
Protein structure determination using metagenome sequence data

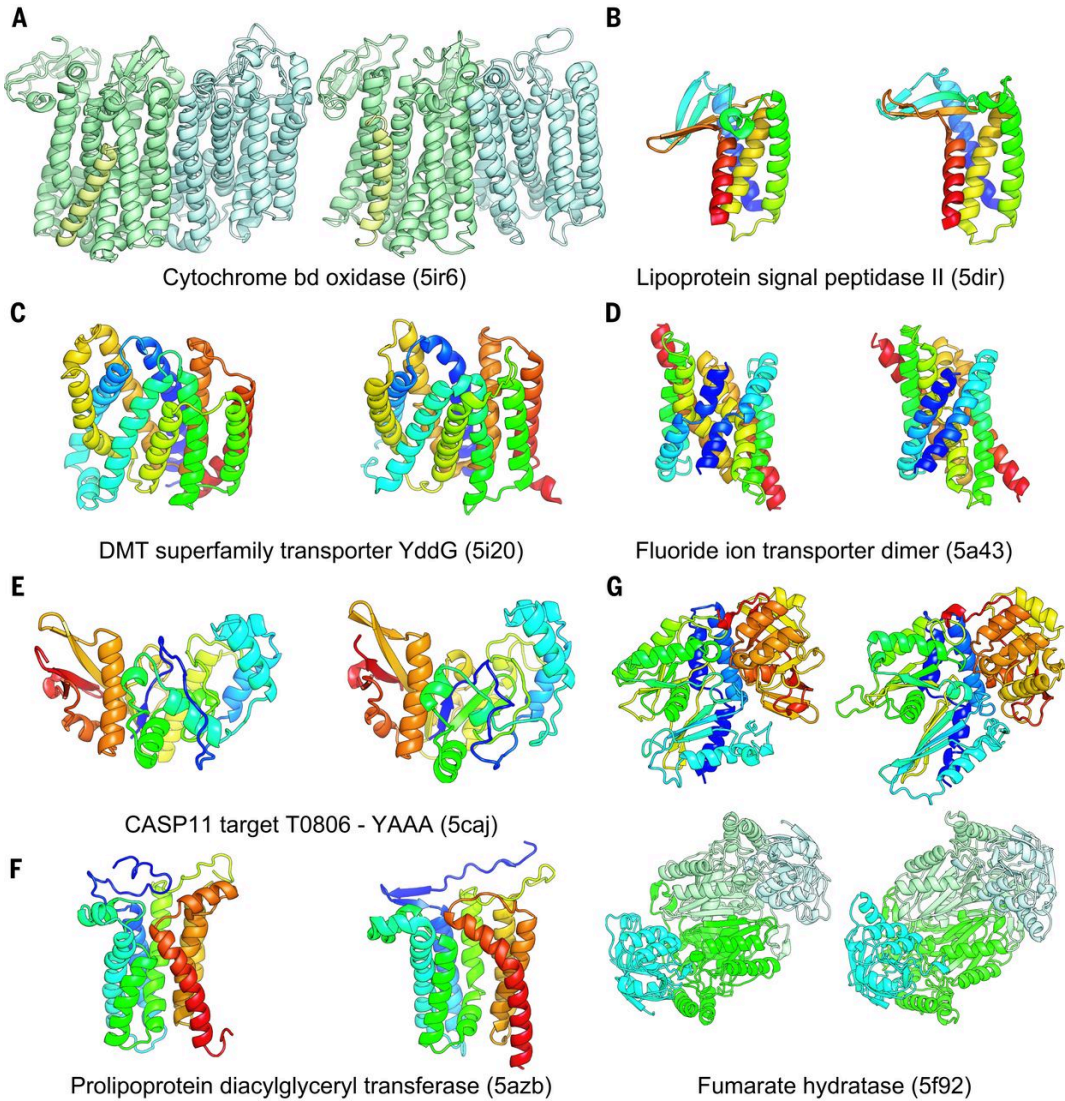
Sergey Ovchinnikov[1,2,3], Hahnbeom Park[1,2], Neha Varghese[4], Po-Ssu Huang[1,2], Georgios A. Pavlopoulos[4], David E. Kim[1,5], Hetunandan Kamisetty[6], Nikos C. Kyrpides[4,7], David Baker[1,2,5,*]

+ Author Affiliations

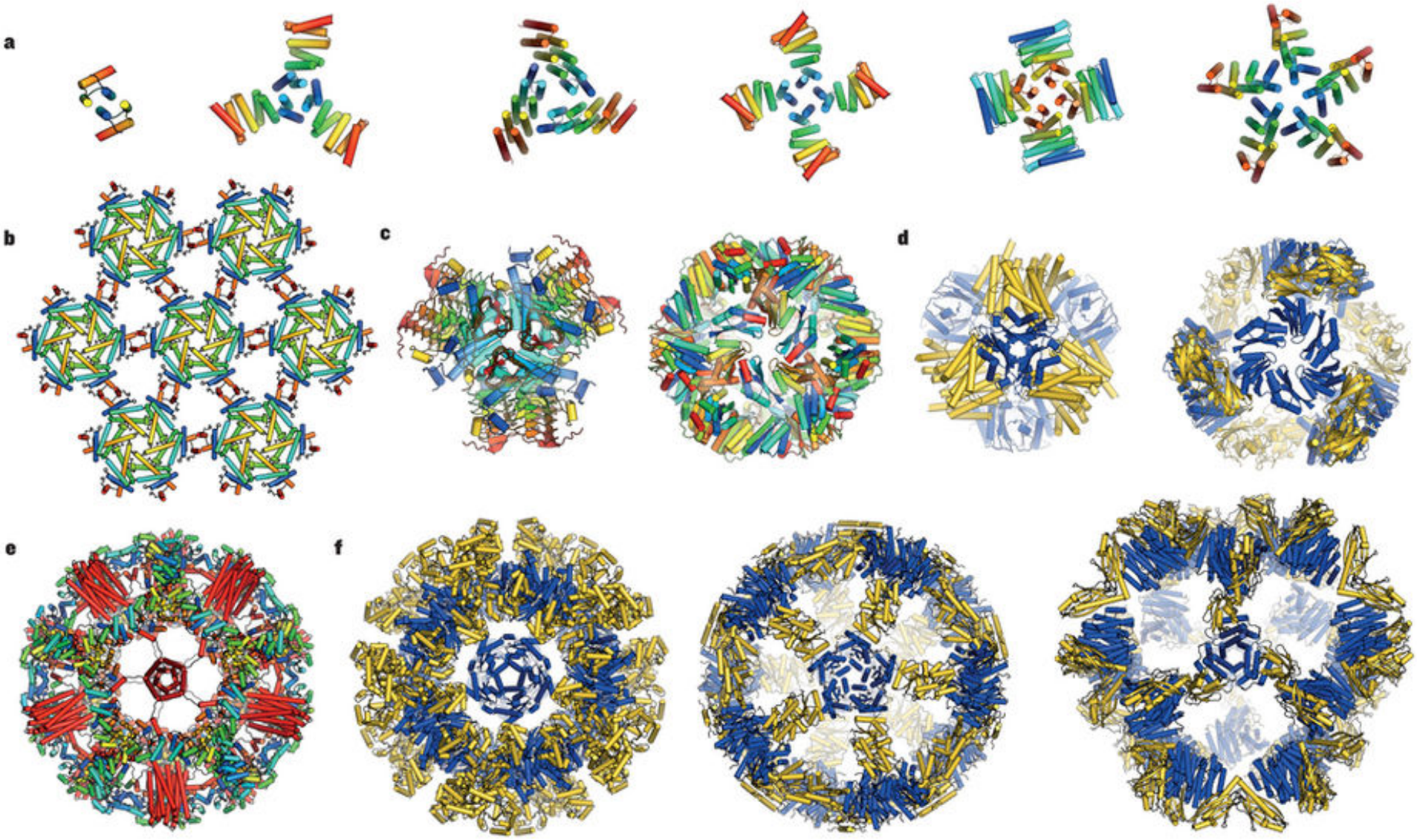
↵*Corresponding author. Email: dabaker@u.washington.edu

Science 20 Jan 2017:
Vol. 355, Issue 6322, pp. 294-298
DOI: 10.1126/science.aah4043





The future: computational protein design



References

- Databases: https://doi.org/10.1007/s12033-010-9372-4
- Proteopedia: http://proteopedia.org/wiki/index.php/Main_Page
- Coevolution: https://www.nature.com/articles/nbt.2419
- De novo protein design: https://www.nature.com/articles/nature19946

Thank you!

Have any questions, comments?
Email me: asm63@cam.ac.uk