

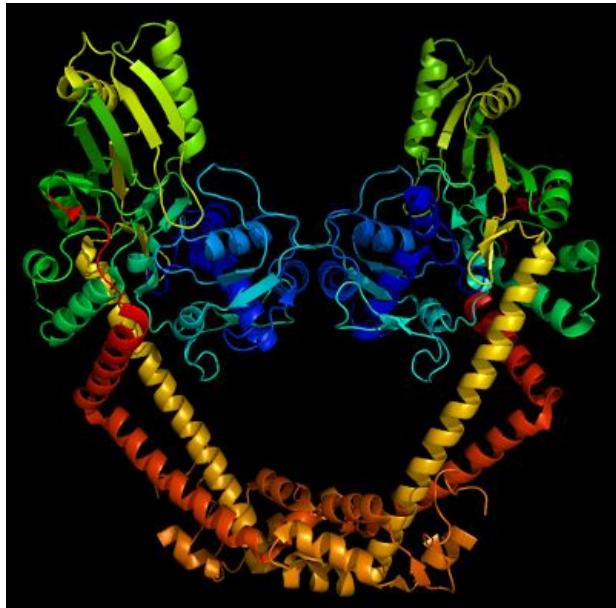
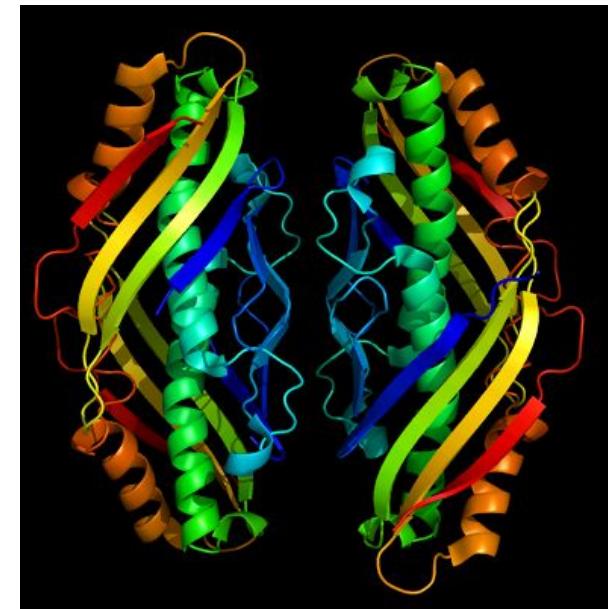
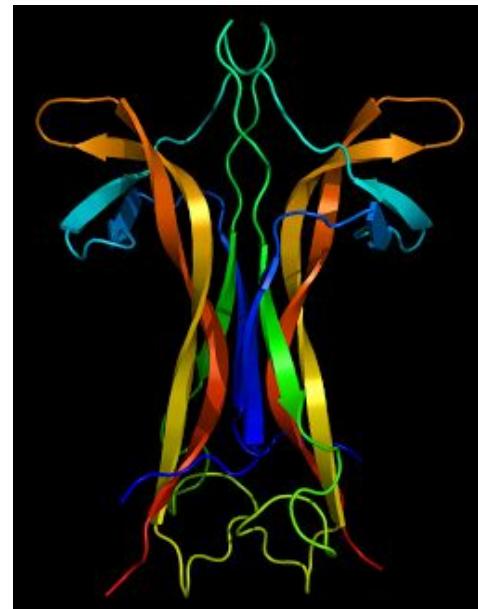
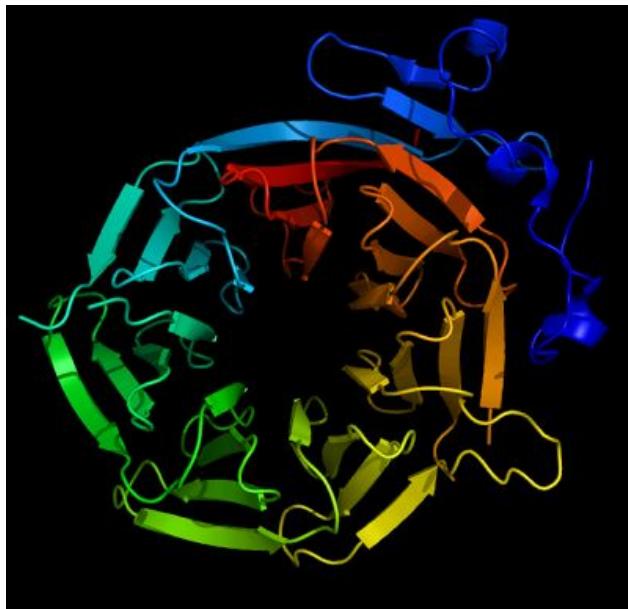
Structural Bioinformatics

Alexey Morgunov

Manifold Research Ltd.

University of Cambridge, UK

University of Tartu, Estonia



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



Aims of this lecture

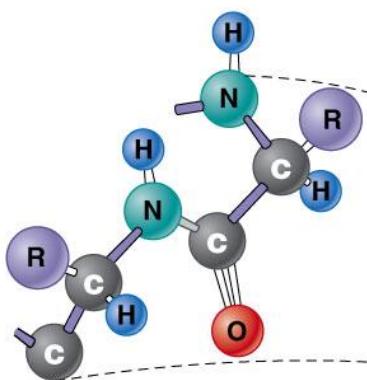
- A superficial introduction to a large and diverse field of structural bioinformatics
- Describing the landscape of resources available
- Focus on what knowledge can be gained by analysing protein structures
- Introducing some current directions and applications

What is structural bioinformatics?

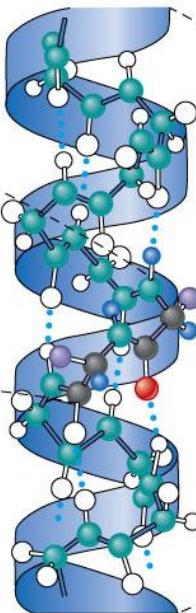
- Categorisation and descriptive analysis
- Comparative methods and techniques
- Theoretical foundations
- Predictive and design capability

Protein structure refresher

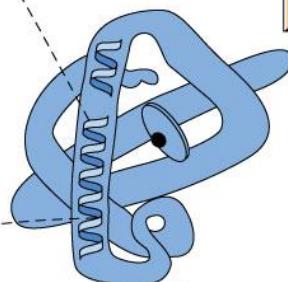
(a) Primary structure. The primary structure of a protein is a sequence of amino acids linked together by peptide bonds, forming a polypeptide.



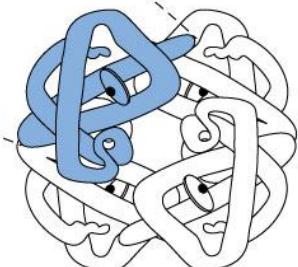
(b) Secondary structure. Local regions of the resulting polypeptide can then be coiled into an α helix, one form of secondary structure.

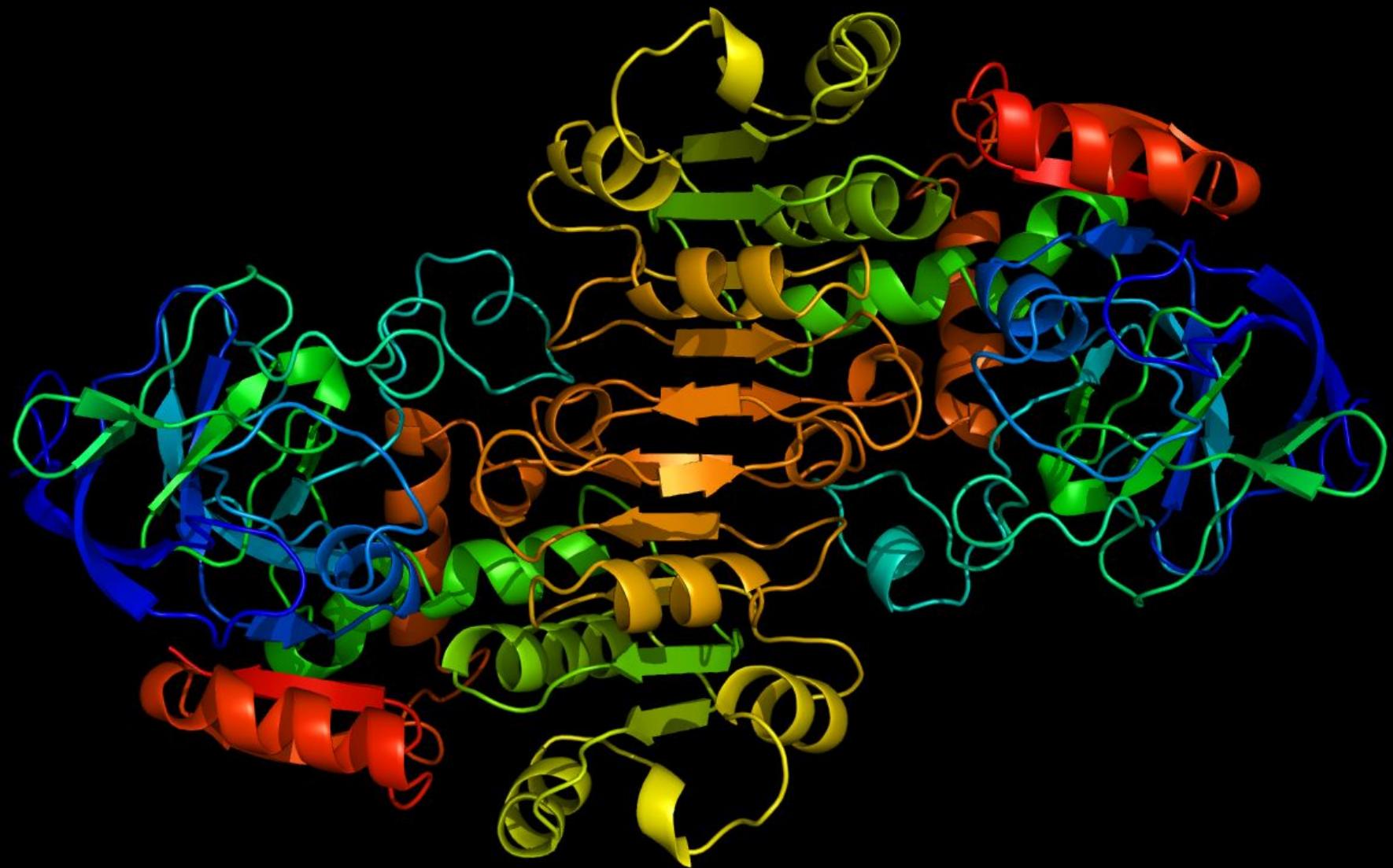


(c) Tertiary structure. Regions of secondary structure associate with each other in a specific manner to form the tertiary structure, which describes the final folding of the polypeptide.

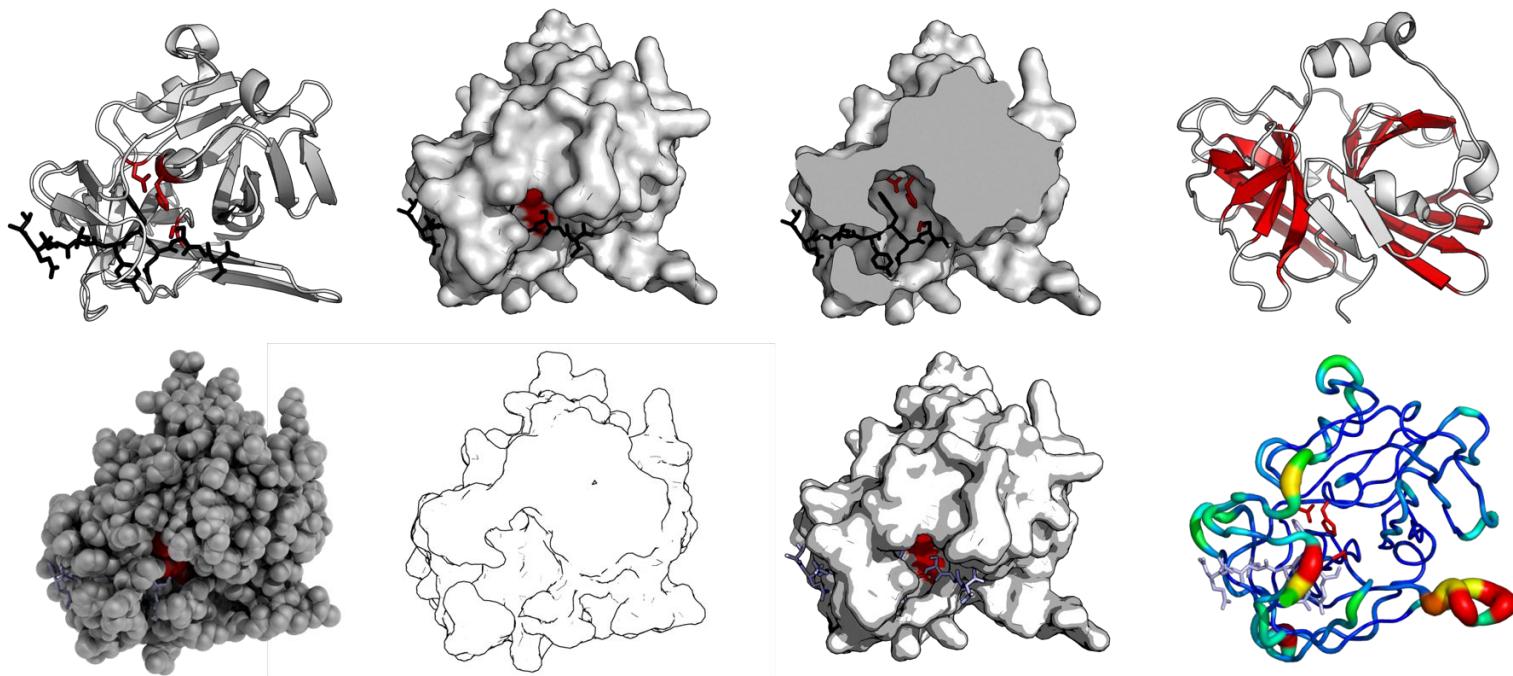


(d) Quaternary structure. For multimeric proteins, the quaternary structure describes the association of two or more polypeptides as they interact to form the final, functional protein.





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.).

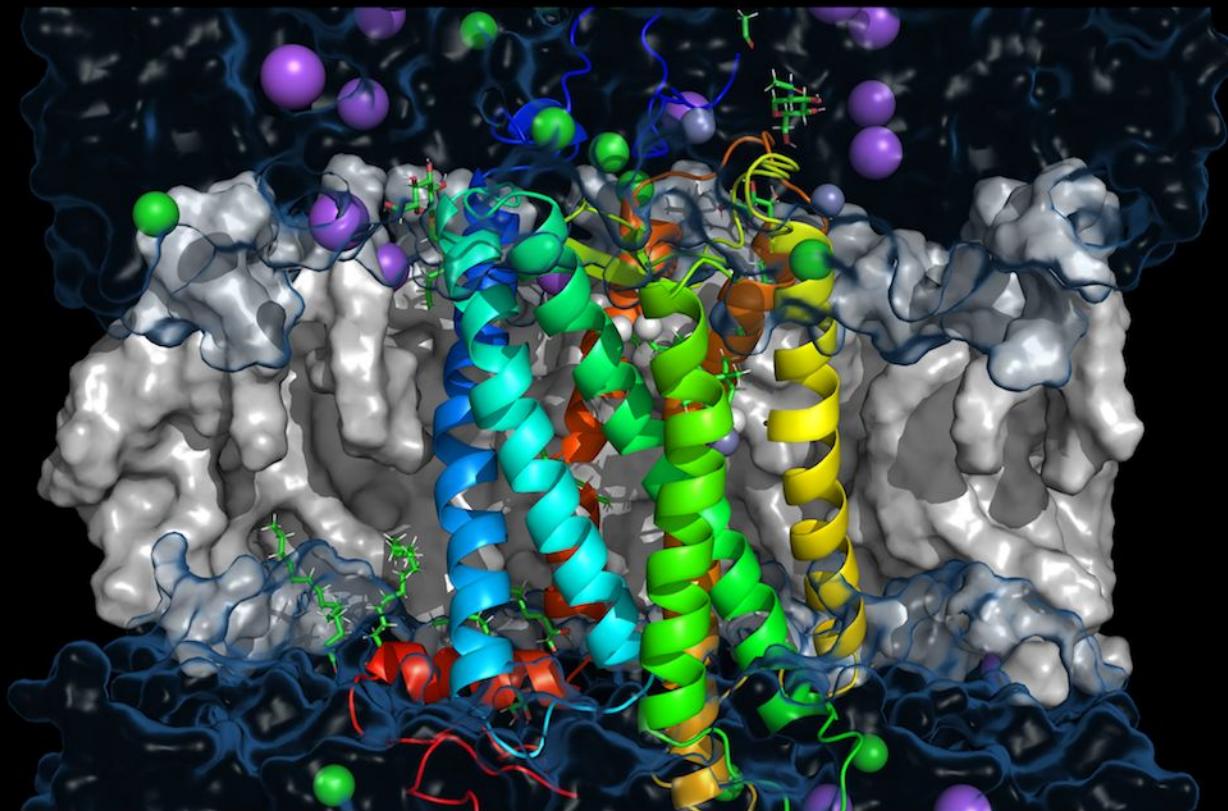
Reset Zoom Orient Draw/Ray

Unpick Deselect Rock Get View

< < Stop Play > > MClear

Builder Properties Rebuild

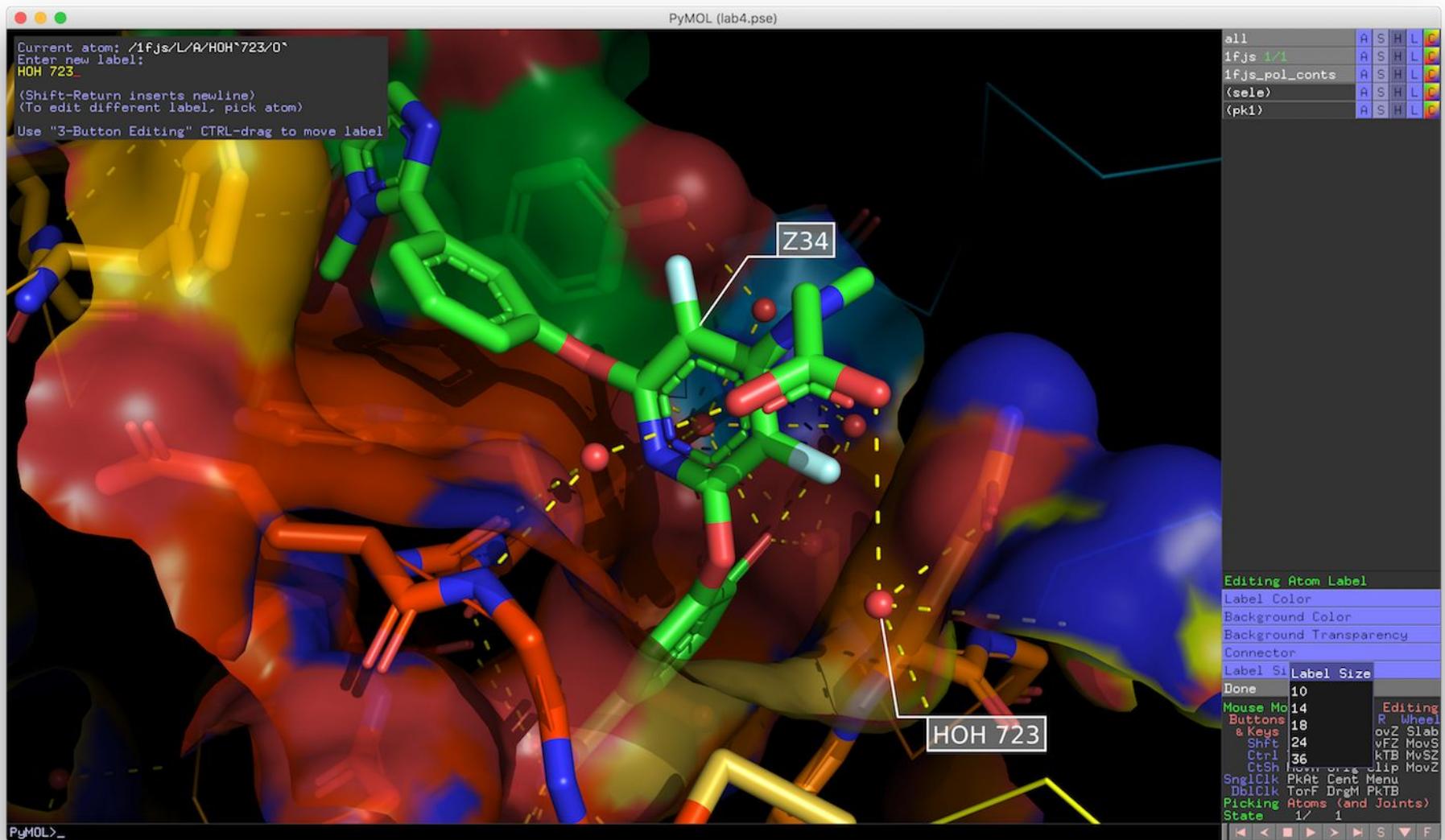
PyMOL>



all	A	S	H	L	F
ions 1/1	A	S	H	L	F
(sele)	A	S	H	L	C
water 1/1	A	S	H	L	C
protein 1/1	A	S	H	L	C
membrane1 1/1	A	S	H	L	C
membrane2 1/1	A	S	H	L	C

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 Selz Drig Clip MovZ
SnglClk // Cent Menu
DblClk Menu - PkAt
Selecting Residues
State 1/ 1

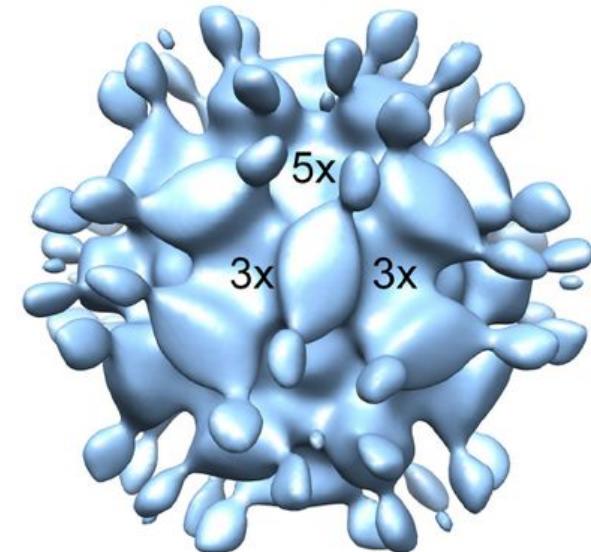
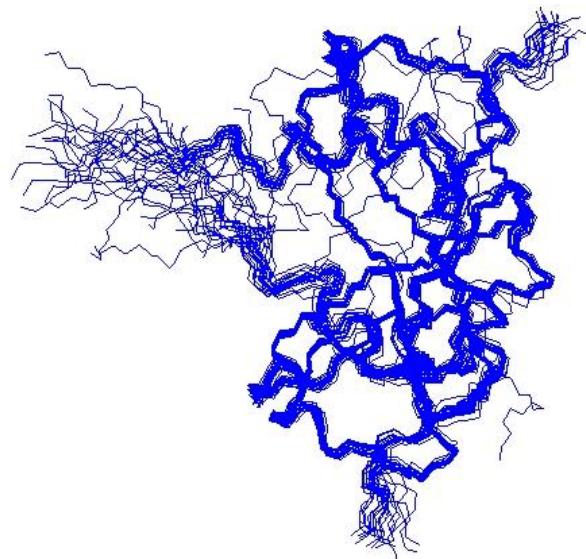
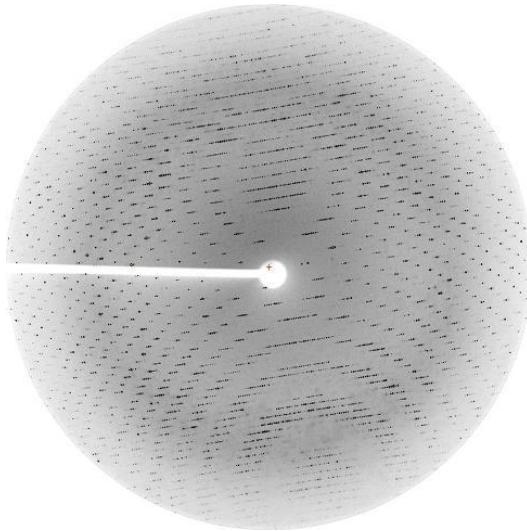
PyMOL>_



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>

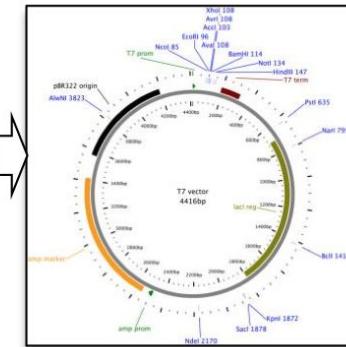


X-ray crystallography

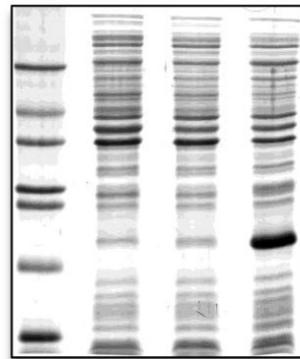
Sequence analysis

133 ATGGCCGAGTATTCTGAGGCGATCTTCTGAGGGATCCAA 177
134 I N A A V L E S I F L K R S Q 178
135 CAGAAAGAGAACATCACCTCTAACCTTAACCTTCAAGGGCCTTTT 222
136 O K K K T S P L N F K K R L F 30
223 CCTTGGACCGTACAAACCTCCACTATGATGTGATTGAA 267
31 L L T V H K L S Y Y E D F 45
268 CGTGGGAGAGGGACTAGAAGGGTTCATATGTTGAGG 312
46 R G R R S K G S I D V E K 68
313 ATCACTTGTTGAACAGTGGTCTGAAAAAATCTCTCCA 357
61 T C V T V V P E K N P P P 75
358 GAAAGAGATTCCGAGAAGGGAGAGCTGAGTGAATGGG 492
78 T R Q T E 96
493 CAACAGGATTCGAAAGTCTCTTCCCTTCCGTTCA 457
91 C C S I Z E R F P Y P F Q V V 185
448 TATGATGAAAGGCTCTTAGCTCTTCTCCAACTGAAAGTA 492
106 Y D E G P L Y V V F S P T E E L 120
492 AGGAGCGGTGATTCAACGCTAAAMAACTTAATCCGTCAC 537
121 R K R W I H Q L K N V I R Y N 135
538 AGTGTGTTGAAATAACCTCTTCTGATCGATGGG 582
136 S D L V Q K Y H P C F M W 150
583 CAGATCTGCGTCTCAAGACGCAAATGCTATGGCTGC 627
151 Q Y L C C S Q T A K N A M G C 165

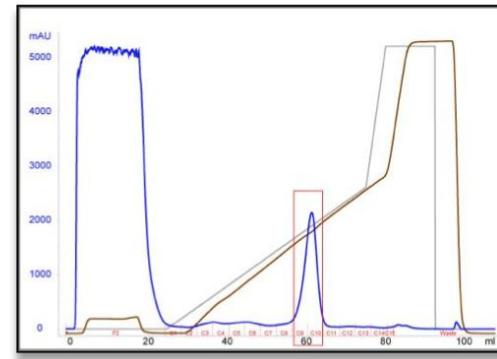
Expression construct



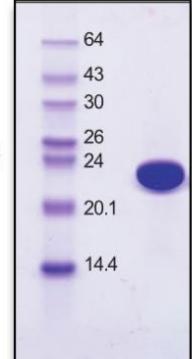
Recombinant expression



Protein purification



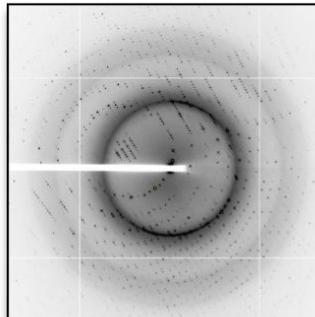
Homogenous protein



Crystallisation

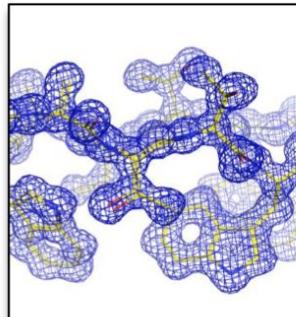


Diffraction data collection

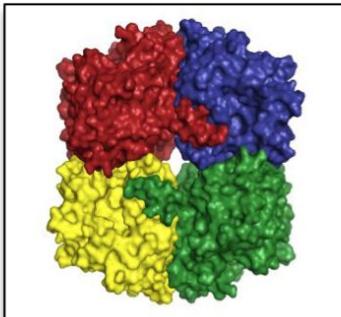


Phase problem solution

Fit model to electron density



The final structure



Cryo-EM

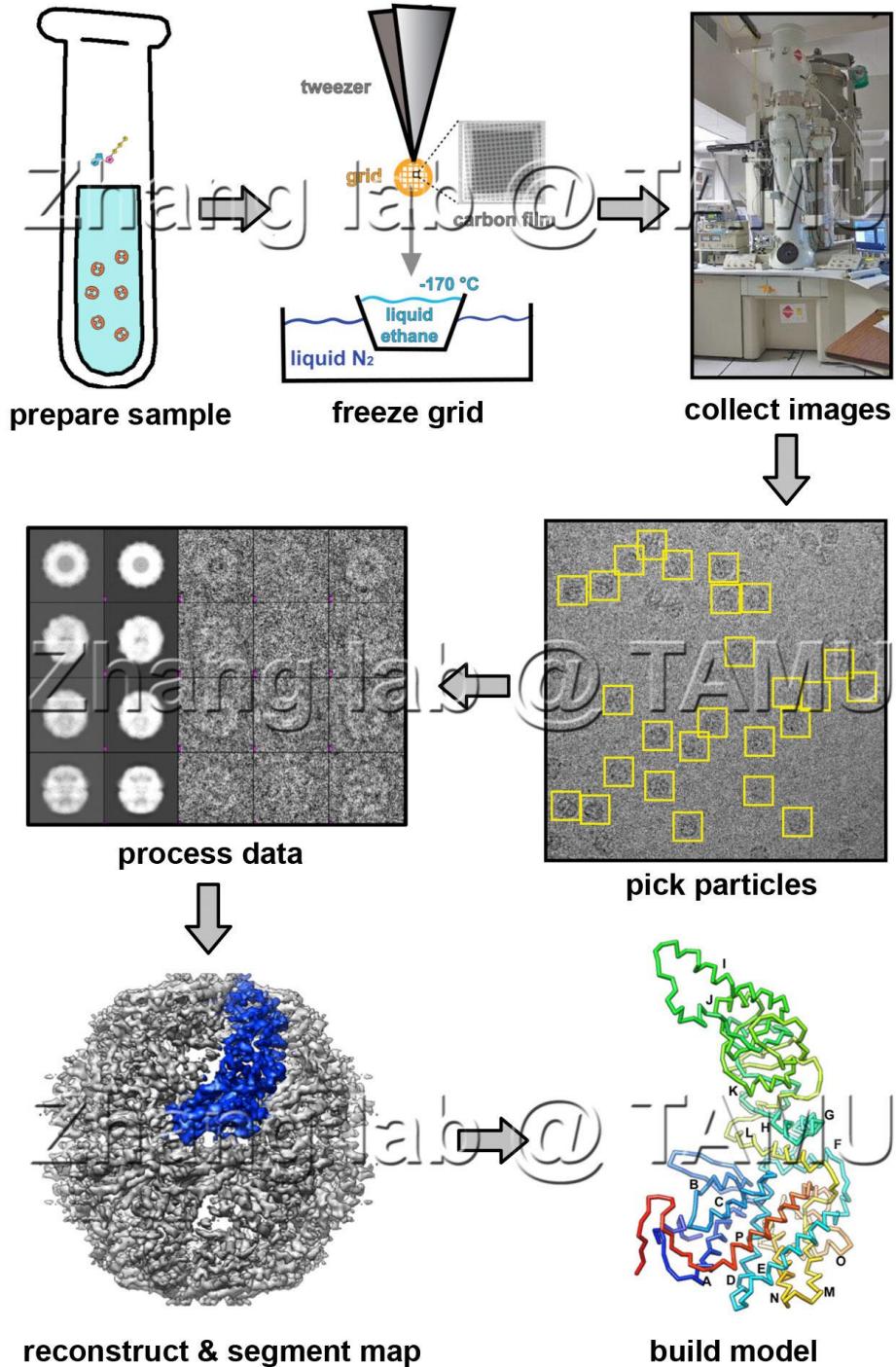
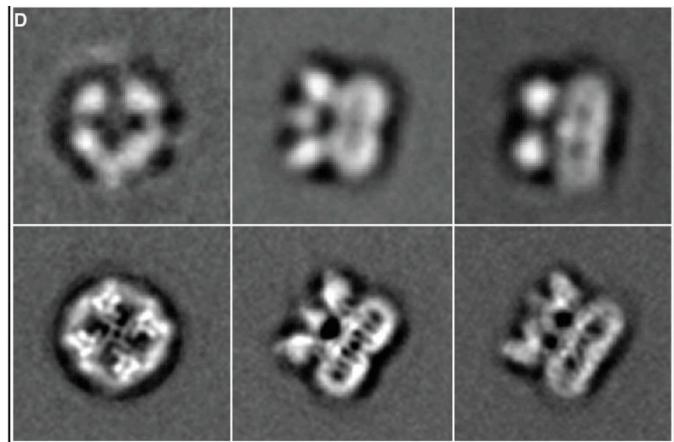
Can now do very high resolution

Less sample than X-ray

No need for crystals

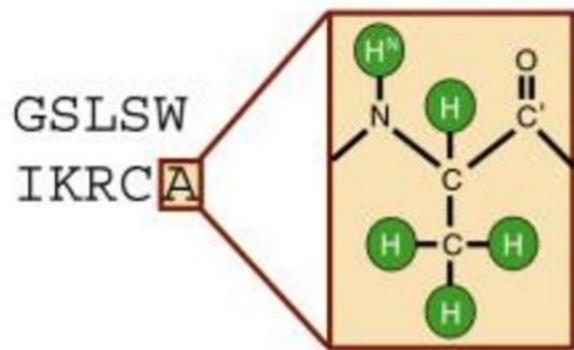
Much faster

[[doi:10.1016/j.tibs.2014.10.005](https://doi.org/10.1016/j.tibs.2014.10.005)]

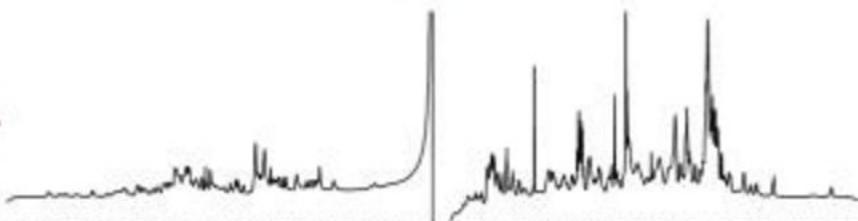


NMR

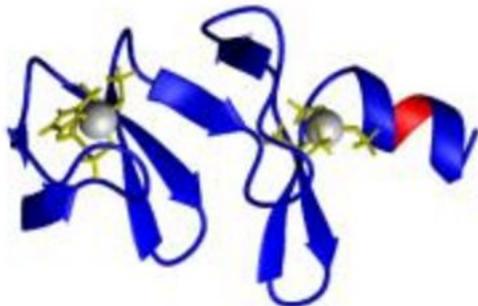
1. amino acid sequence



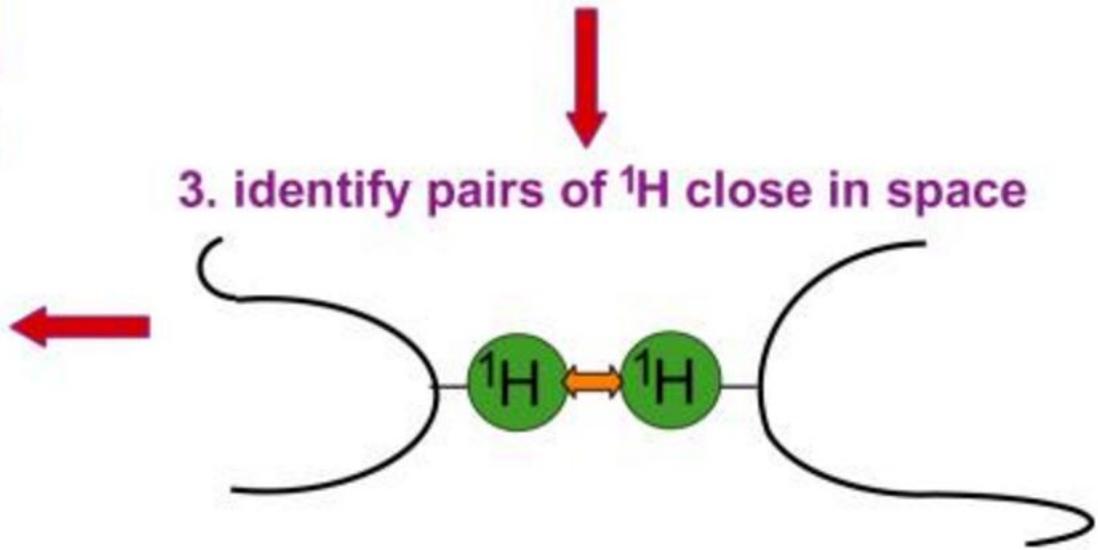
2. assign ^1H signals



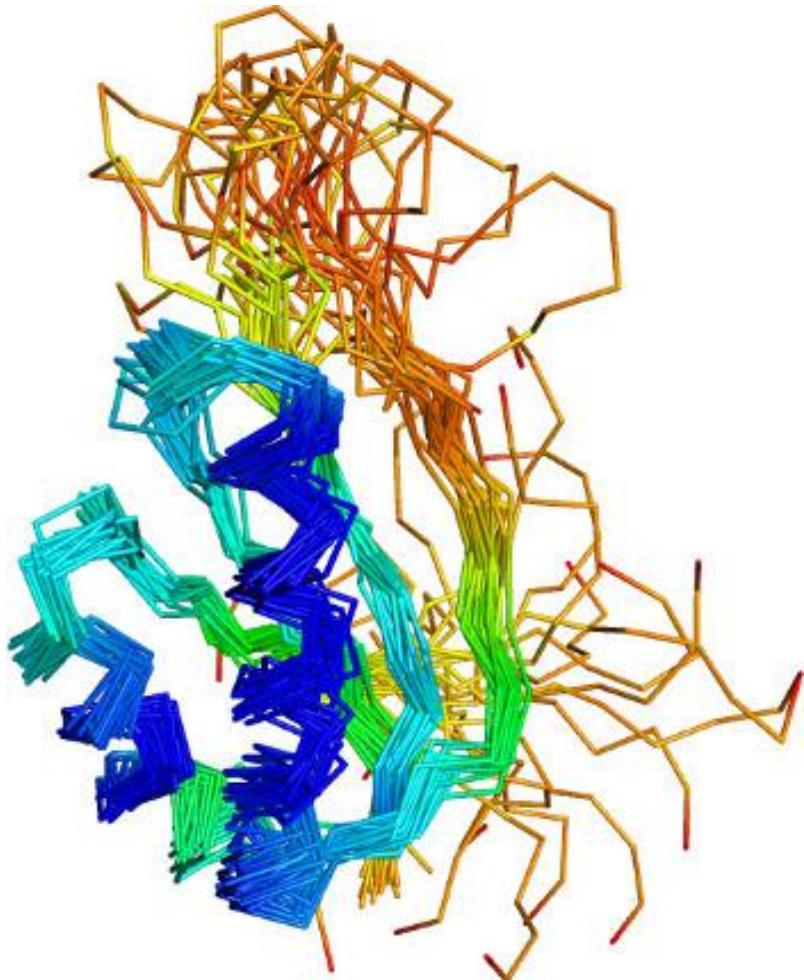
4. final structure solved



3. identify pairs of ^1H close in space



NMR



Ensemble of solutions to distance constraints

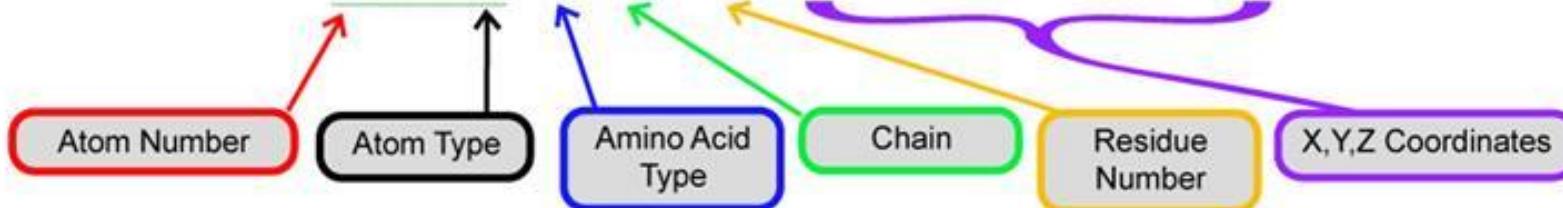
More rigid regions are more precisely determined

Useful in studying protein dynamics

How do we store a structural model?

PDB file format

ATOM	1132	NH1	ARG	A	149	31.814	-31.597	16.995
ATOM	1133	NH2	ARG	A	149	32.203	-32.934	18.816
ATOM	1134	N	ASN	A	150	29.346	-24.359	18.812
ATOM	1135	CA	ASN	A	150	28.480	-23.190	18.933
ATOM	1136	C	ASN	A	150	28.606	-22.168	17.808
ATOM	1137	O	ASN	A	150	27.803	-21.276	17.678
ATOM	1138	CB	ASN	A	150	28.732	-22.524	20.282
ATOM	1139	CG	ASN	A	150	28.284	-23.389	21.447
ATOM	1140	OD1	ASN	A	150	27.205	-23.981	21.430
ATOM	1141	ND2	ASN	A	150	29.110	-23.463	22.466
ATOM	1142	N	LEU	A	151	29.629	-22.313	16.996
ATOM	1143	CA	LEU	A	151	29.868	-21.415	15.894
ATOM	1144	C	LEU	A	151	29.953	-22.205	14.597
ATOM	1145	O	LEU	A	151	30.149	-23.422	14.614
ATOM	1146	CB	LEU	A	151	31.208	-20.735	16.100
ATOM	1147	CG	LEU	A	151	31.436	-19.884	17.337
ATOM	1148	CD1	LEU	A	151	32.846	-19.333	17.256



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/>

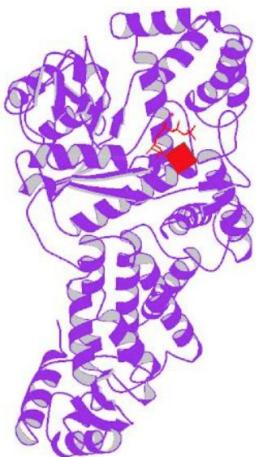


Biological Assembly 1

2ZUE

Display Files

Download Files



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.

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

DOI: [10.22110/pdb2ZUE/pdb](https://doi.org/10.22110/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

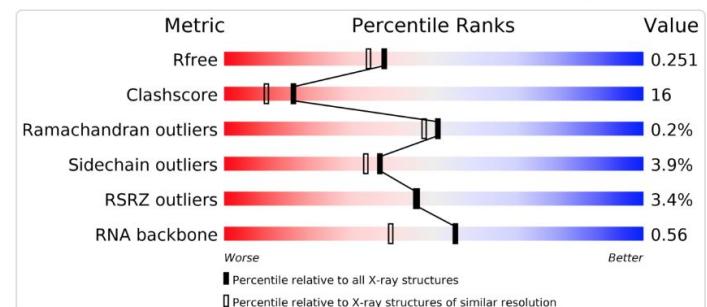
Resolution: 2 Å

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))

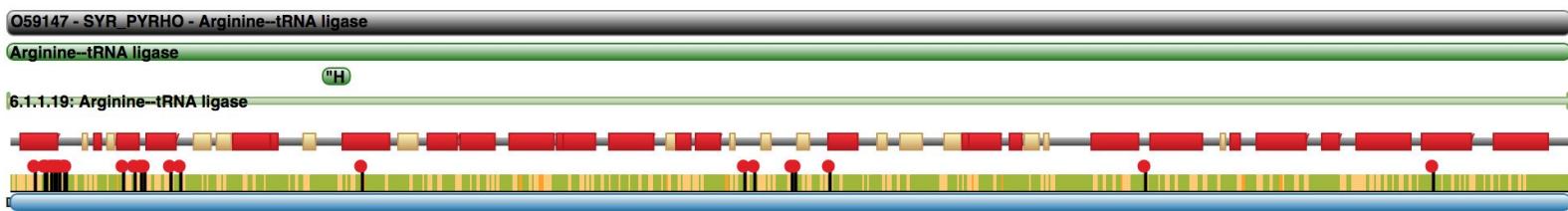
Go to UniProtKB: [O59147](#)

Protein Feature View

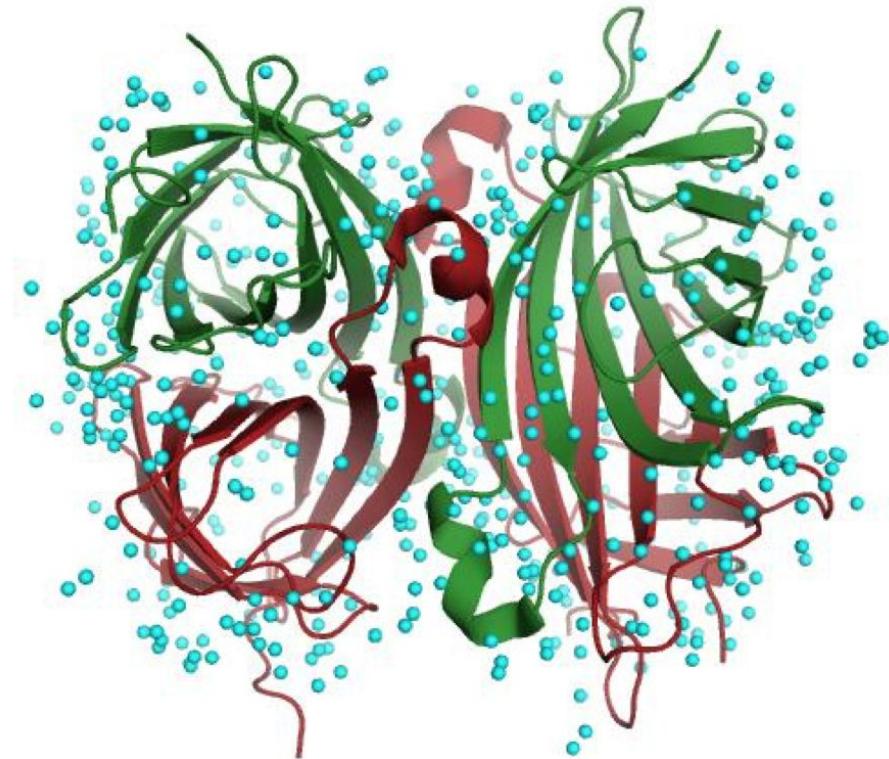
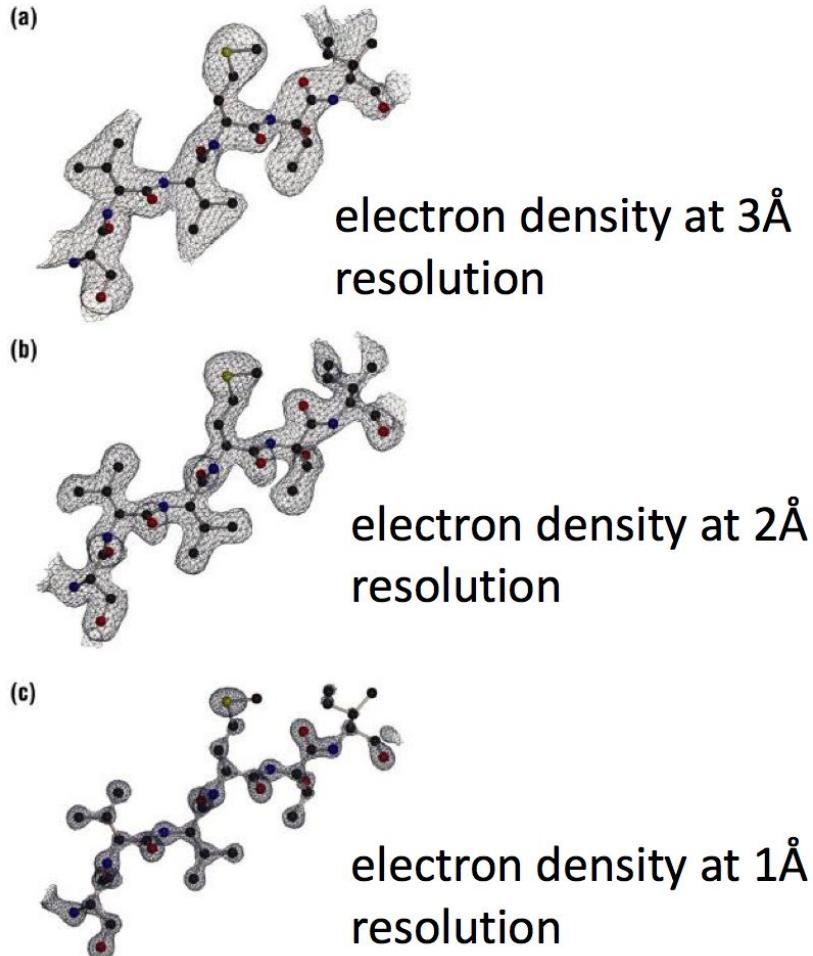


Full Protein Feature View for [O59147](#)

O59147
Mole. Processing
Motif
E.C.
Secstruc
PDB Validation
PDB
2ZUE.A
+



What is a good structural model?



Extensive water network seen in a high resolution structure

At high resolution, (1 \AA 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.* [[doi:10.3389/fnhum.2018.00037](https://doi.org/10.3389/fnhum.2018.00037); [doi:10.1107/S0907444907033847](https://doi.org/10.1107/S0907444907033847)]
- 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 Å² signifies confidence in the position of a residue

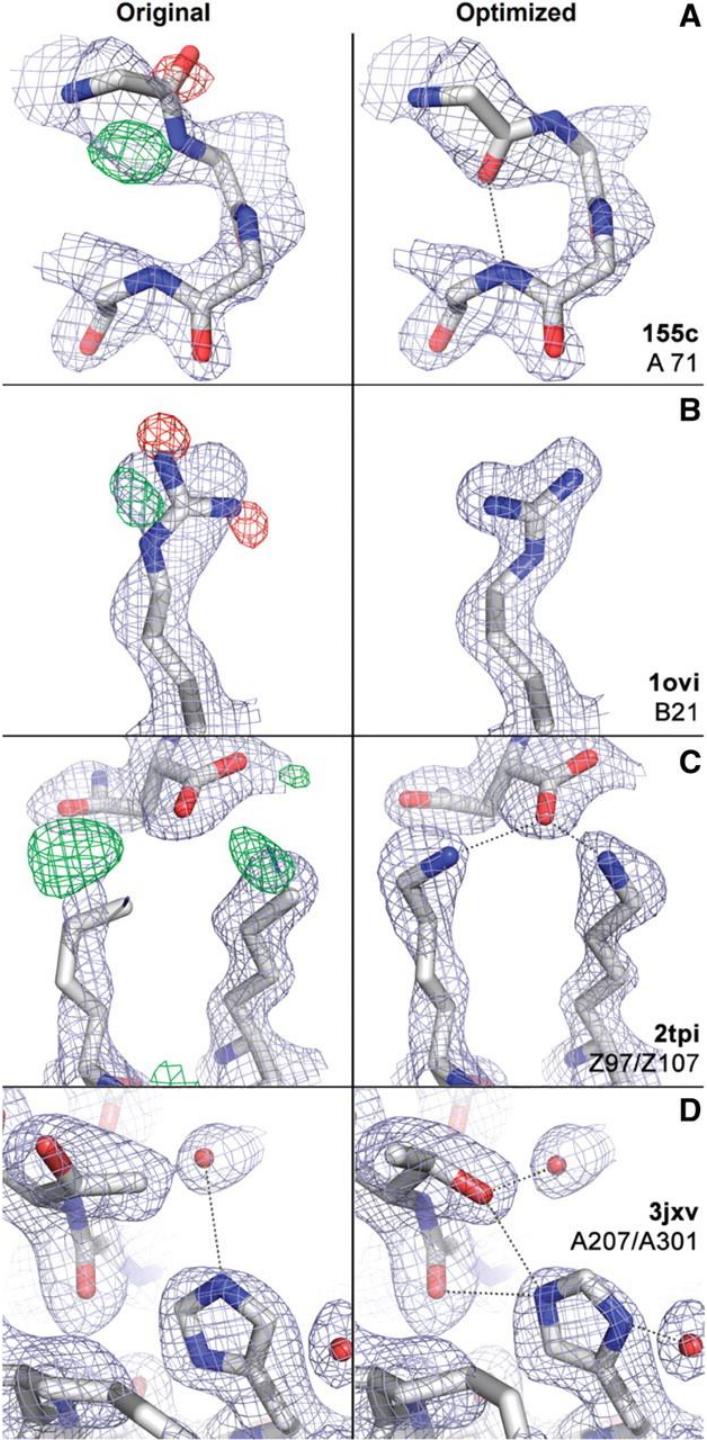
2ZUE coloured by B factor



PDB-REDO

- Optimised structures
- Refinement, rebuilding and validation
- Typically, these are better quality structures and should be used instead of originals!
- The method introduces consistency – better for large scale analyses

<https://pdb-redo.eu>



What can we learn from the structural model? Atlases

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

EBI > Databases > Structure Databases > PDBsum

PDBsum Pictorial database of 3D structures in the Protein Data Bank

PDBsum is a pictorial database that provides an at-a-glance overview of the contents of each 3D structure deposited in the Protein Data Bank (PDB). It shows the molecule(s) that make up the structure (ie protein chains, DNA, ligands and metal ions) and schematic diagrams of the interactions between them. [Read more ...](#)



PDB code (4 chars) **Find** Example: "1kfv"

Text search **Search**

Scans all TITLE, HEADER, COMPND, SOURCE and AUTHOR records in the PDB (eg to find a given protein by name).

Search by sequence

Search

Perform FASTA search vs all sequences in the PDB to get a list of the closest matches.

Search by

UniProt id:

(eg P03023, LAC1_ECOLI, etc)

Search

Pfam id:

(eg PF07992)

Search

Ensembl id:

(eg ENSG00000086205, ENST00000256999)

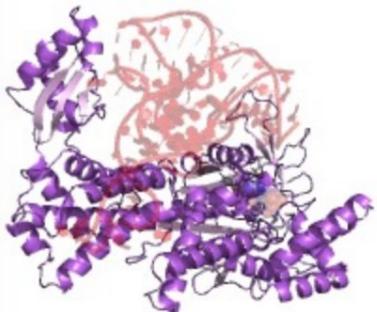
Search

Chain A (628 residues)

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



Protein chain A highlighted
(click to view)

Jmol Snap

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

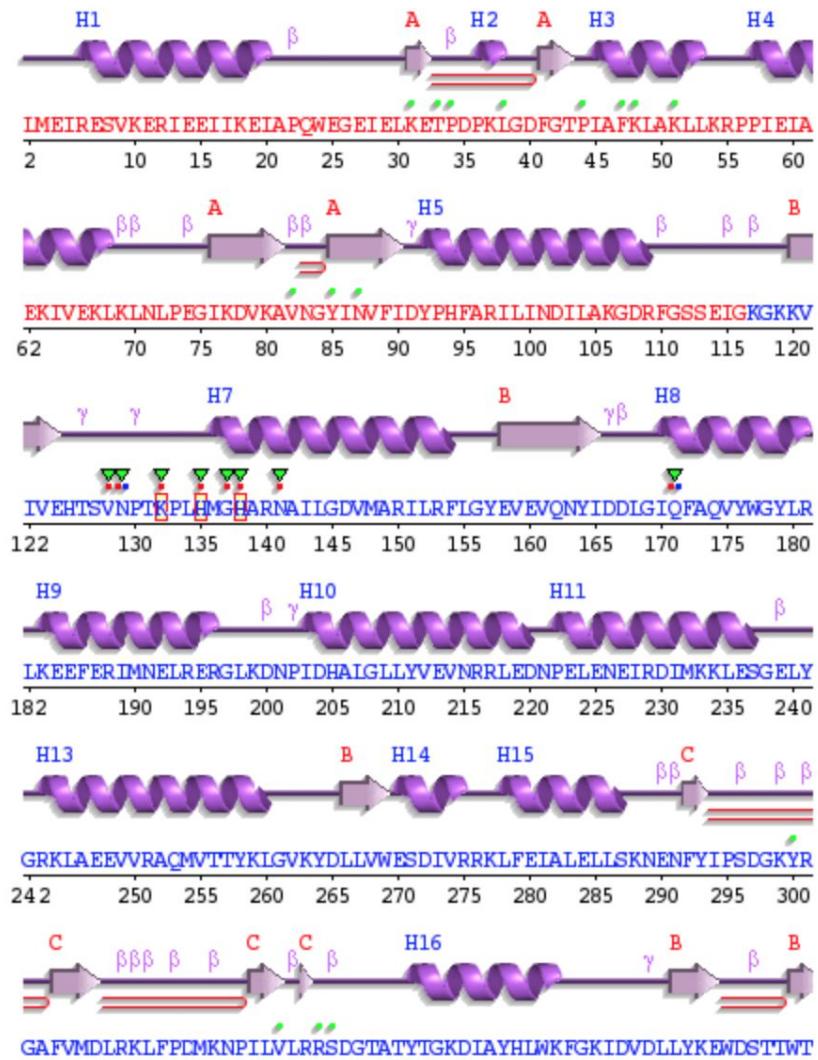
40 helix-helix interacs

43 beta turns

7 gamma turns

Catalytic residues

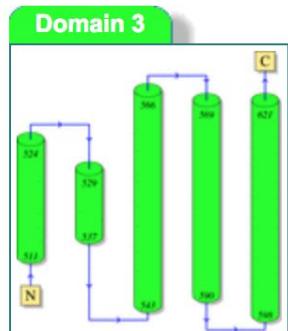
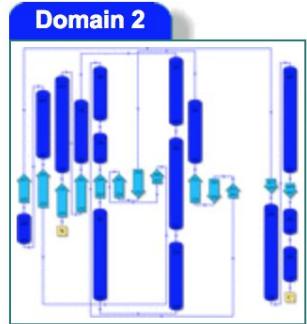
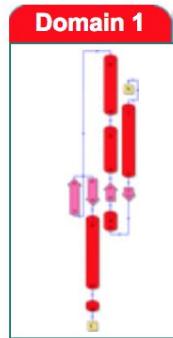
H135-H138-K132



Jmol



FASTA file

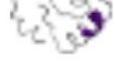
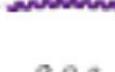


What can we learn from the structural model?

Classification

- Structural Classification of Proteins (SCOP)
 - SCOP2 (recently released):
<http://scop2.mrc-lmb.cam.ac.uk/>
 - SCOPe/ASTRAL (extension of v1, most used):
<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)

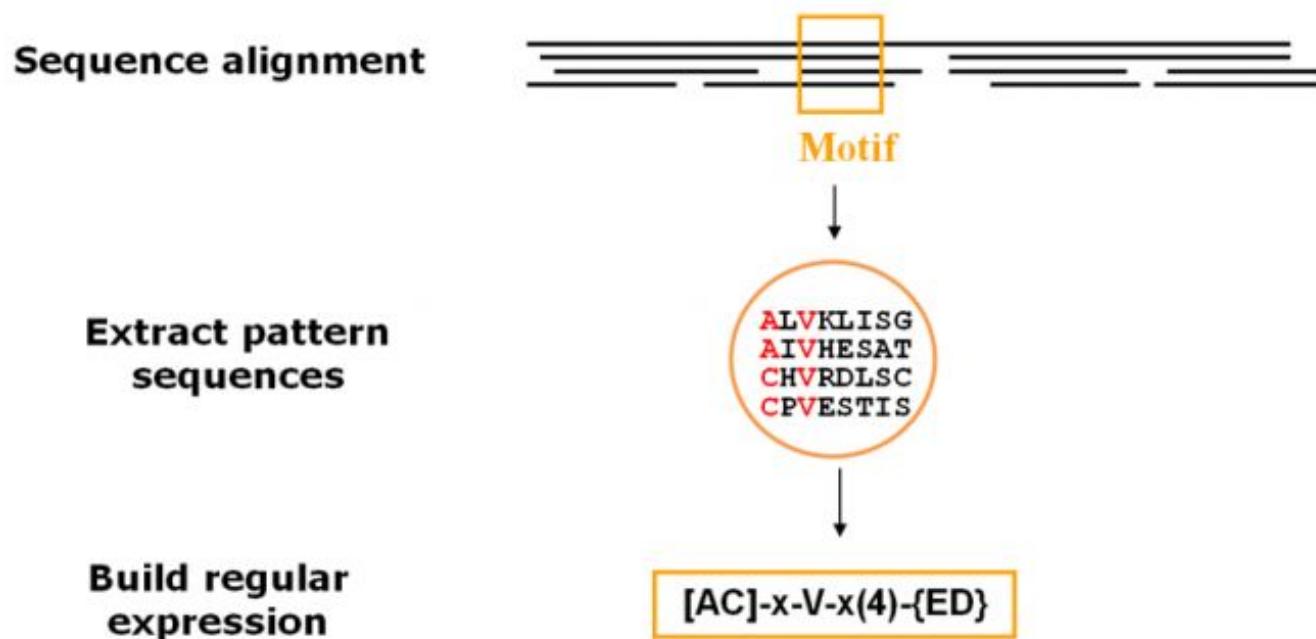
What can we learn from the structural model?

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

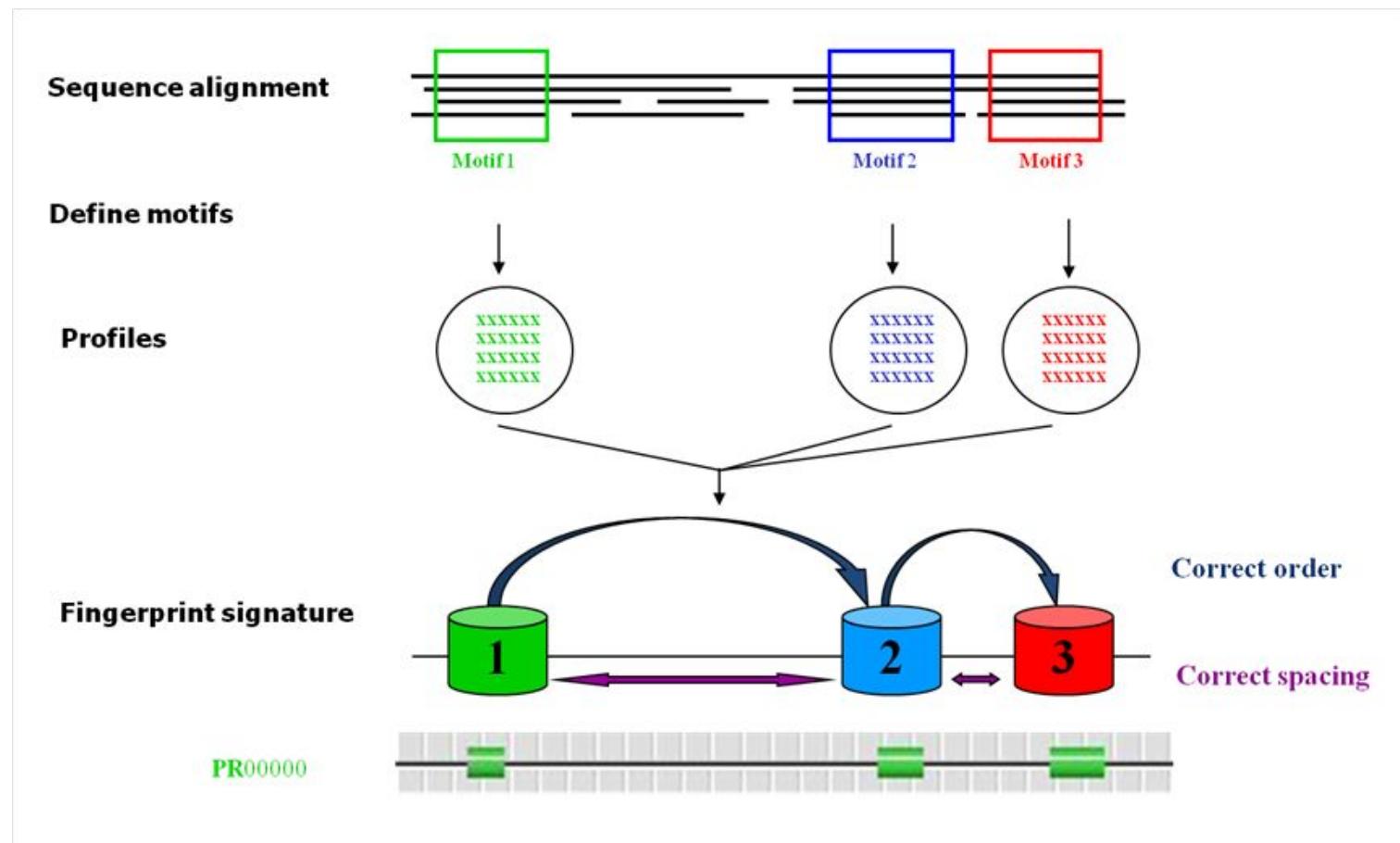
Classification is based on protein signatures

- **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

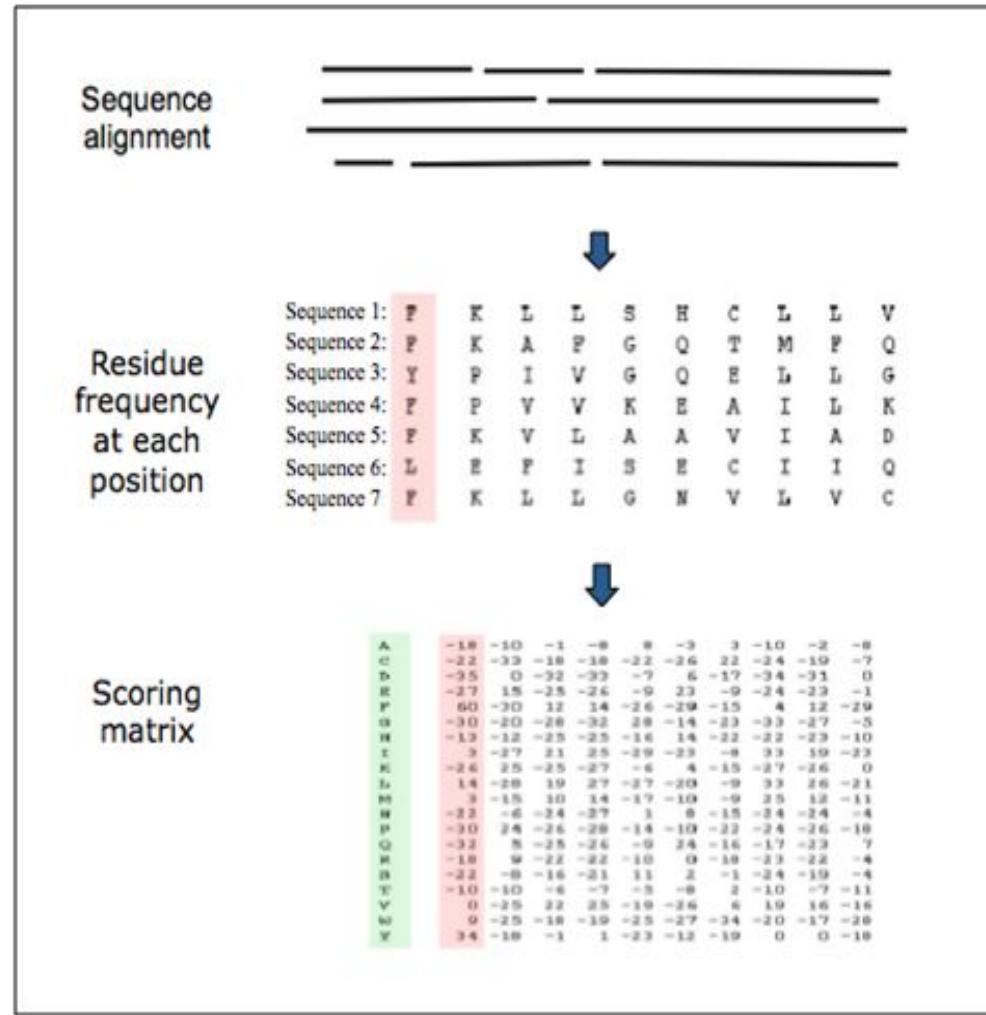
- **Fingerprints** – consist of multiple short conserved motifs, which are drawn from sequence alignments (e.g. [PRINTS](#))



Classification is based on protein signatures

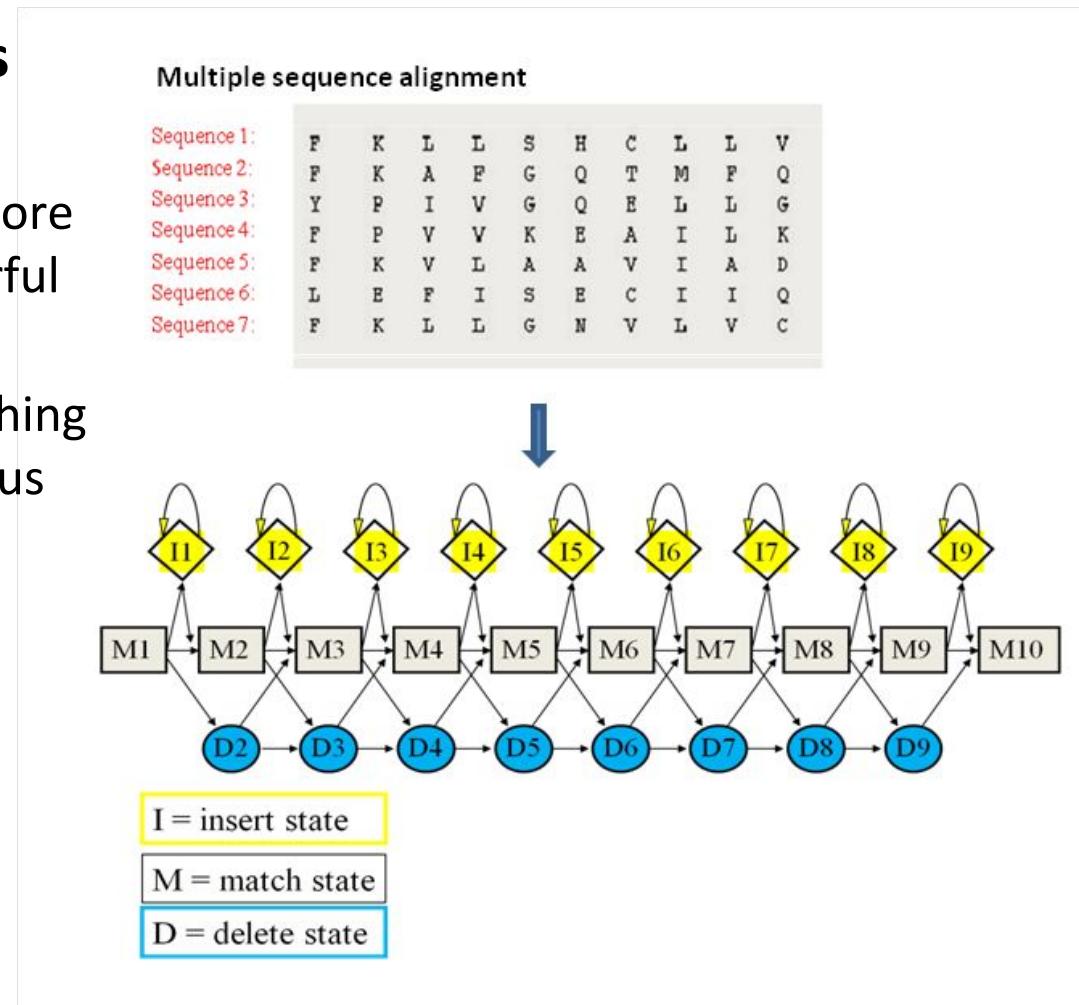
- **Profiles**

- built by converting multiple sequence alignments into position-specific scoring matrices (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 to these scores.
- E.g. [HAMAP](#), [PROSITE](#)



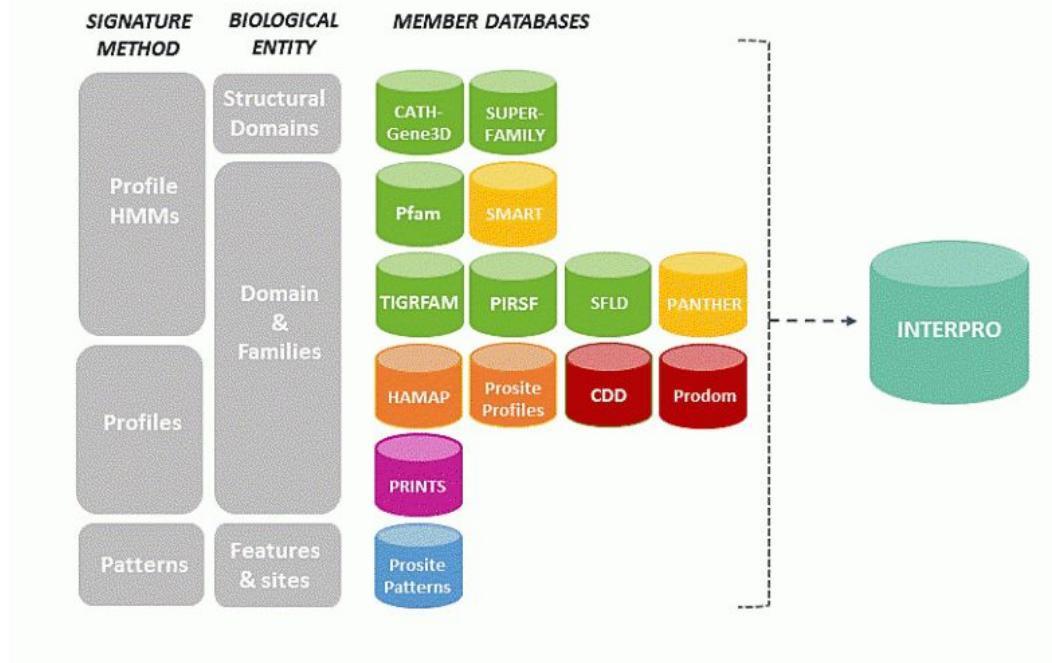
Classification is based on protein signatures

- **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](#)



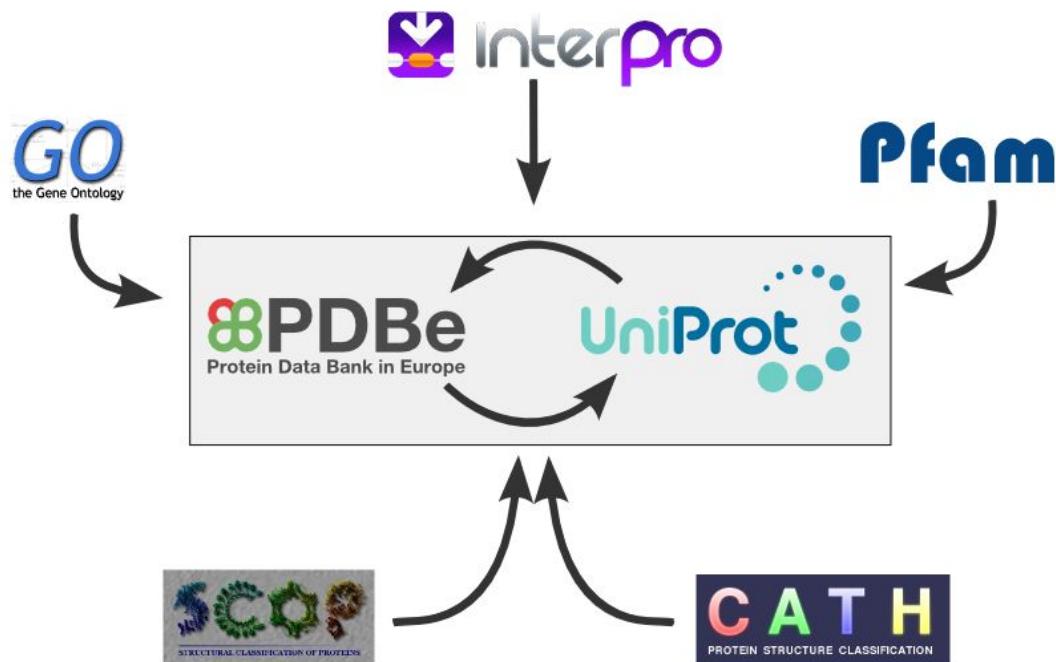
Integrating resources

- 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>

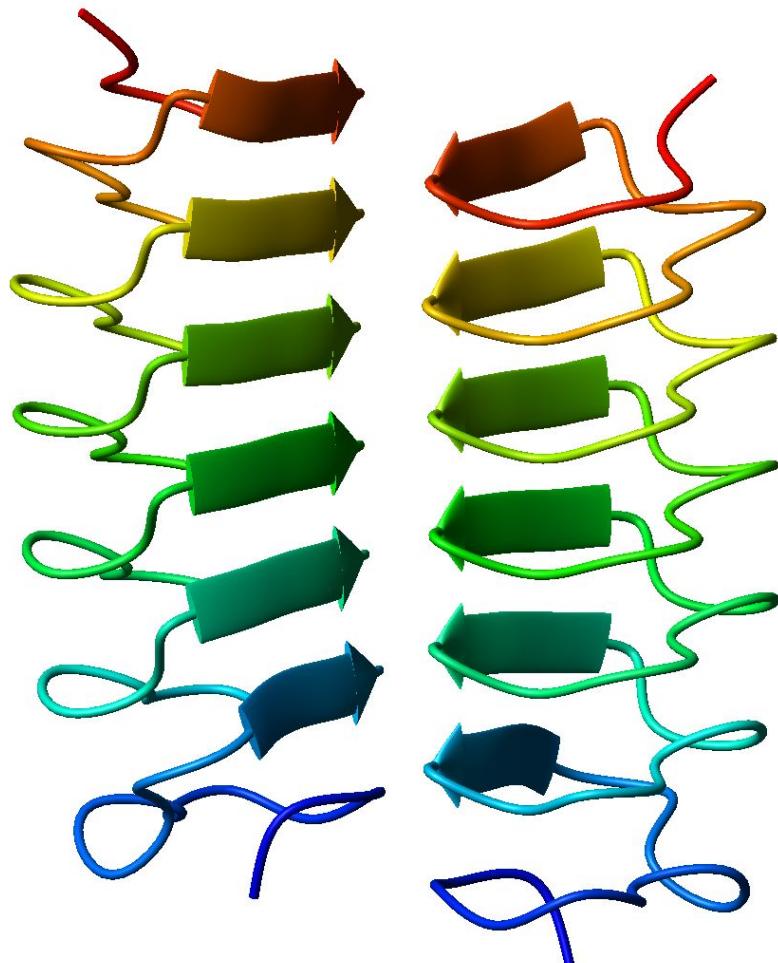


Integrating resources

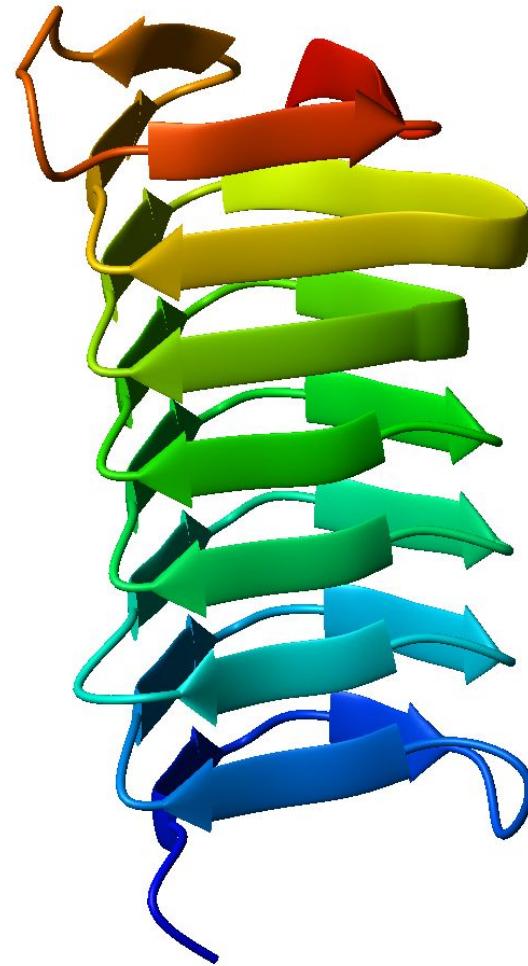
- Structure Integration with Function, Taxonomy and Sequence (SIFTS) - <https://www.ebi.ac.uk/pdbe/docs/sifts/index.html>



Intermission! BETA helices

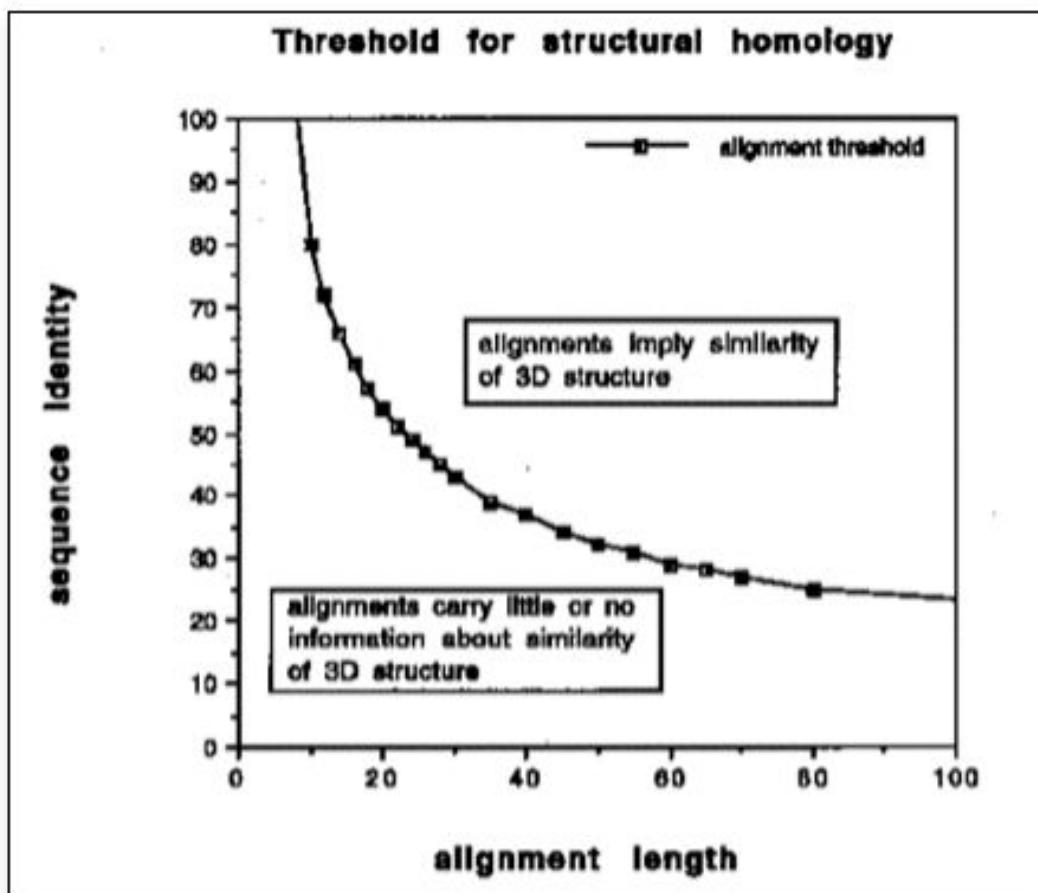


1EZG



1M8N

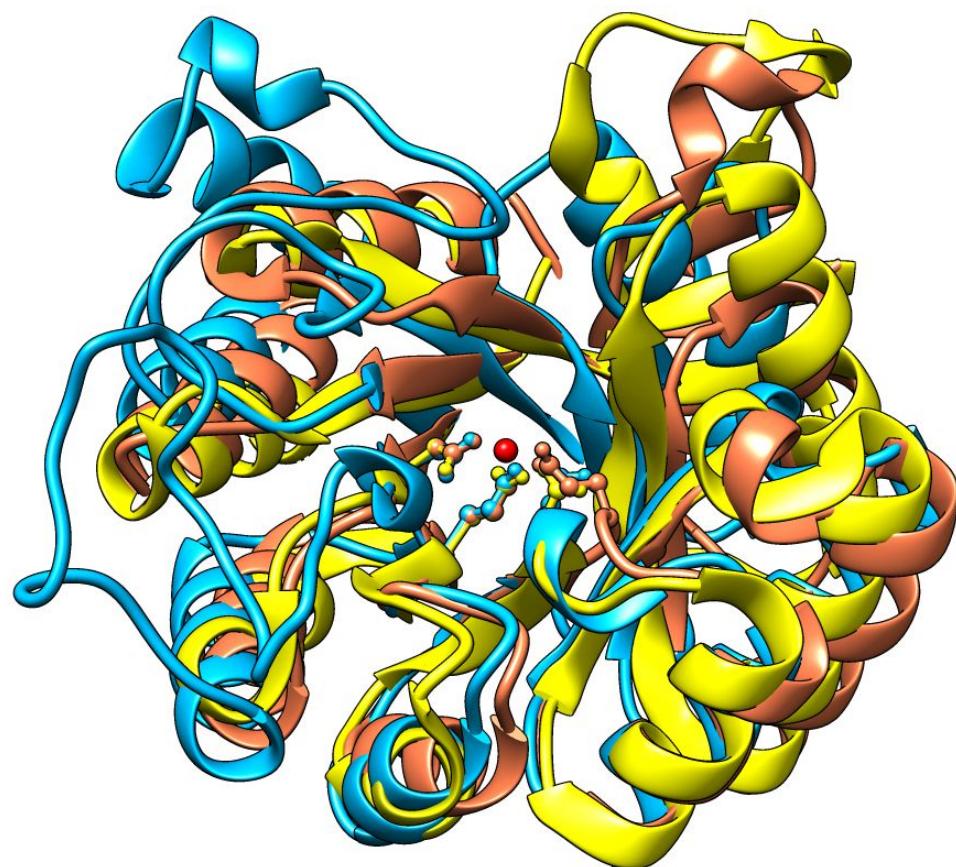
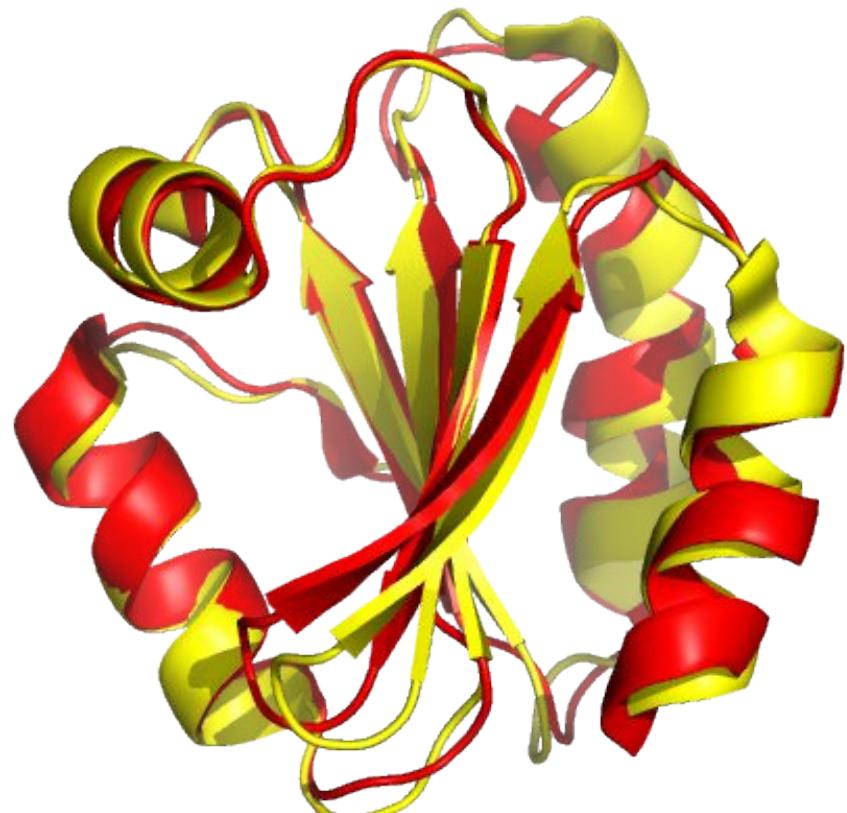
Structure is more conserved than sequence



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

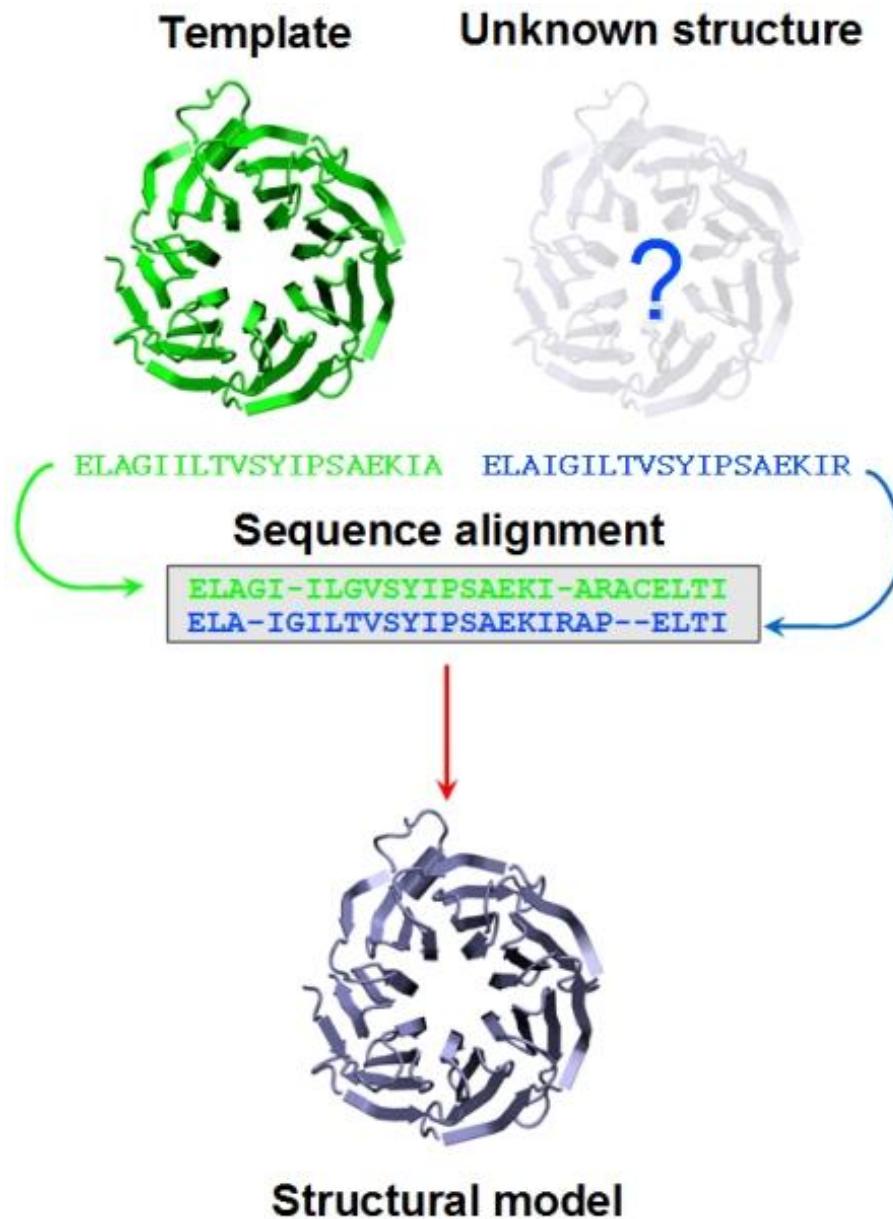
$$\text{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2}$$



What if I don't have a structure?

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

Homology modelling



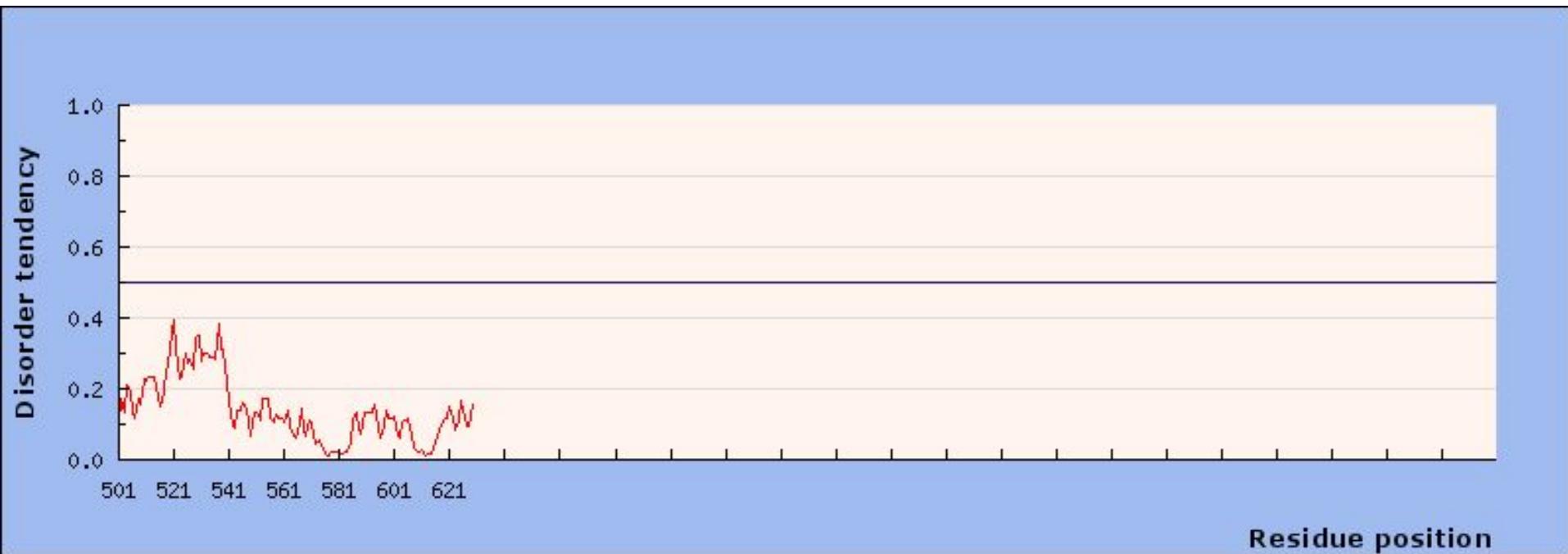
Homology modelling

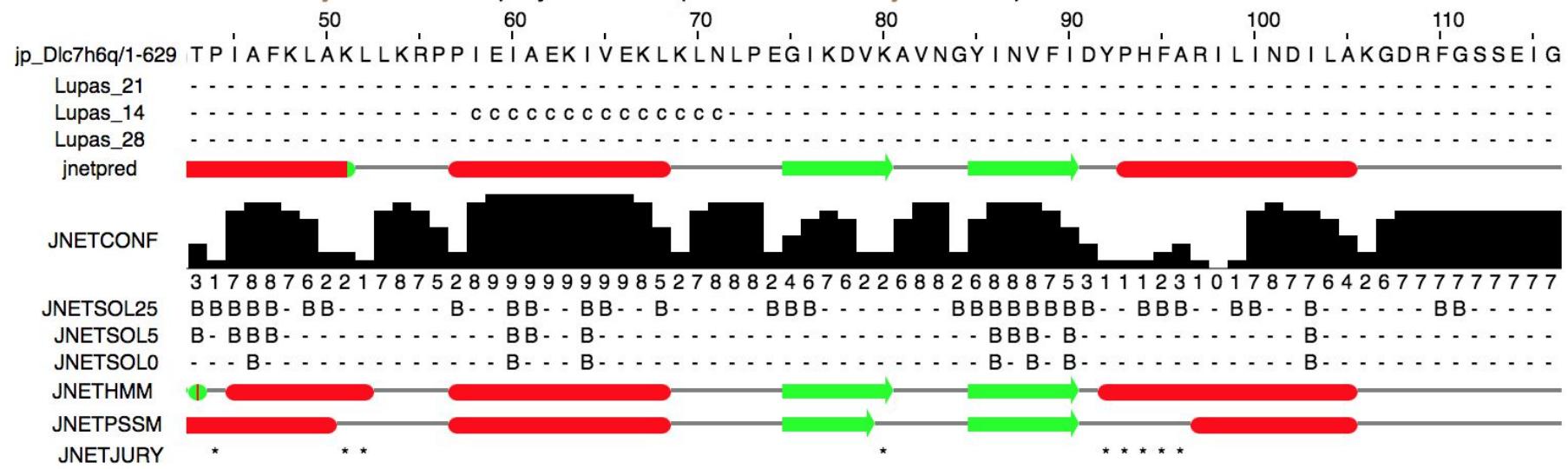
- Select a template structure – needs to share at least 30% sequence similarity with your protein, preferably 50%
- Align the sequences
- Generate geometric criteria and “fit” the new sequence into the structure
- Assess the model according to statistical potentials or physics-based energy calculations
- 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 – <https://iupred2a.elte.hu/>

>sp|059147|SYR_PYRHO Arginine--tRNA ligase





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?

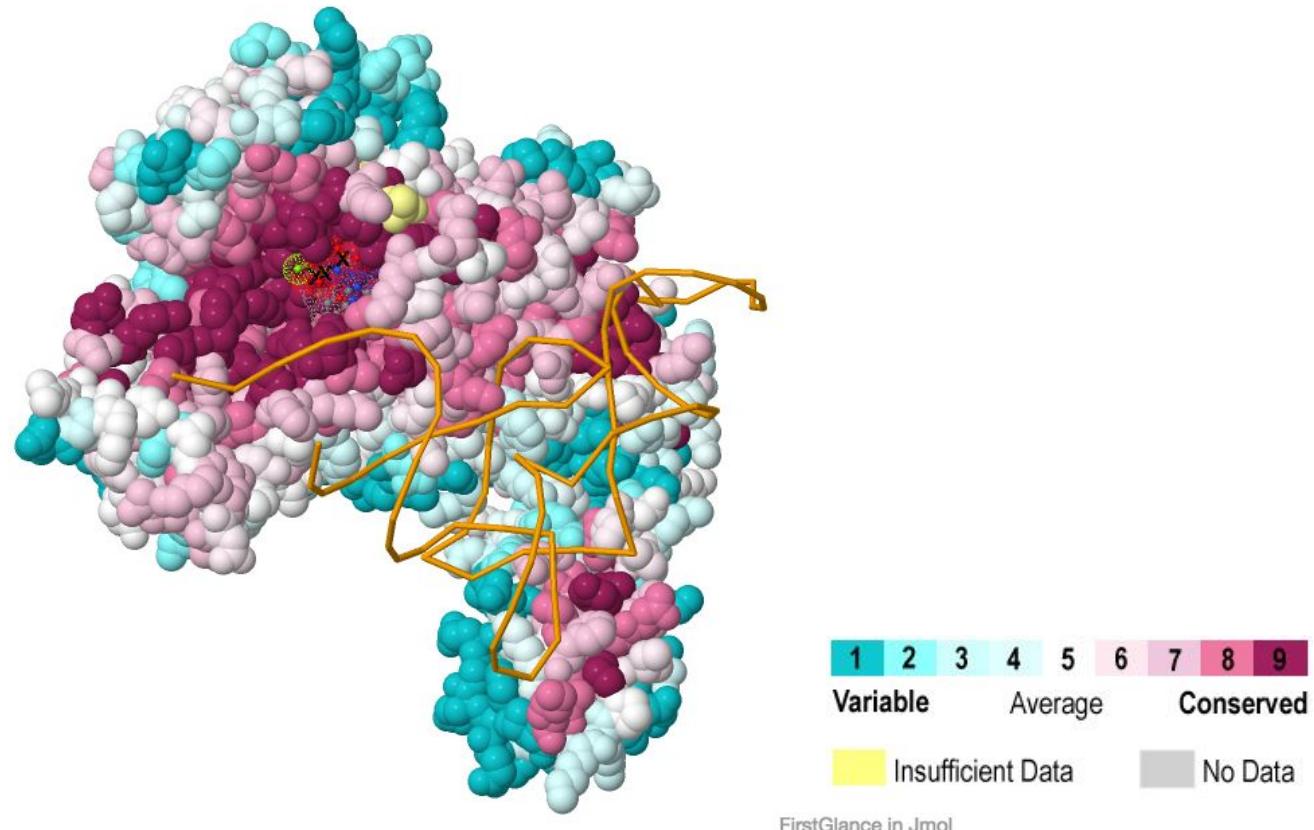
Interactions

- 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 modelling of biomolecular complexes

What can I do with a structure?

Function

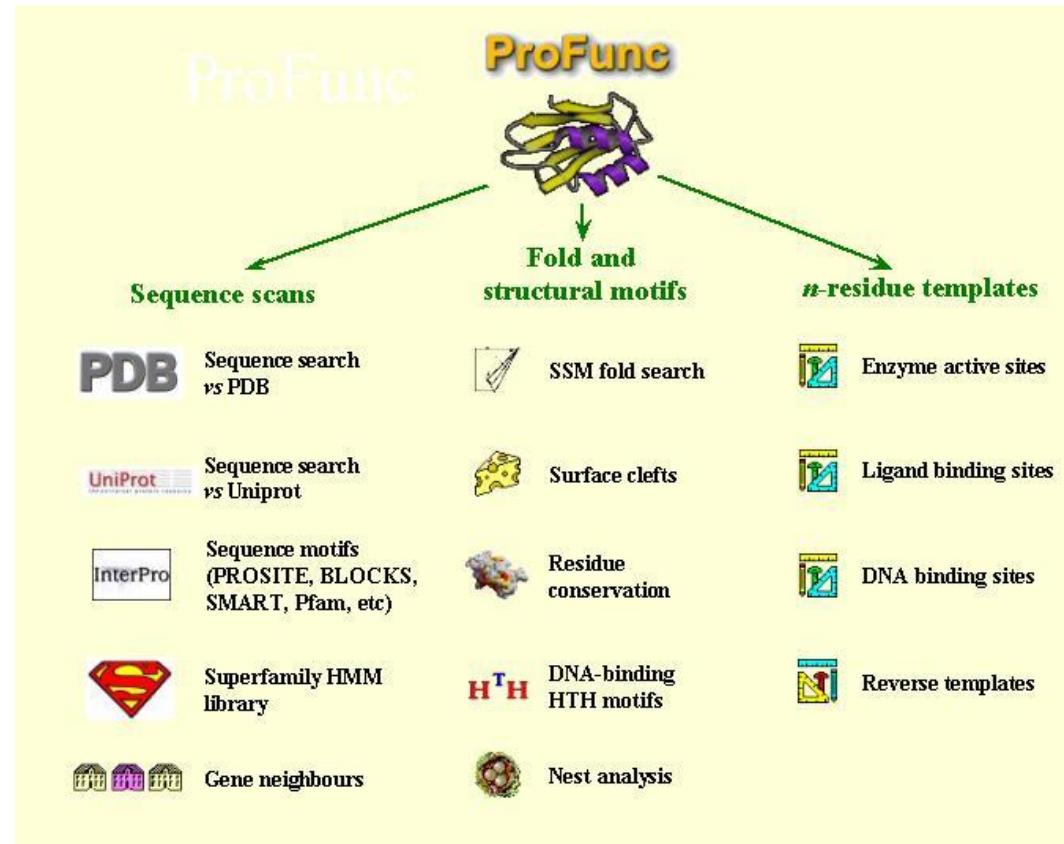
- ConSurf & ConSurf-DB
 - <http://consurf.tau.ac.il/> and <http://consurfdbs.tau.ac.il/>
 - Function based on evolutionary conservation - looking for evolutionarily conserved patches of amino acids in a 3D protein structure is a good way to locate functional sites



What can I do with a structure?

Function

- ProFunc -
<https://www.ebi.ac.uk/thornton-srv/databases/profunc>
L
 - 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/saab/dap/>
 - likely structural effects of a mutation
- Ensembl Variant Effect Predictor (VEP) -
<https://www.ensembl.org/info/docs/tools/vep/index.html>

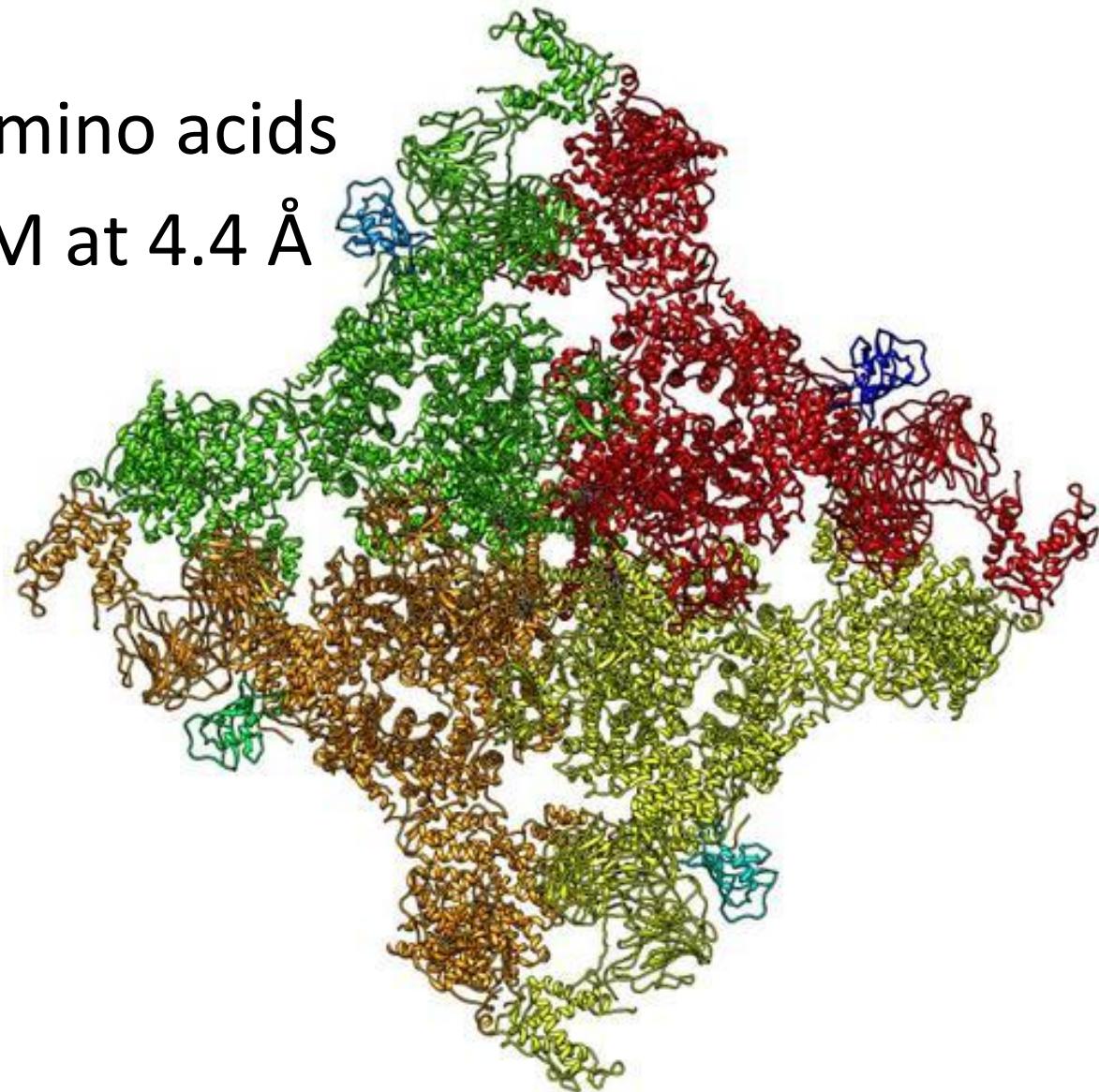
Machine Learning is changing everything...



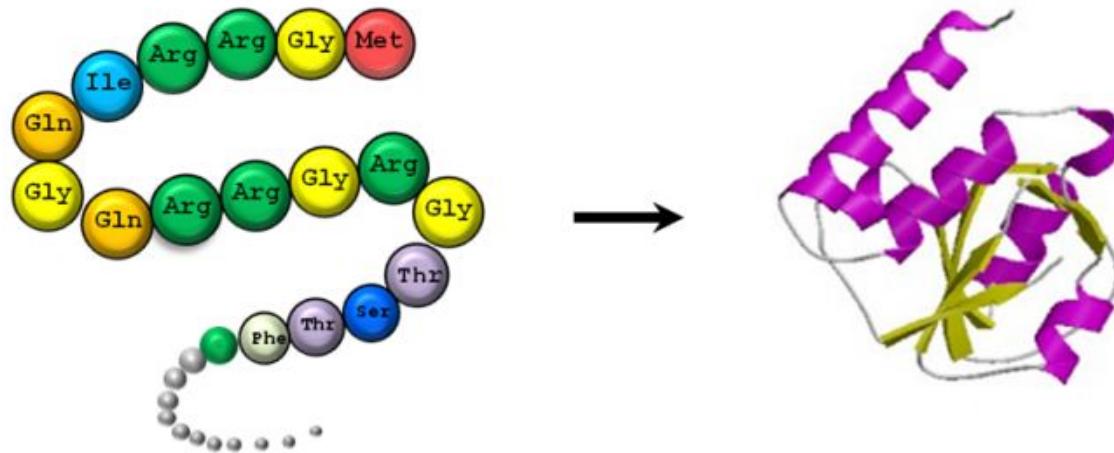
...but we won't go there today as it is still a wild wild west...

Intermission! Largest structure

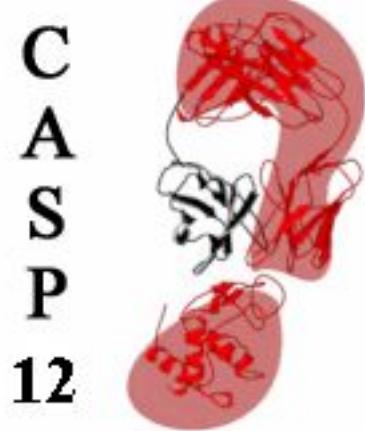
- 5TA3 - 39,852 amino acids
- RYR1 by Cryo-EM at 4.4 Å



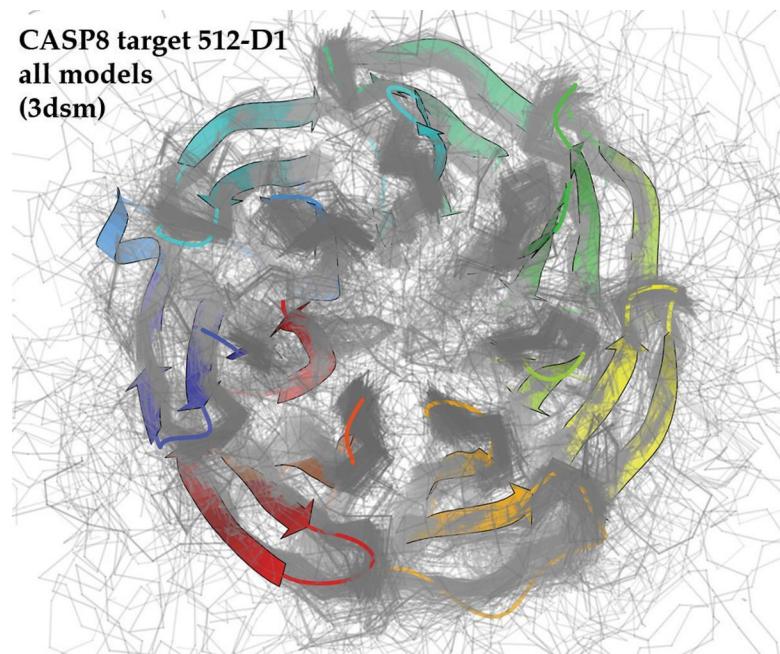
The Holy Grail of structural biology – structure prediction from sequence



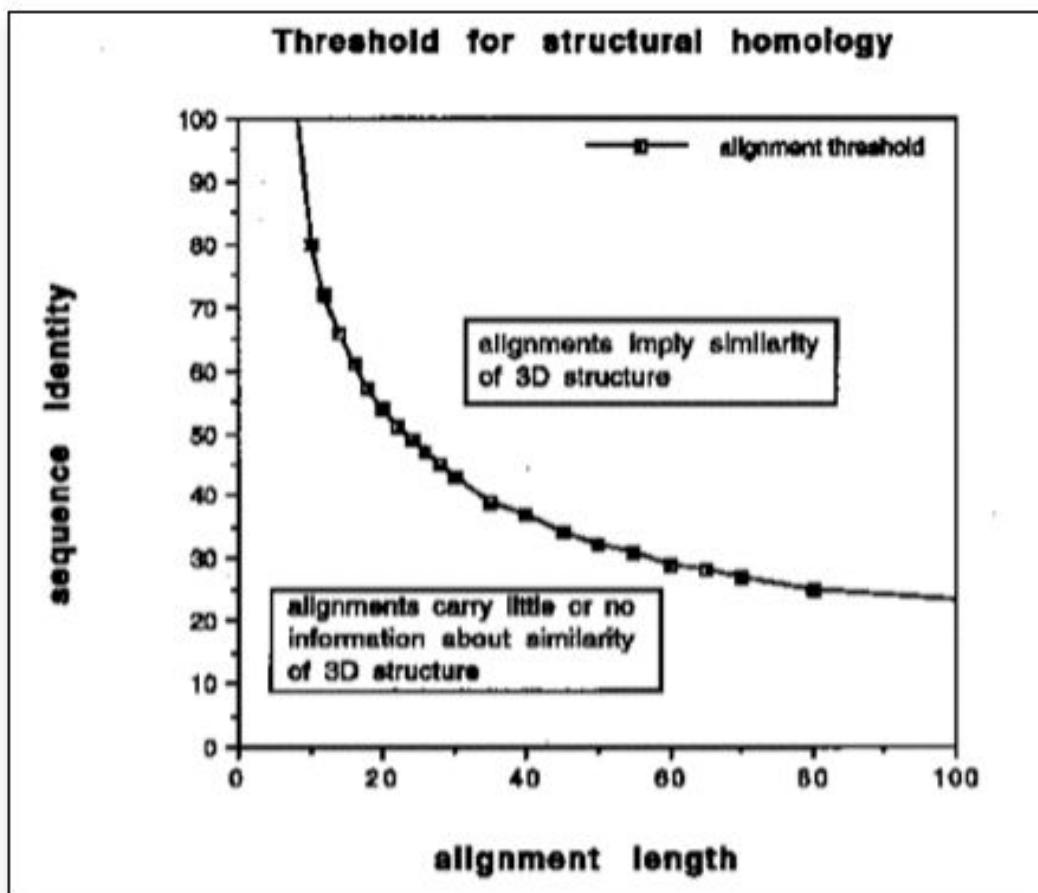
CASP – Critical Assessment of techniques for protein Structure Prediction

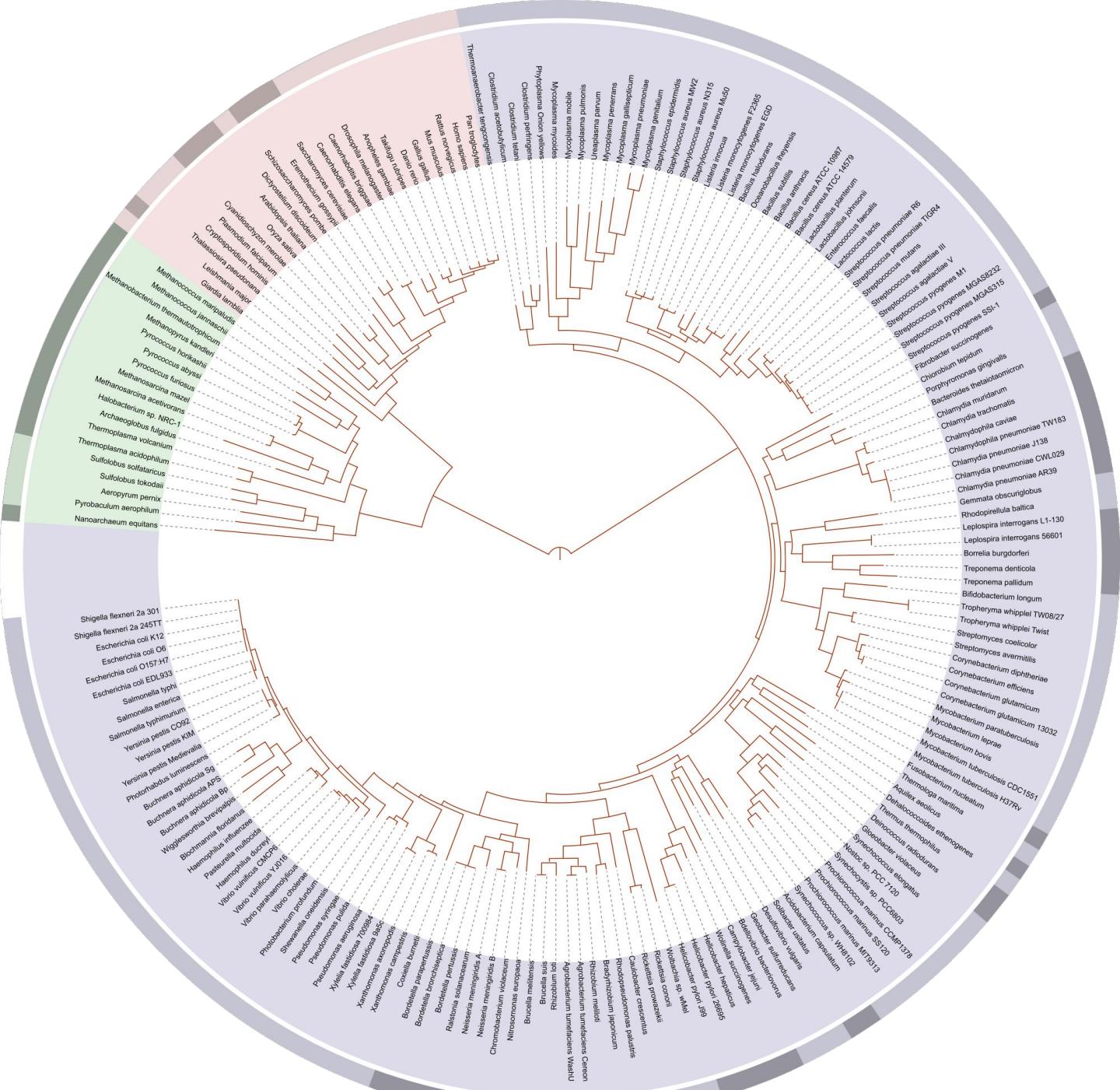


CASP8 target 512-D1
all models
(3dsm)

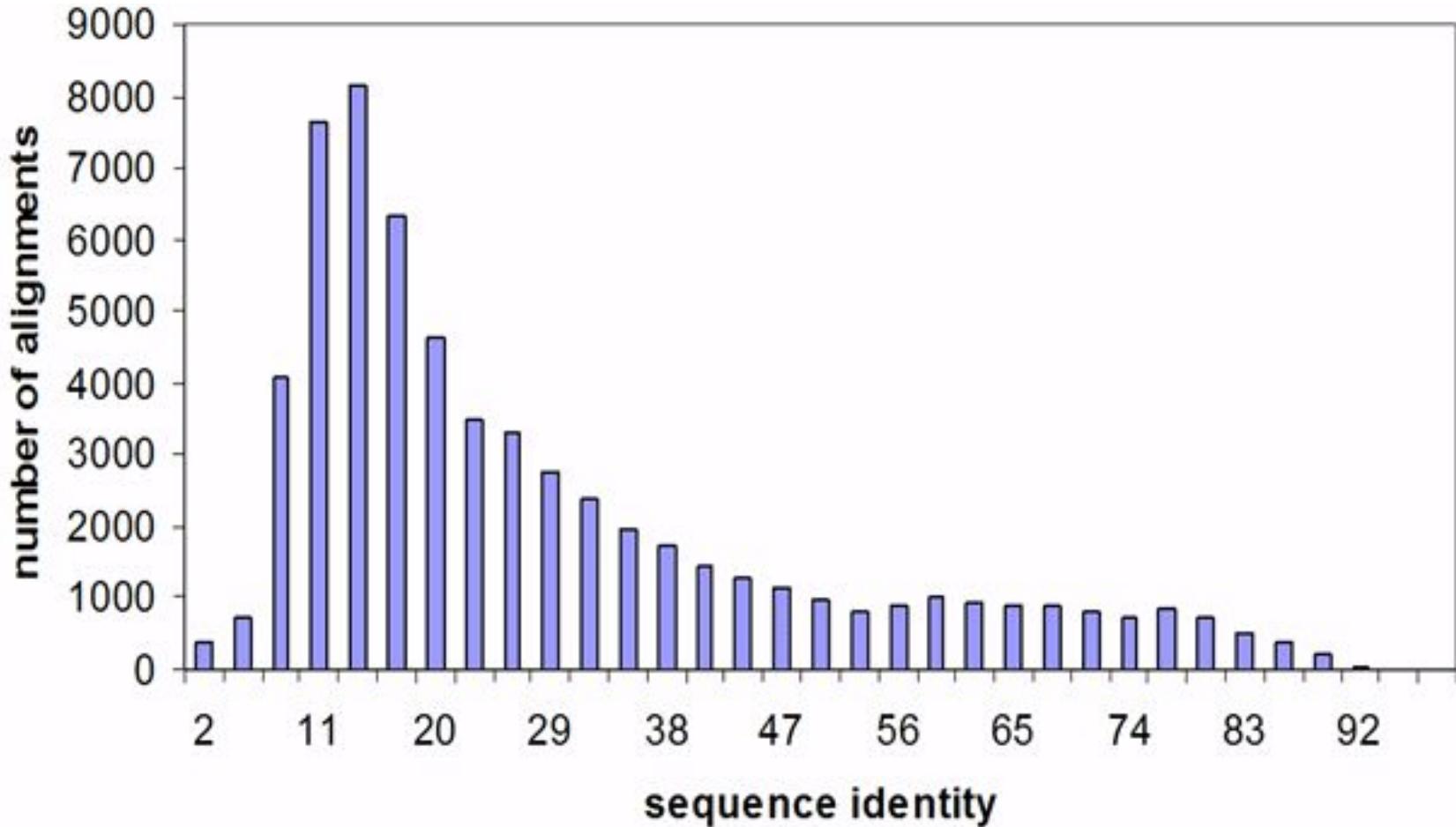


Structure is more conserved than sequence





Structure is more conserved than sequence



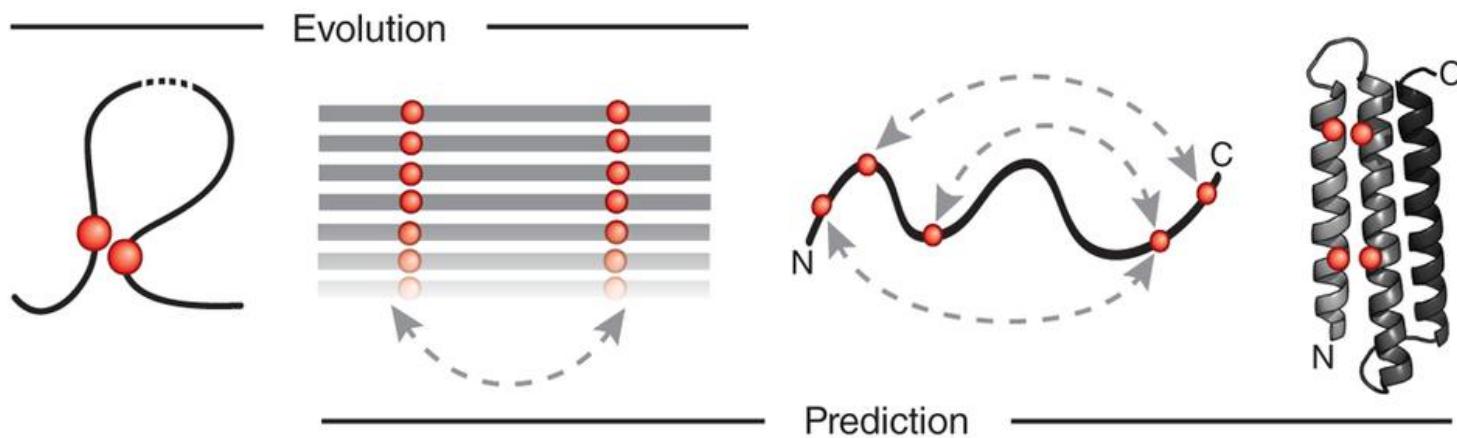
Hypothesis

- 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?

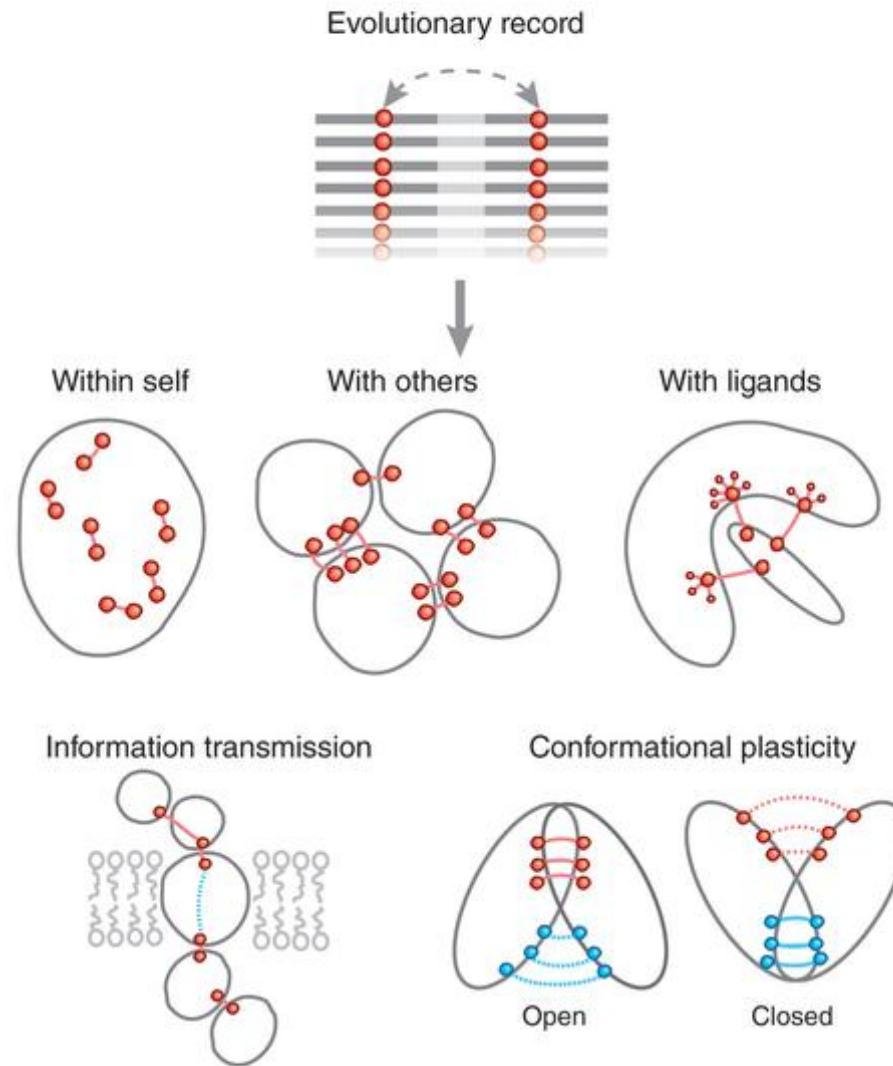
Coevolution



Co-evolutionary signal carries information

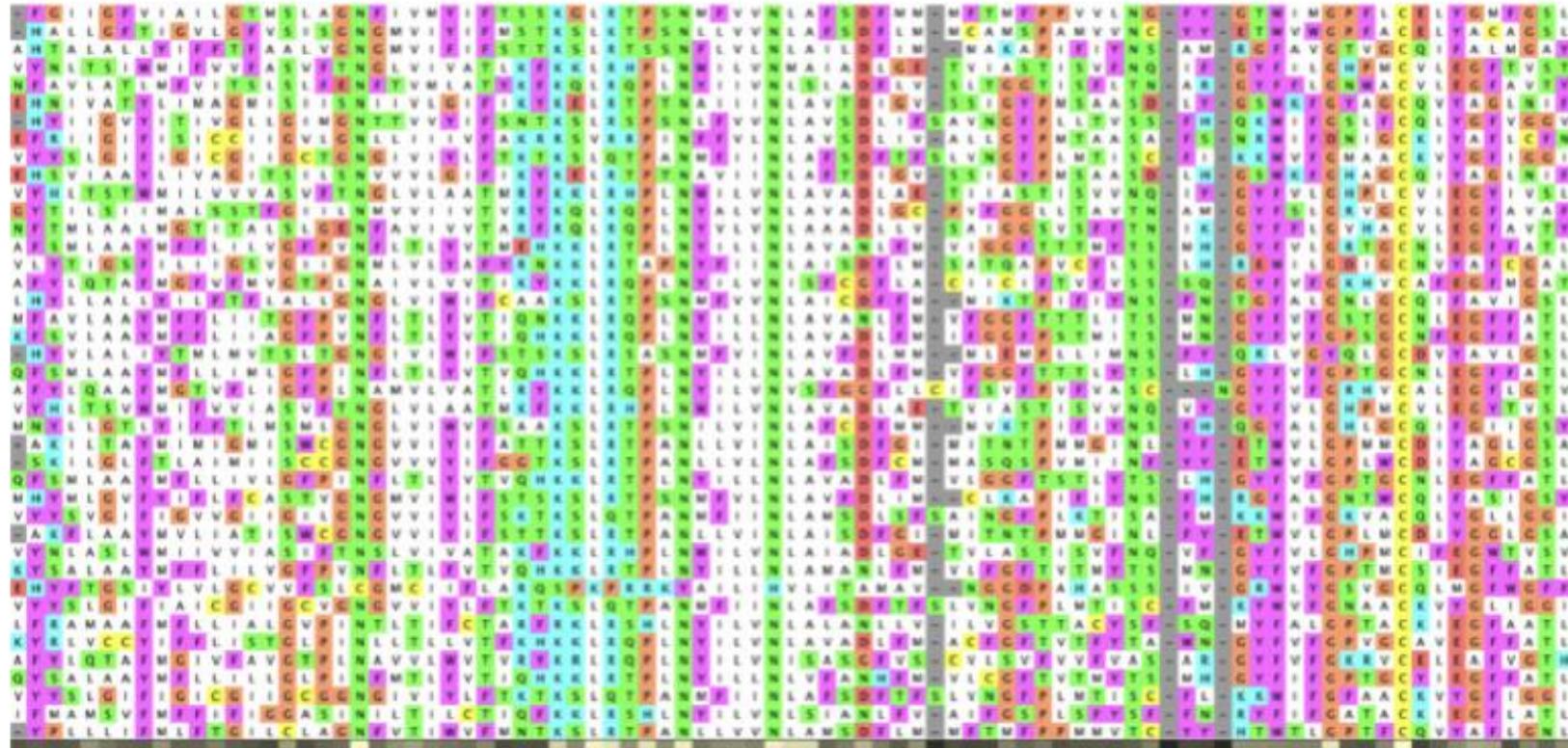


Co-evolutionary signal carries information

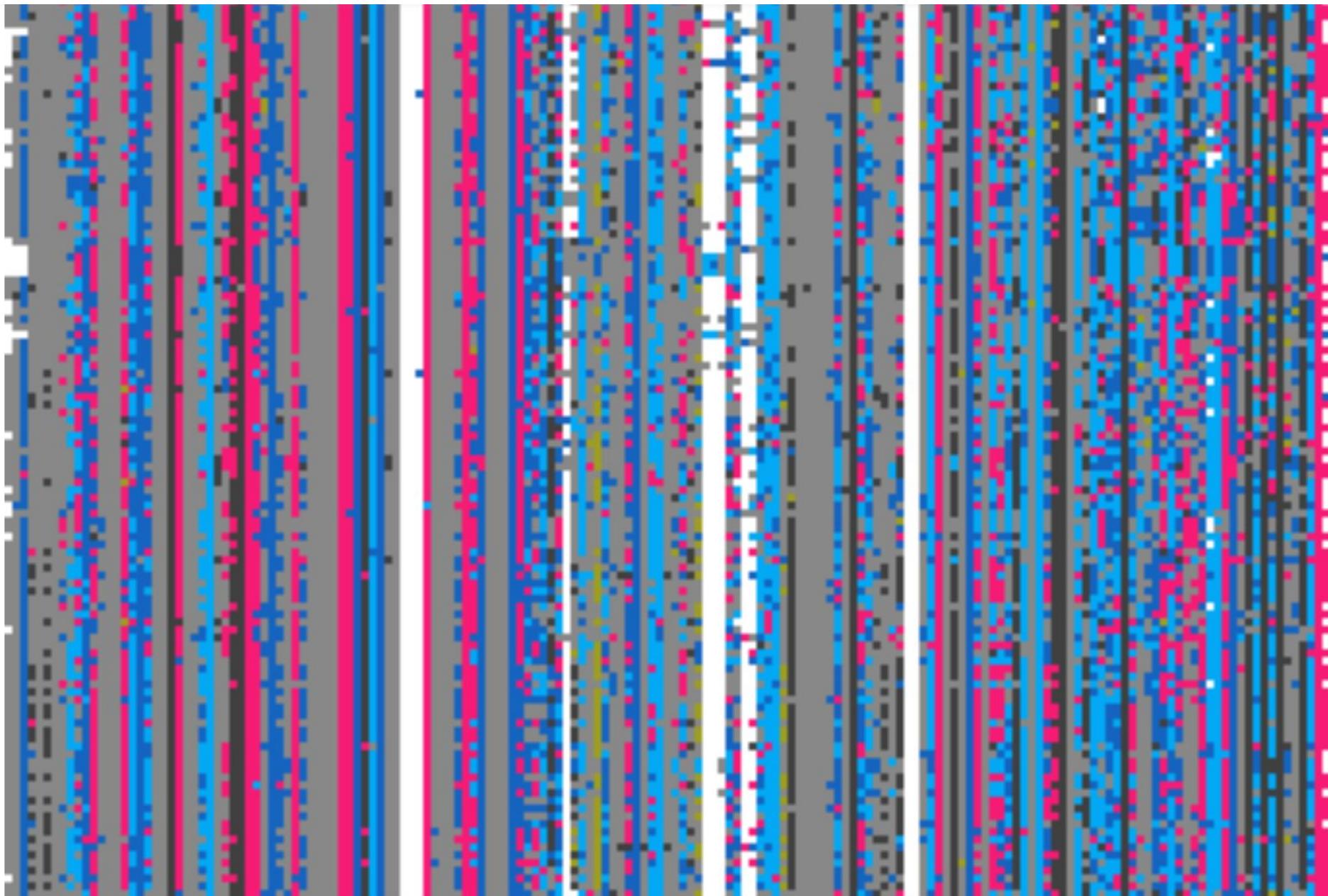


Thanks to cheap DNA sequencing!

1 Position in the alignment n

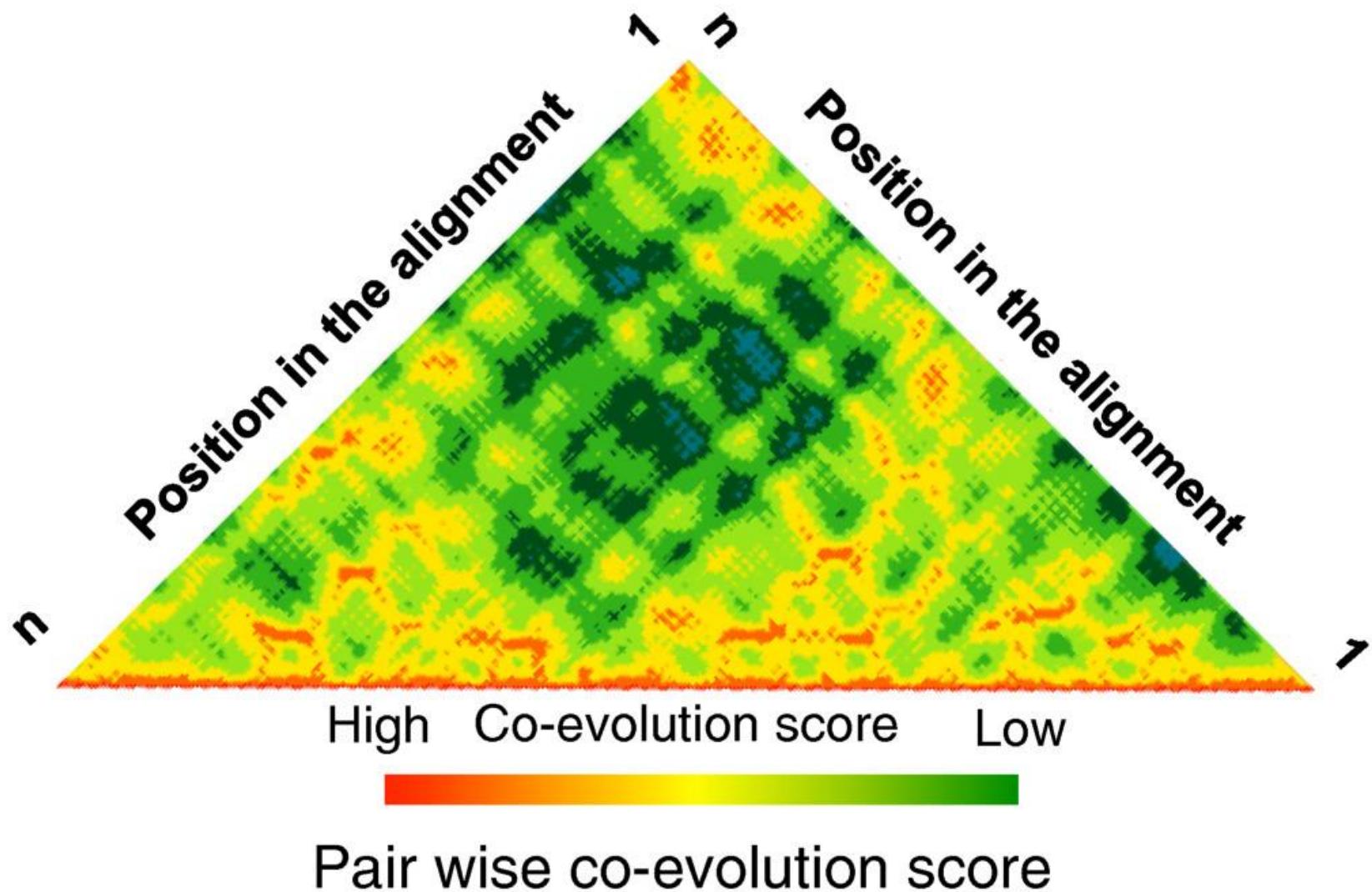


Multiple sequence alignment



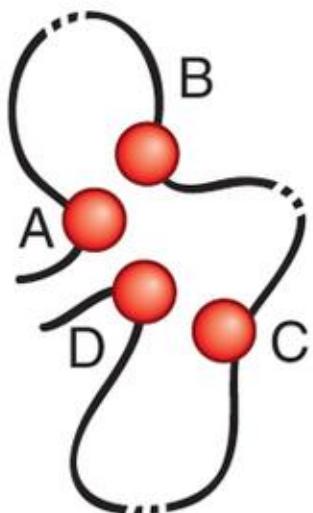


How to score co-evolution?

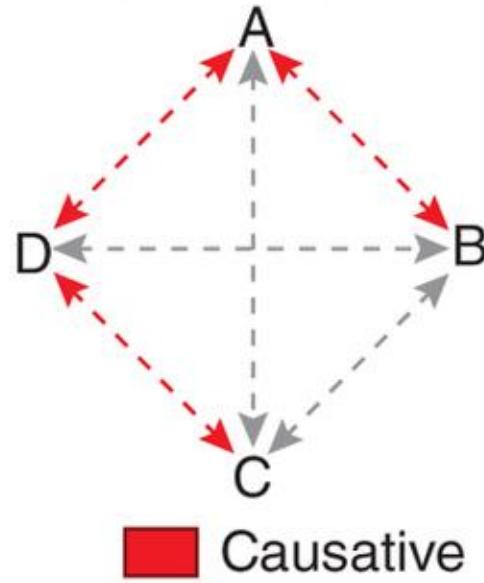


Causative vs transitive correlations (inverse Ising problem)

Physical contacts



Observed correlations



Predicted contacts

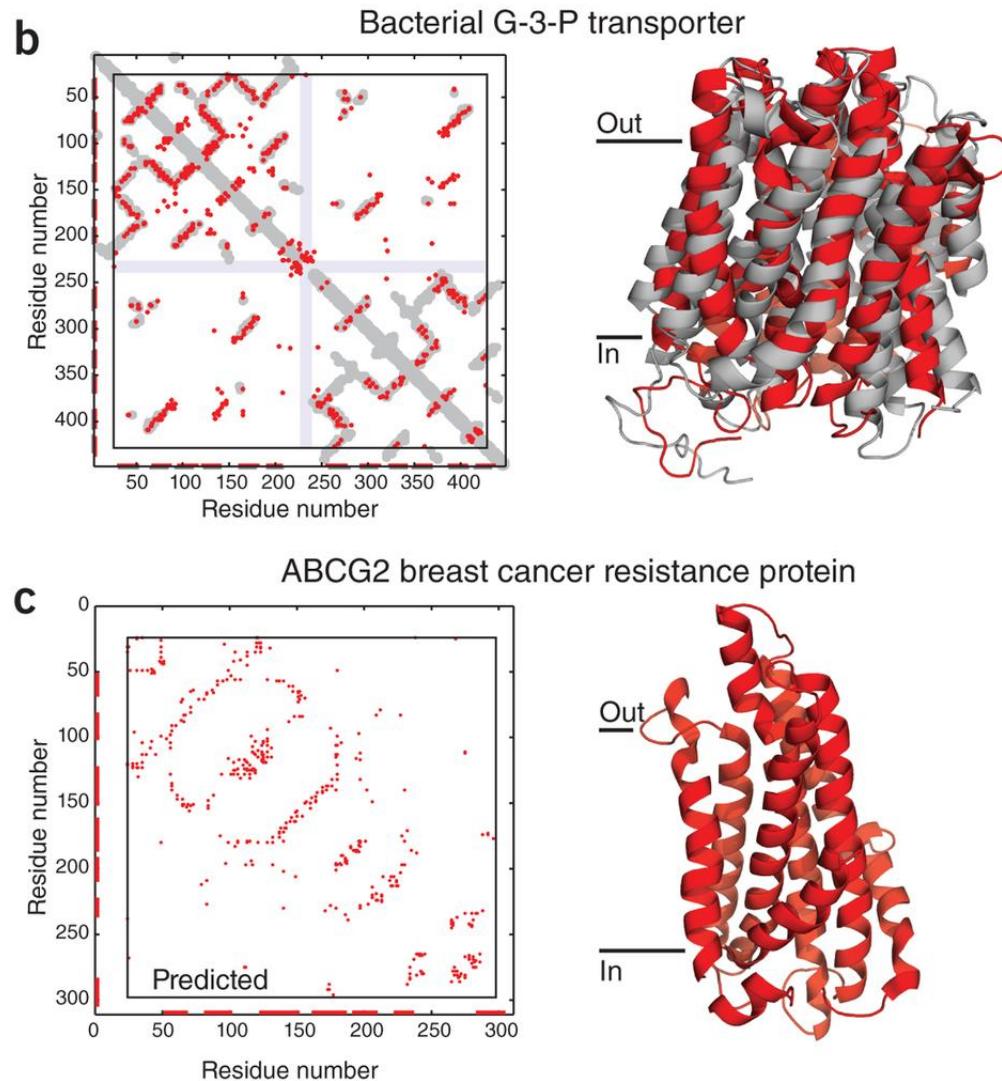
	A	B	C	D
A				
B				
C				
D				

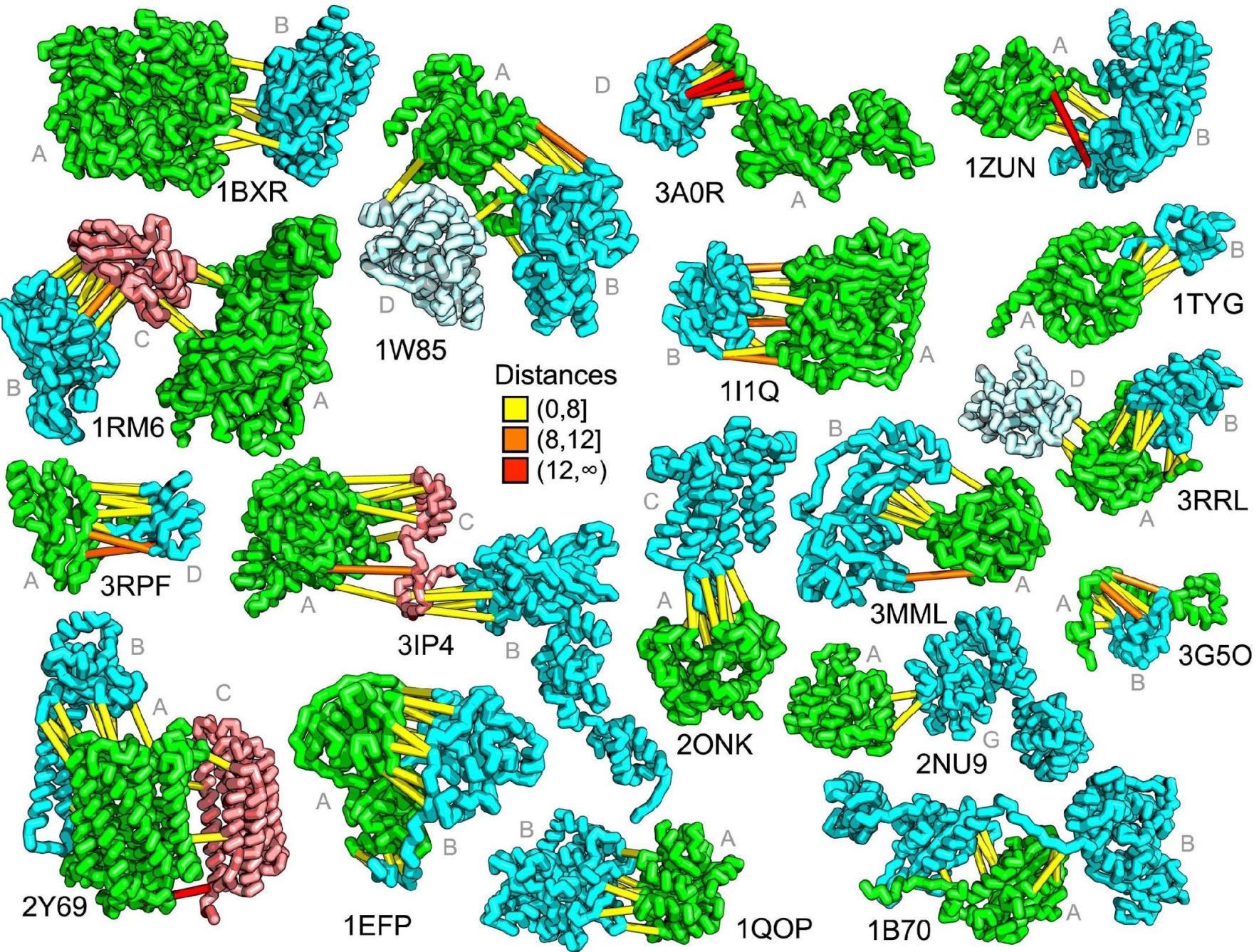
Legend: Red square = Causative; Gray square = Transitive

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)
 - Gaussian approximation (gDCA)
- Machine Learning based approaches
 - CNNs, ResNets, AlphaFold

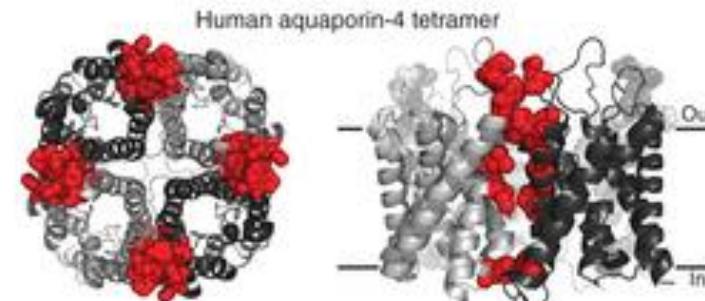
Early success story: membrane protein structure prediction



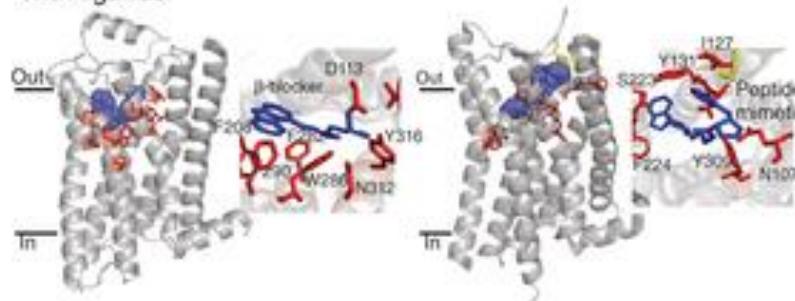


Evolutionary couplings reveal function

With others



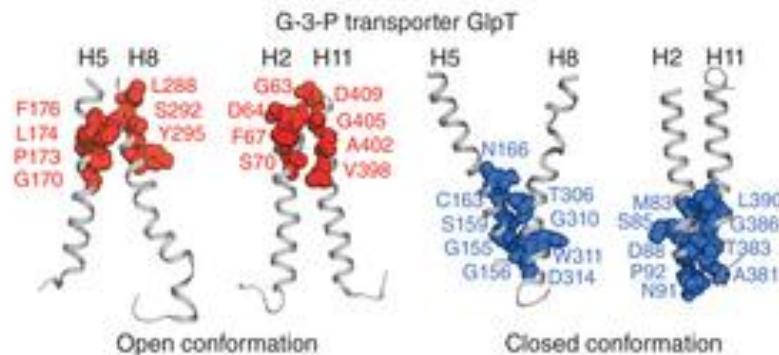
With ligands



Human $\beta 2$ adrenergic receptor

Human nociception receptor

Conformational plasticity



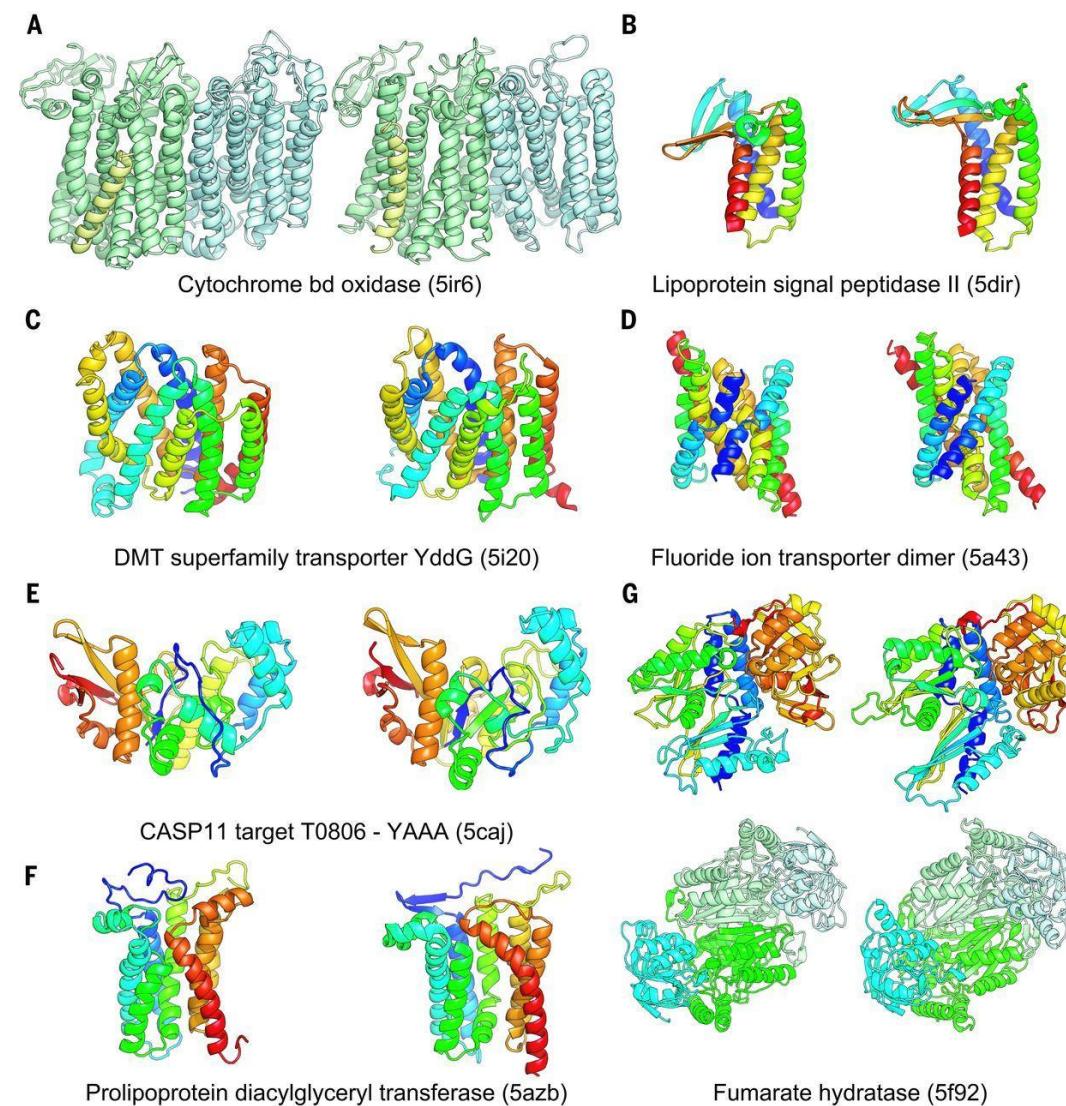
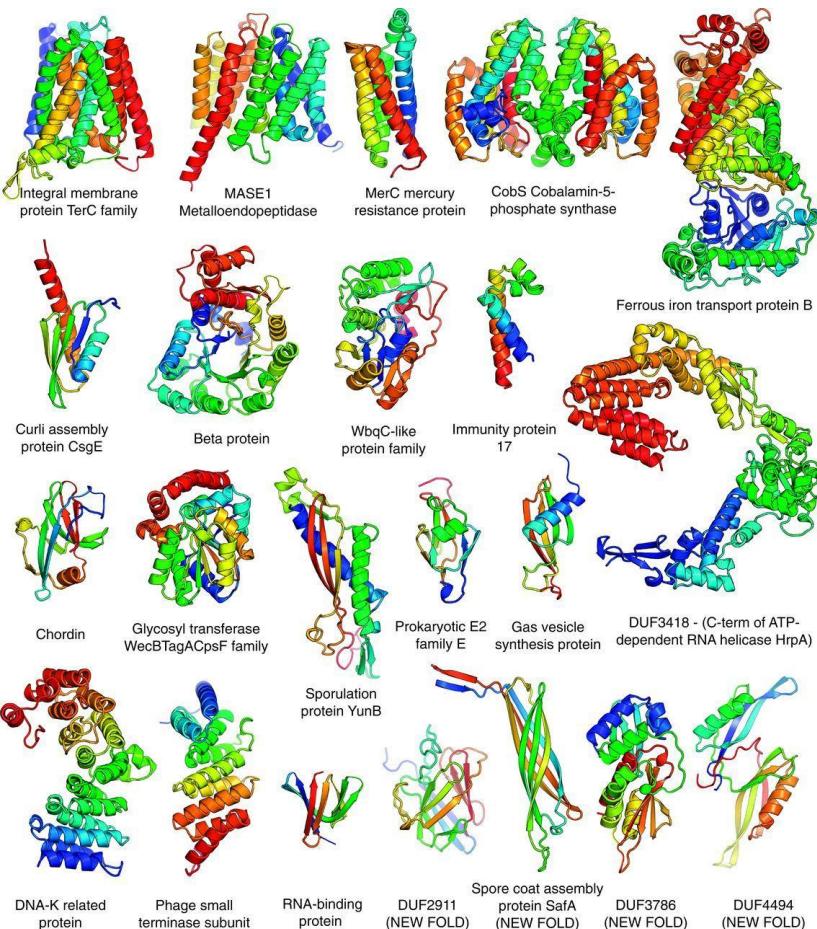
Protein structure determination using metagenome sequence data

Sergey Ovchinnikov^{1,2,3}, Hahnbeom Park^{1,2}, Neha Varghese⁴, Po-Ssu Huang^{1,2}, Georgios A. Pavlopoulos⁴, David E. Kim^{1,5}, Hetunandan Kamisetty⁶, Nikos C. Kyrpides^{4,7}, David Baker^{1,2,5,*}

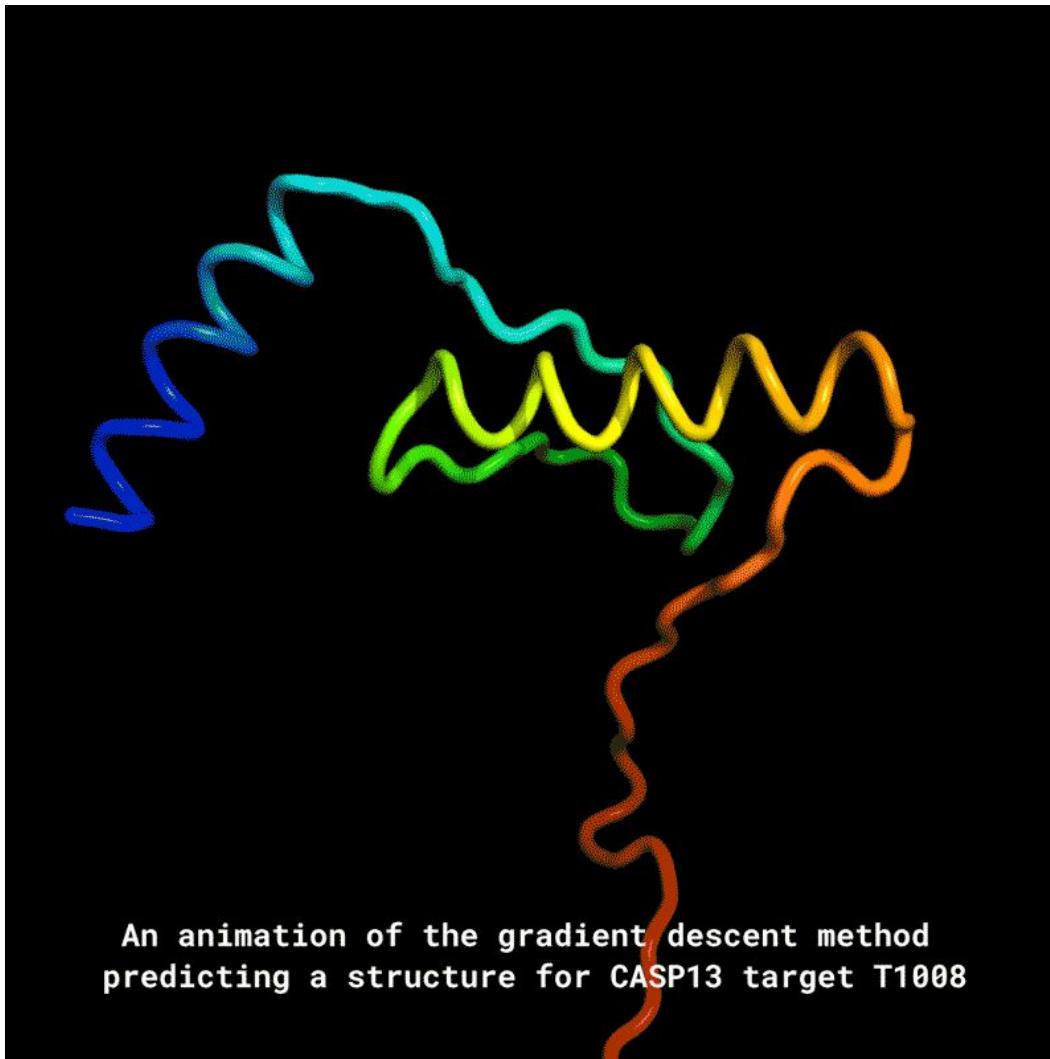
+ Author Affiliations

✉ Corresponding author. Email: dabaker@u.washington.edu

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

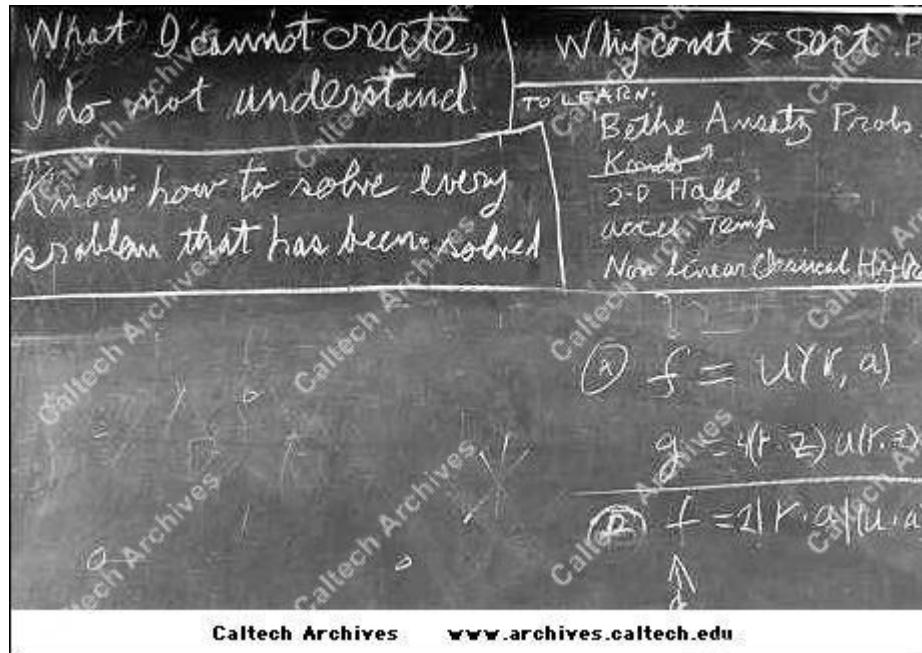


AlphaFold



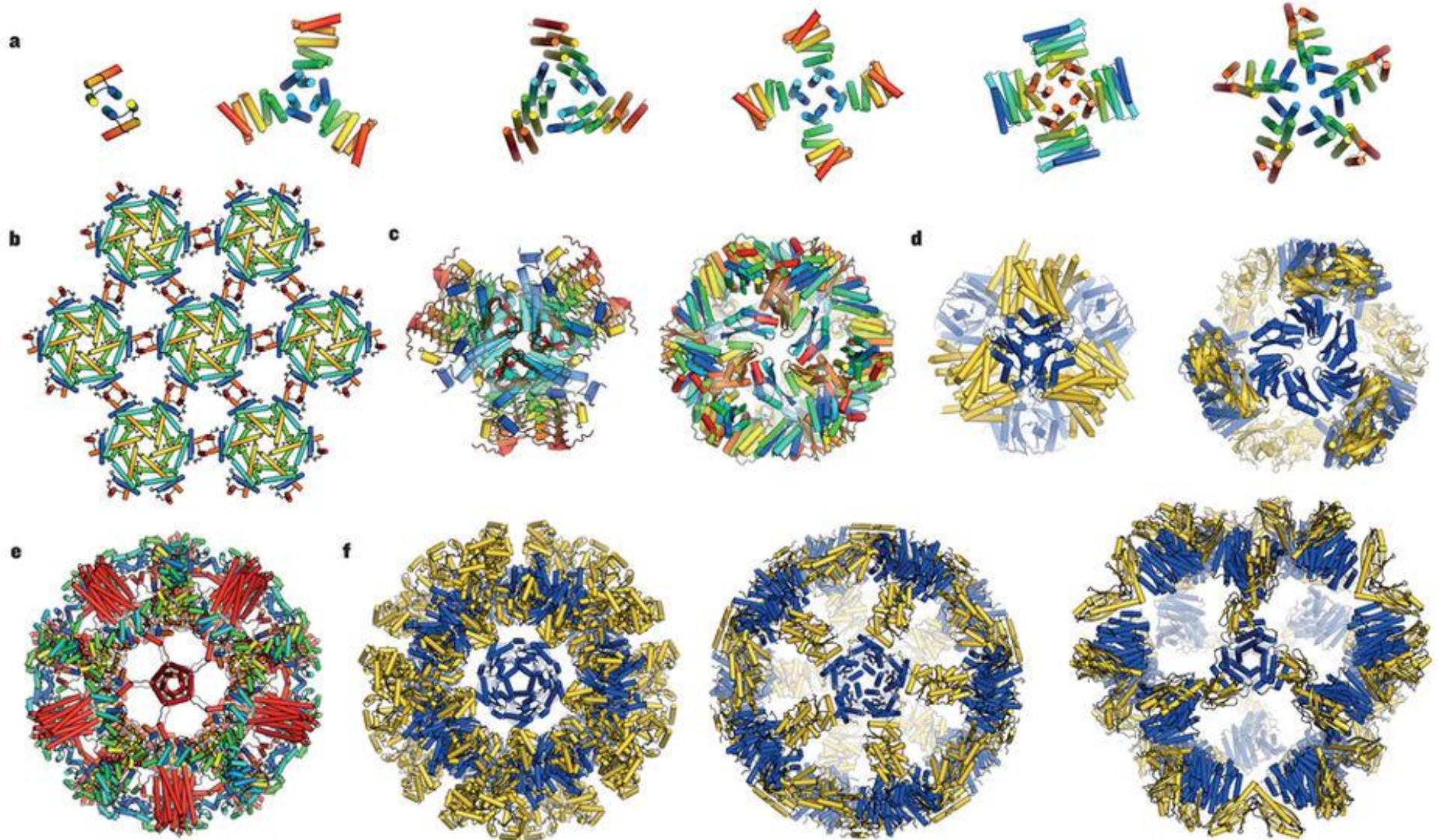
An animation of the gradient descent method
predicting a structure for CASP13 target T1008

Richard Feynman's famous “last words”



“What I cannot create, I do not understand.”

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

Extra reading for “novelty/cool” research directions:

- Coevolution: <https://www.nature.com/articles/nbt.2419>
- De novo design: <https://www.nature.com/articles/nature19946>
- Structure prediction & design:
<https://www.nature.com/articles/s41580-019-0163-x>

Questions for the supervisions

- A structural model on its own is not very informative. Discuss this statement with reference to a variety of structural bioinformatics databases and resources.
- How does the experimental pipeline or choice of technique for structure determination impact the quality and validity of the structural model?
- *(2018 exam)* Protein structure models are only as good as the methods used to generate them. Discuss this statement in relation to experimental techniques, homology modelling and *de novo* structure prediction.
- *(2019 exam)* For the following questions assume you can only rely on bioinformatics tools and no new experiments.
 - Given the amino acid sequence of a protein, describe the steps you would take and the tools you would use to learn about its secondary and tertiary molecular structure. [60%]
 - Subsequently, a structural model of the same protein is published. Describe how you would first check if the model is reliable, and then how you would compare it to your predictions from a). Describe the steps you would take and the tools you would use. [40%]

Lecture slides and practical material

github.com/alexeymorgunov/structbioinfo

Thank you!

Have any questions, comments?

Email me: **alexey@manifoldresearch.com**