PROTEIN STRUCTURE: INSTRUCTIONS FOR USE

Luciana Esposito

Istituto di Biostrutture e Bioimmagini
Consiglio Nazionale delle Ricerche
Via Mezzocannone 16, Napoli
E-mail: luciana.esposito@cnr.it

Naples, Italy
29 Aug – 3 Sep 2019

Outline

- Sequence – Structure – Function Paradigm
- Important information in a PDB entry that can be relevant for a correct use of a protein structure
- The relevance of the knowledge of a 3D structure in Drug Design, Molecular Docking, Molecular Dynamics simulations
- From 3D structure to function: the complement of sequence to function relationships
- Brief outline of Integrative Structural Biology

The centrality of Structural Biology

Biology - biomolecules

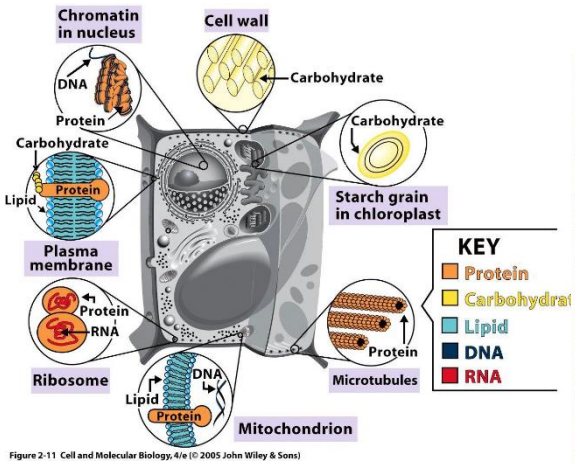
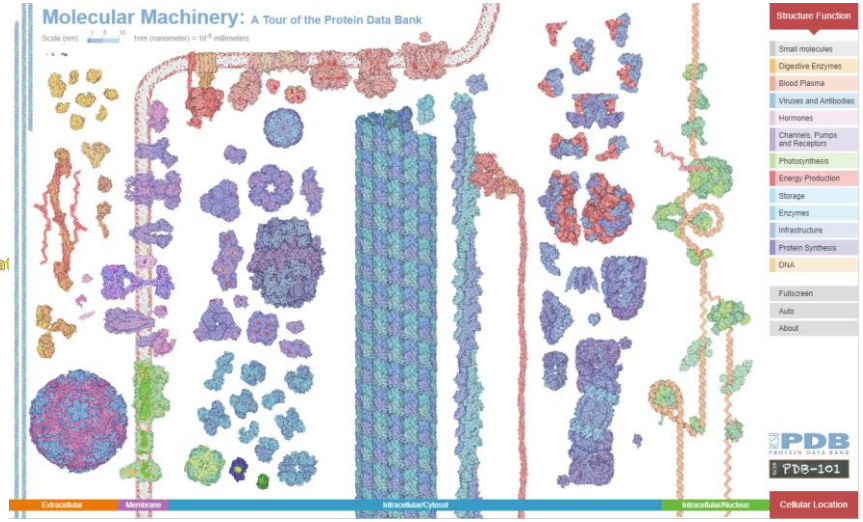


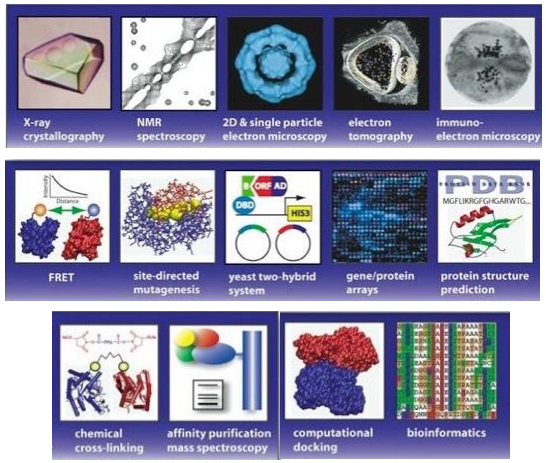
Figure 2-11 Cell and Molecular Biology, 4/e (© 2005 John Wiley & Sons)

Structural Biology

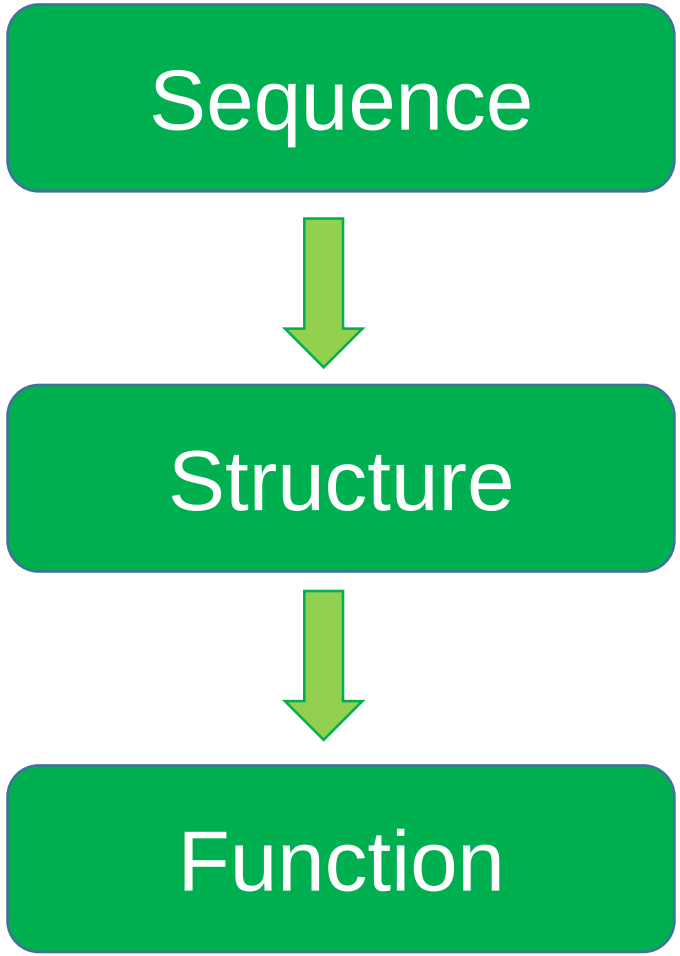


Structural Biology aims to understand how biology works at the molecular level.
Structural biology is the study of the molecular structure and dynamics of biological macromolecules, and how alterations in their structures affect their function.

Integrative Structural Biology



Paradigm



The aminoacid sequence encodes the structure

The structure determines the function

The Structure – Function relationship

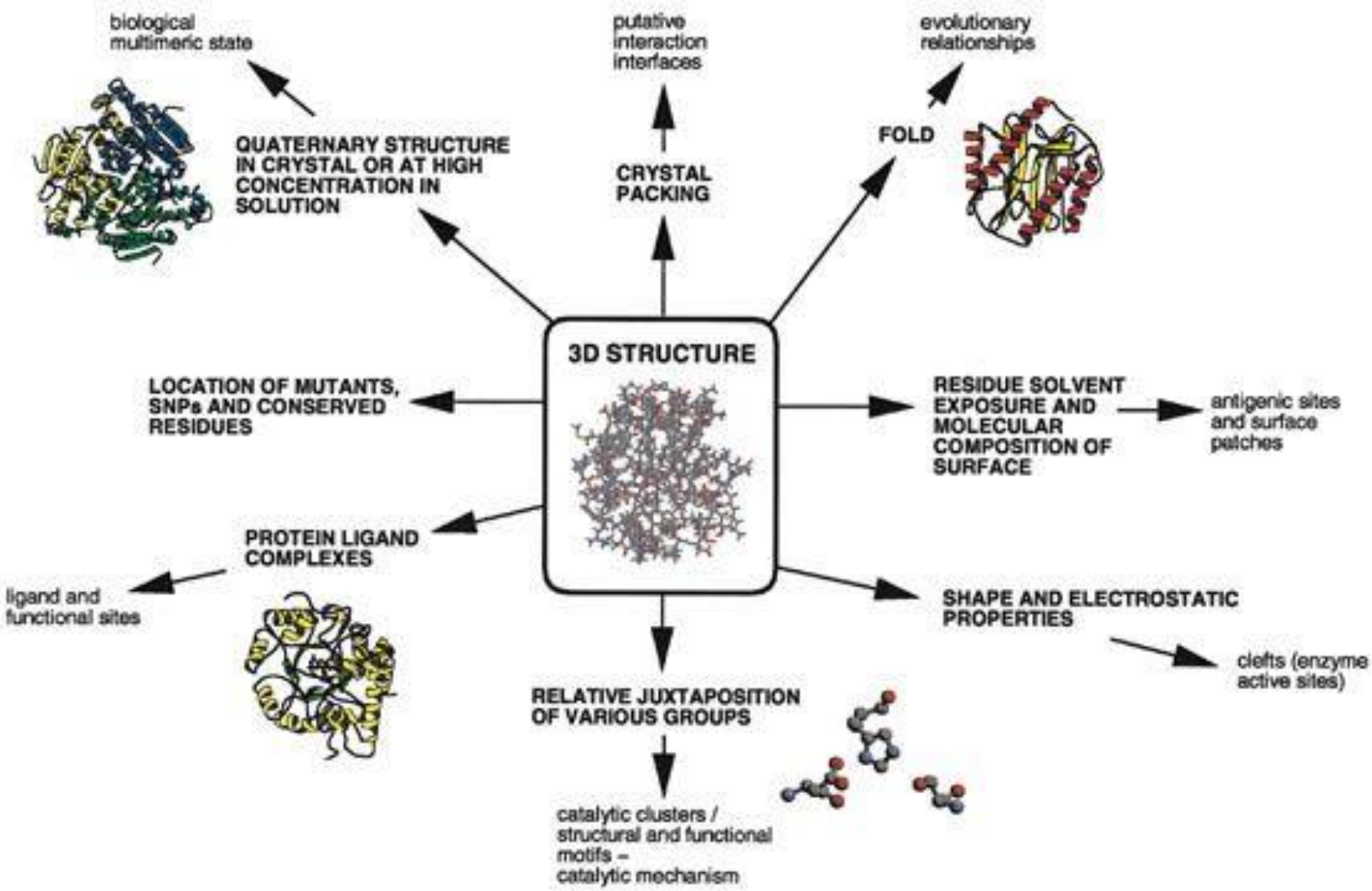
Function → Structure

- **The classical way**
  - A function is discovered and studied
  - The gene responsible for the function is identified
  - Product of this gene is isolated, crystallized and the structure solved
  - The structure is used to "rationalize" the function and provide molecular details

Structure → Function

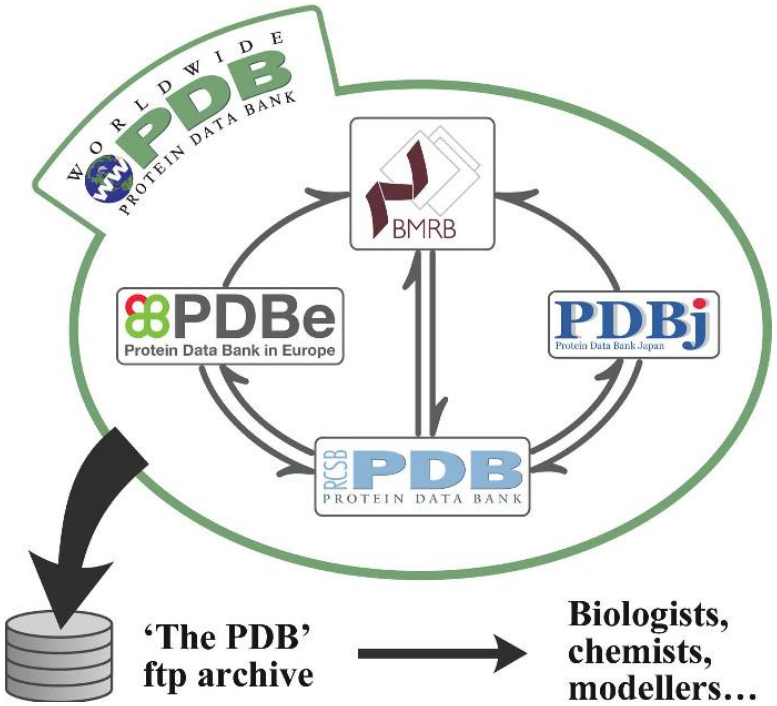
- **Post-genomic**
  - A new, uncharacterized gene is found in a genome
  - Predictions or high-throughput methods select this gene for further studies
  - The protein is expressed and has to be studied in detail
  - The structure is solved and can be the first experimental information about the "hypothetical" protein whose function is unknown

Summary of Information Derived from 3D-structure



Thornton JM et al., *Nat. Struct. Biol.* 2000

wwwPDB



Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

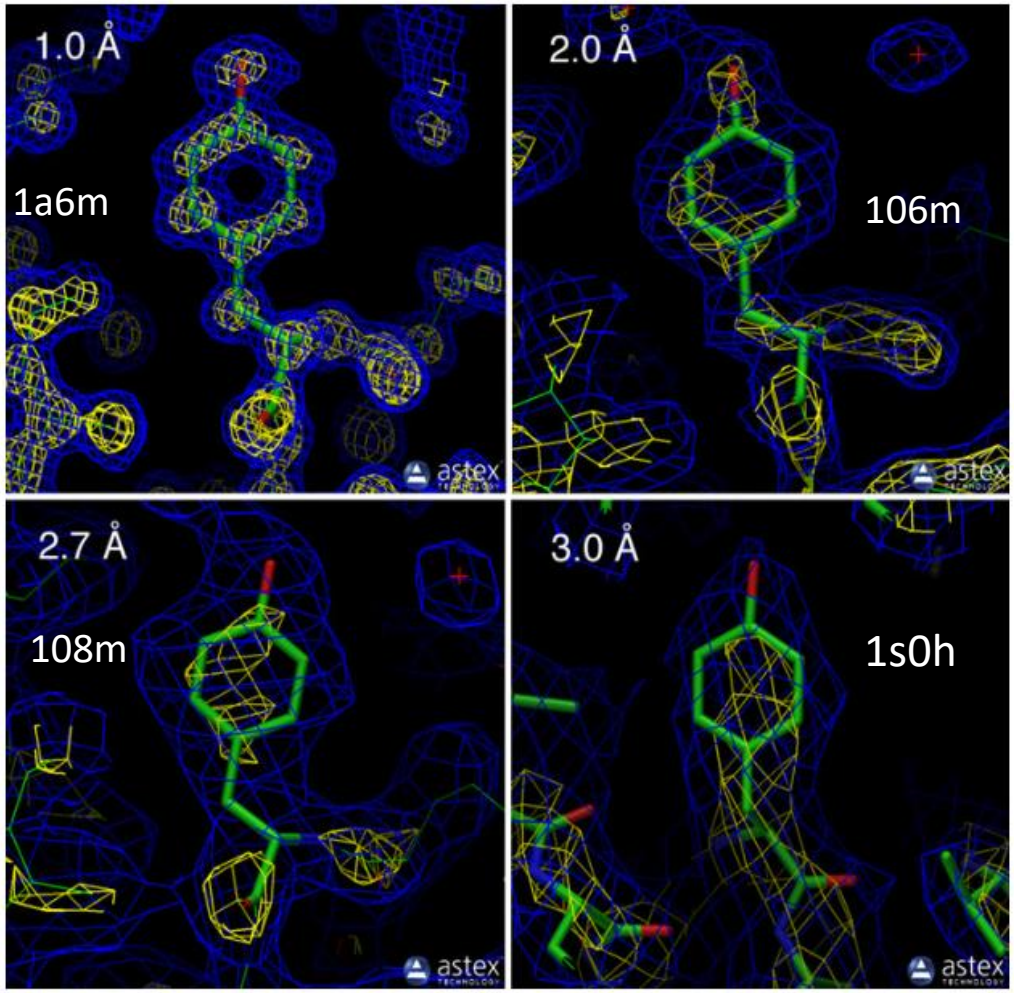
✓ Let's analyze important information that can be retrieved from a PDB entry and that have to be consider when using the structure to address functional questions and/or expand the understanding of the protein system by means of other tools:

- Drug design
- Docking
- MD simulations
- …

Resolution

```
REMARK    2
REMARK    2 RESOLUTION.    1.65 ANGSTROMS.
```

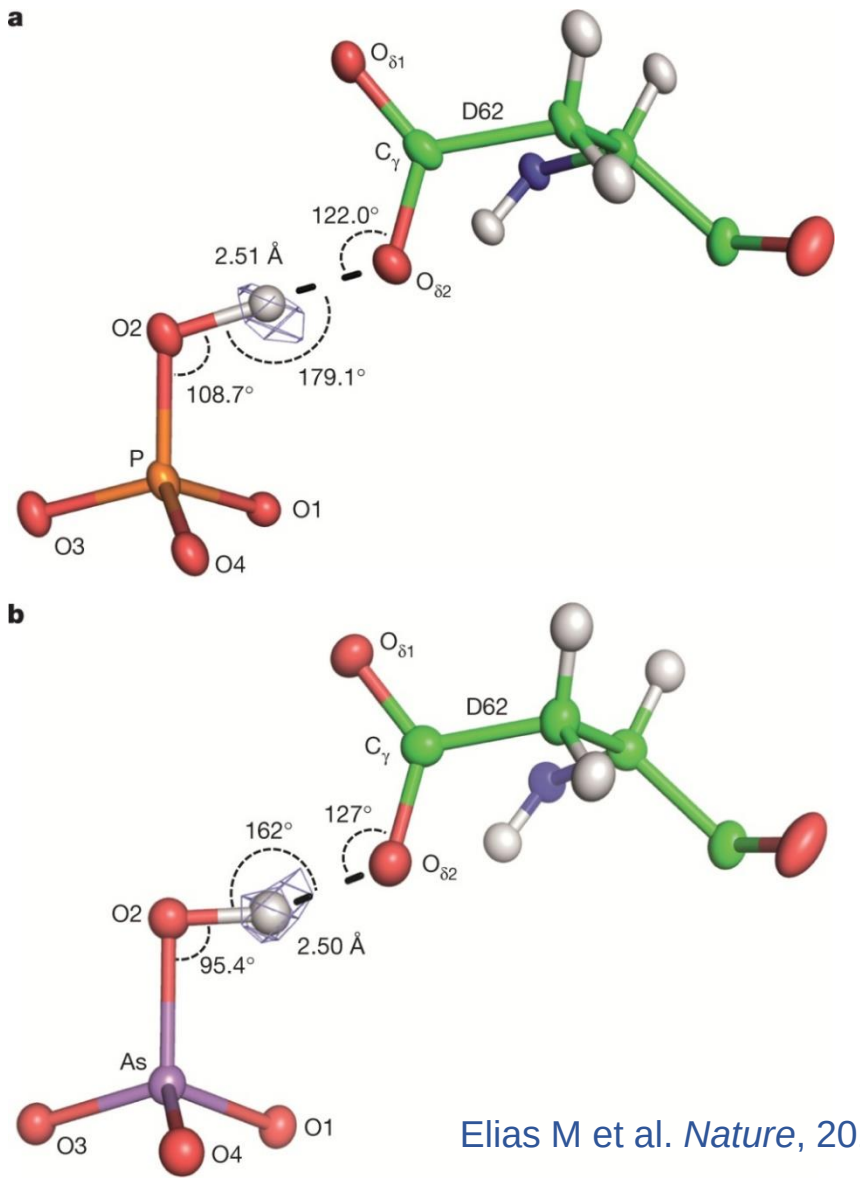
mmCIF **Category name** refine
Item name _refine.ls_d_res_high



Resolution is a measure of the level of detail present in the diffraction pattern and the level of detail that will be seen in the electron density map.

RCSB PDB-101
Educational portal of RCSB PDB

The need of Ultra-high Resolution
→ very subtle details



Elias M et al. *Nature*, 2012

The molecular basis of phosphate discrimination in arsenate-rich environments

Periplasmic **phosphate-binding protein** PBP: arsenate-bound and phosphate-bound structures determined at 0.96 Å and 0.88 Å resolution.

The low-barrier Hydrogen Bond (negative-charge-assisted HB) angles are optimal in the phosphate-bound structure but distorted with arsenate. This is the consequence of the longer As-O2 bond than the P-O2 bond.

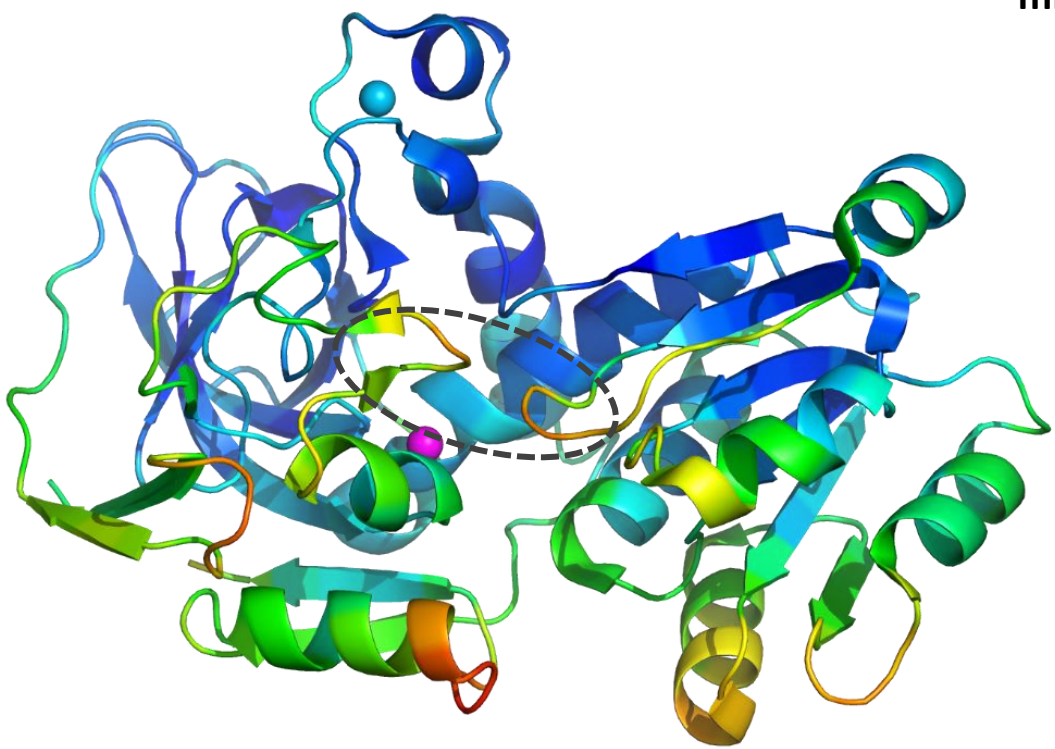
→ Anion selectivity (at least 10^3 excess)
The energy of this short bond is channelled not towards anion-binding affinity but rather towards anion selectivity

Atomic displacement parameters (B-factors)

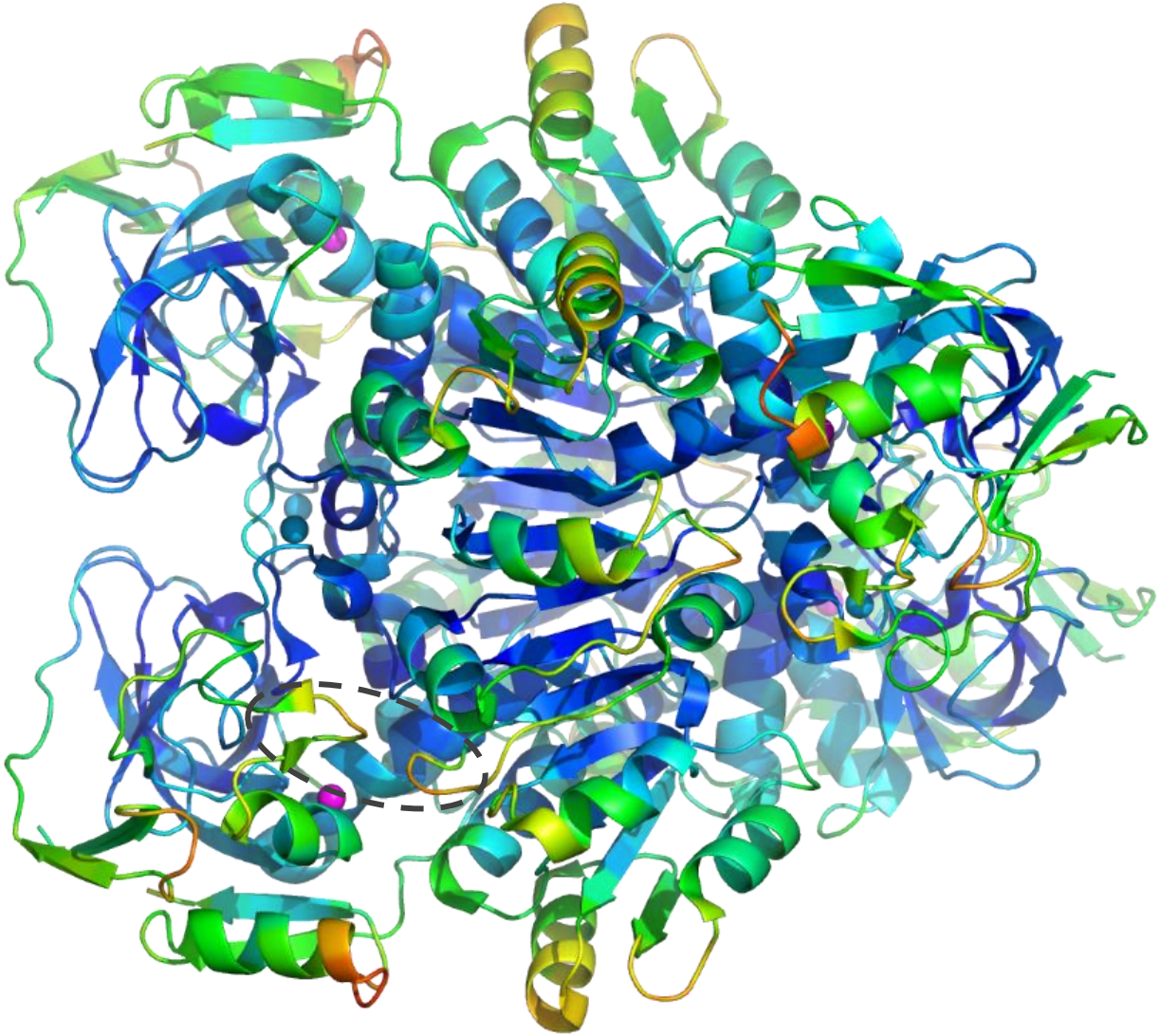
						x,y,z coordinates			Occ.	B-fact	
ATOM	1	N	MET	A	1	67.994	-36.765	49.452	1.00	41.39	N
ATOM	2	CA	MET	A	1	66.958	-35.932	50.131	1.00	41.99	C
ATOM	3	C	MET	A	1	65.607	-36.625	50.045	1.00	41.97	C
ATOM	4	O	MET	A	1	65.386	-37.463	49.173	1.00	41.53	O
ATOM	5	CB	MET	A	1	66.850	-34.562	49.447	1.00	41.54	C
ATOM	6	CG	MET	A	1	66.192	-34.614	48.064	1.00	41.11	C
ATOM	7	SD	MET	A	1	65.929	-32.957	47.297	1.00	40.03	S
ATOM	8	CE	MET	A	1	65.389	-33.399	45.649	1.00	34.90	C

$$B = 8\pi^2 \overline{u^2}$$

mmCIF **Category name** atom_site
Item name _atom_site.B_iso_or_equiv



Atomic displacement parameters (B-factors)

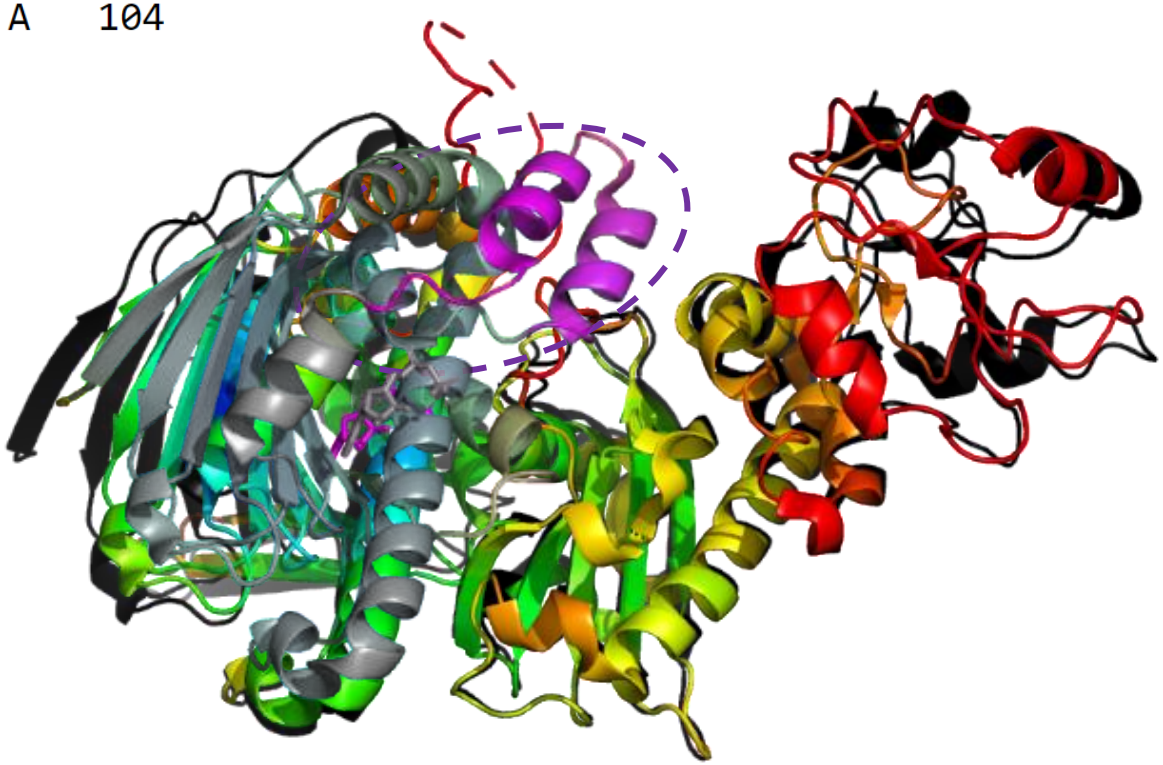


Color scheme:
Blu → Red
Low → High B

B-factors / Missing residues

mmCIF **Category name** pdbx_missing_residue_list

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     LYS A   103
REMARK 465     SER A   104
```

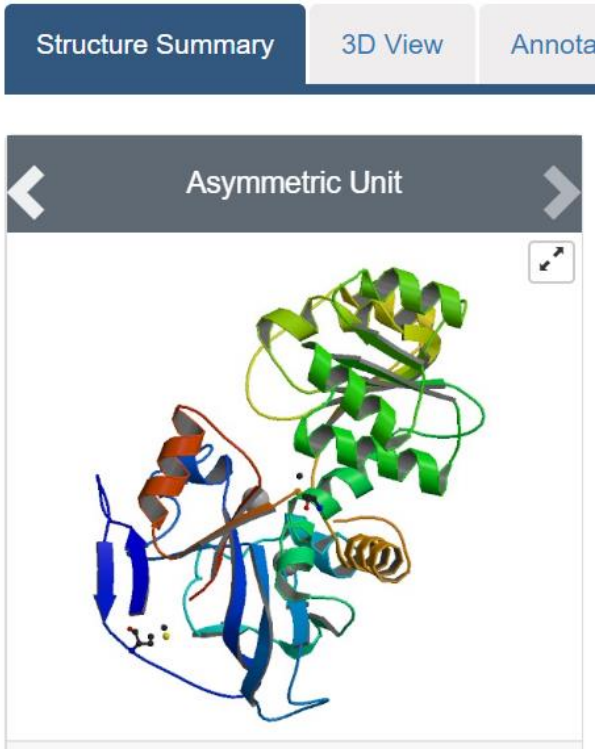


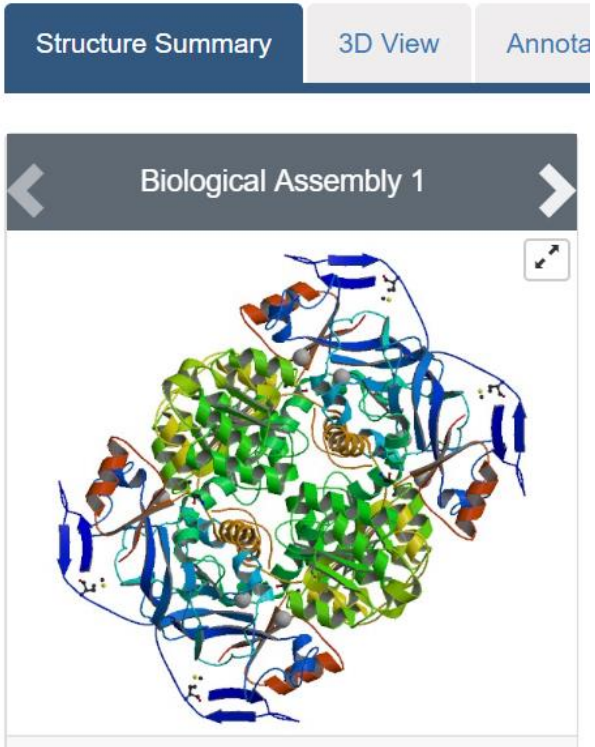
Color scheme:
Blu → Red
Low → High B

Biological Assembly

The biological assembly is the macromolecular assembly that has either been shown to be or is believed to be the functional form of the molecule.

mmCIF **Category name** struct_biol
Category name pdbx_struct_assembly



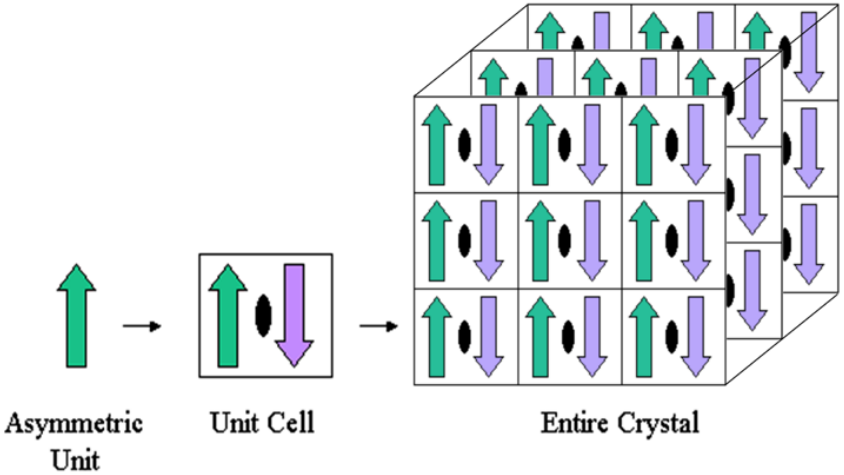



```
REMARK 300 BIOMOLECULE: 1
REMARK 300 SEE REMARK 350 FOR THE AUTHOR PROVIDED AND/OR PROGRAM
REMARK 300 GENERATED ASSEMBLY INFORMATION FOR THE STRUCTURE IN
REMARK 300 THIS ENTRY. THE REMARK MAY ALSO PROVIDE INFORMATION ON
REMARK 300 BURIED SURFACE AREA.
REMARK 300 REMARK: THE BIOLOGICAL ASSEMBLY IS A TETRAMER GENERATED FROM THE
REMARK 300 MONOMER IN THE ASYMMETRIC UNIT BY THE OPERATIONS: -Y,-X,-Z, Y,X,
REMARK 300 -Z, -X,-Y,Z
```

```
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW.  BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: TETRAMERIC
```

Biological Assembly

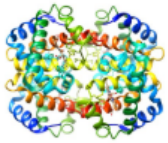
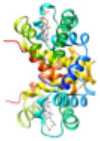
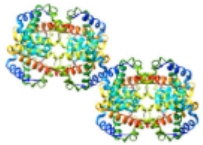
The biological assembly is the macromolecular assembly that has either been shown to be or is believed to be the functional form of the molecule.



A crystal asymmetric unit may contain:

- one biological assembly
- a portion of a biological assembly
- multiple biological assemblies

Hemoglobin

Asymmetric unit with one biological assembly	Asymmetric unit with a portion of a biological assembly	Asymmetric unit with multiple biological assemblies
		
Entry 2hhb contains **one** hemoglobin molecule (**4 chains**) in the asymmetric unit.	Entry 1out contains **half** a hemoglobin molecule (**2 chains**) in the asymmetric unit. A crystallographic two-fold axis generates the other 2 chains of the hemoglobin molecule.	Entry 1hv4 contains **two** hemoglobin molecules (**8 chains**) in the asymmetric unit.

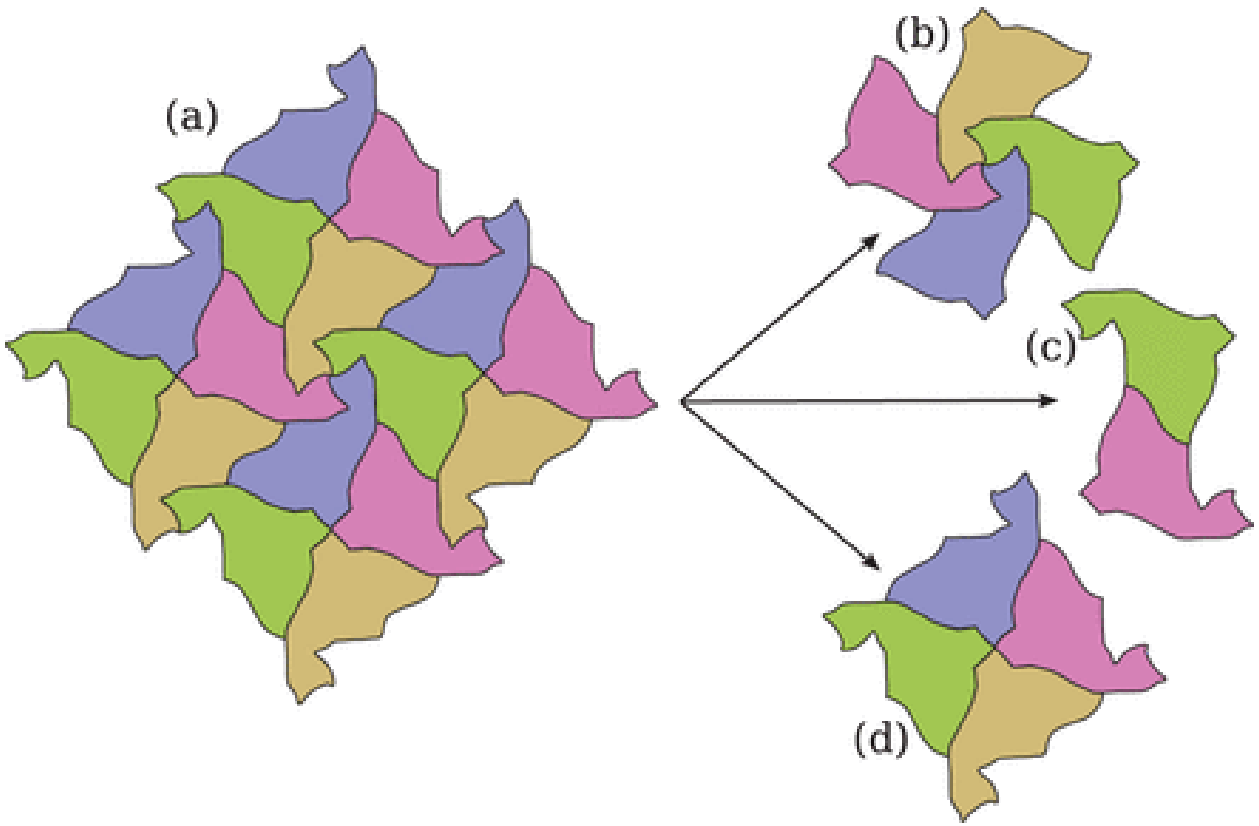
In all structures of Hemoglobin, the biological assembly is a tetramer.

RCSB PDB-101

Educational portal of RCSB PDB

Interfaces in the lattice are biologically relevant?

Interface classification problem. Given (a) as a crystal lattice, you could choose any of (b–d) as the biological unit. Further information is needed to identify which arrangement represents the true biological unit.



Capitani et al., *Bioinformatics* 2016

How to distinguish biologically relevant contacts from crystal packing contacts?

1. Evaluate all protein contacts (interfaces) in crystal
2. Leave only the strongest ones - what you get will have major chances to be a stable protein complex ("biologically relevant") .

→ Method to evaluate the chemical stability of protein assemblies

Protein Data Bank in Europe
Bringing Structure to Biology
PDBePISA

https://www.ebi.ac.uk/pdbe/pisa/

PISA:
based on the **binding energy of the interface** and the **entropy change due to complex formation.**
In the binding energy term: Buried Surface Area (BSA), defined as the difference in Accessible Surface Area (ASA) between uncomplexed and complexed structures; hydrogen bonds, salt bridges and disulfide bonds.

→ 80-90% success rate ; classification is hard in the low range of interface area 400-2000 $Å^2$ – weak biological interfaces more similar to crystal contacts

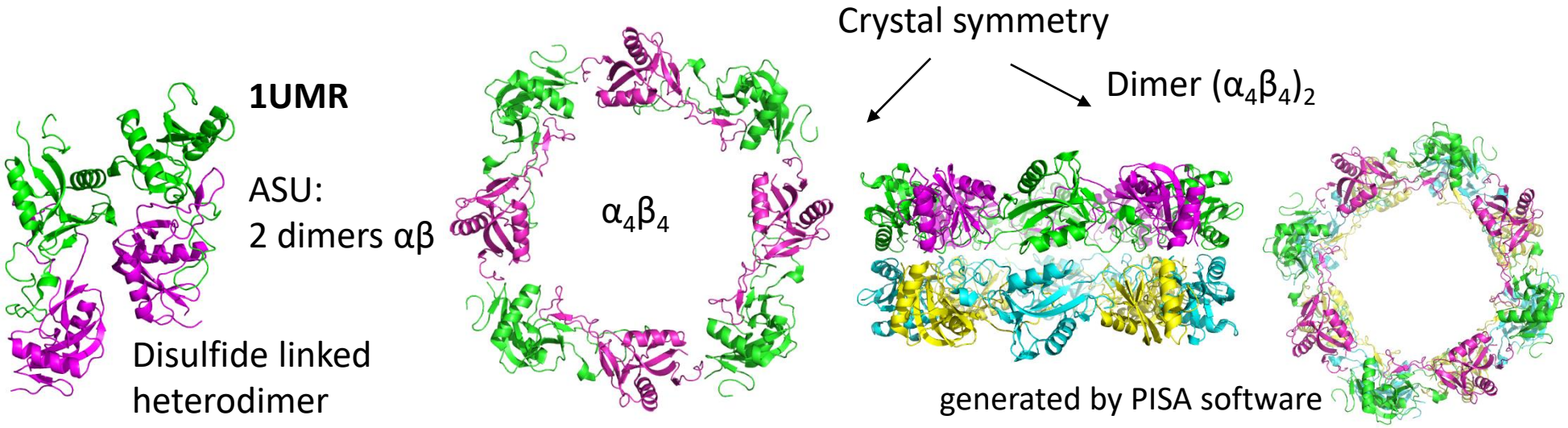
EPPIC: based on geometrical features of the interface but **also on evolutionary features** (conservation of residues at the interfaces by multiple sequence align. of all sequence homologs)

http://www.eppic-web.org

Improving current methods: MD simulations of the interface of interest

Example of importance of crystal packing for information about the biological assembly

Convulxin (CVX) is a C-type lectin-like protein from the venom of the South American rattlesnake that functions as a potent agonist of the platelet collagen receptor glycoprotein VI (GPVI).



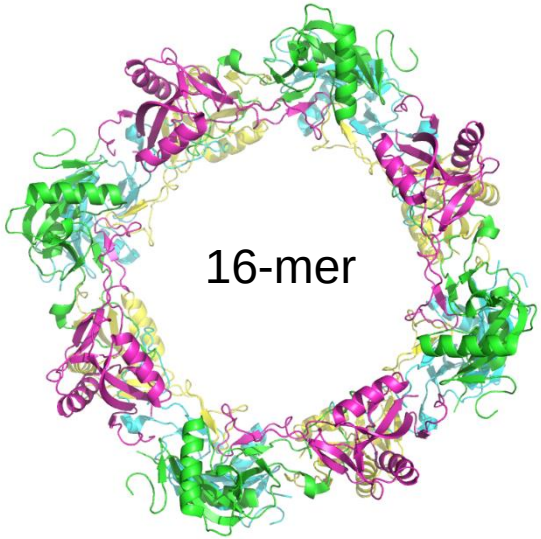
By using Analytical Ultra-Centrifugation it has been shown that CVX exists in solution as a dimer of $\alpha_4\beta_4$ rings, yielding eight potential binding sites for GPVI.
Reanalysis of previously determined crystal structures of CVX revealed the dimer in the available structures.

Horii, Brooks & Herr, *Biochemistry,* 2009

Example of importance of crystal packing for information about the biological assembly

PISA software output:

Assembly



Most probable assembly: 16-mer

Analysis of protein interfaces suggests that the following quaternary structures are stable in solution.

wnload View XML

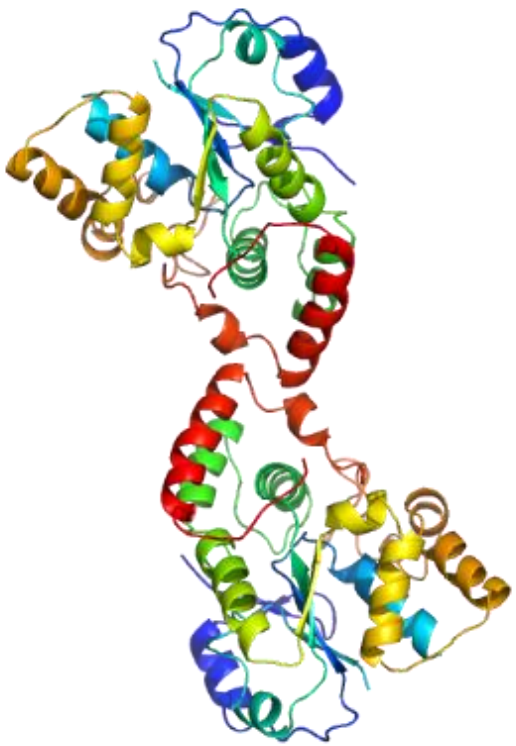
PQS set NN	«»	mm Size	Formula	Composition	Id	Biomol R350	Stable	Surface area, sq. Å	Buried area, sq. Å	ΔG^{int}, kcal/mol	ΔG^{diss}, kcal/mol
1(*)	●	16	A_8B_8	$A_4B_4C_4D_4$	1	–	yes	91690	43810	-240.0	16.8
2	○	8	A_4B_4	B_4D_4	2	2	yes	50190	17550	-115.9	6.1
	○	8	A_4B_4	A_4C_4	2	1	yes	49910	17840	-109.9	4.4
3	○	2	AB	BD	3	–	yes	13270	3670	-24.6	29.2
	○	2	AB	AC	3	–	yes	13220	3720	-23.6	29.2
4(*)	○	2	AB	BD	4	–	yes	16220	720	-4.3	0.5
	○	2	AB	AC	4	–	yes	16190	750	-3.9	0.1

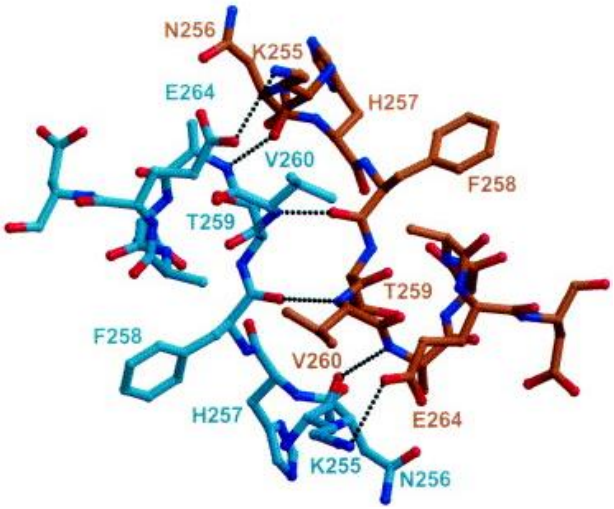
wnload View XML

Difficult cases: small protein-protein interface

Human cytosolic sulfotransferases are classified as monomeric in the PDB structures. PISA and EPPIC failed to find a dimeric stable assembly. It has been confirmed experimentally that the biological unit is a dimer which is present in the crystal structure.

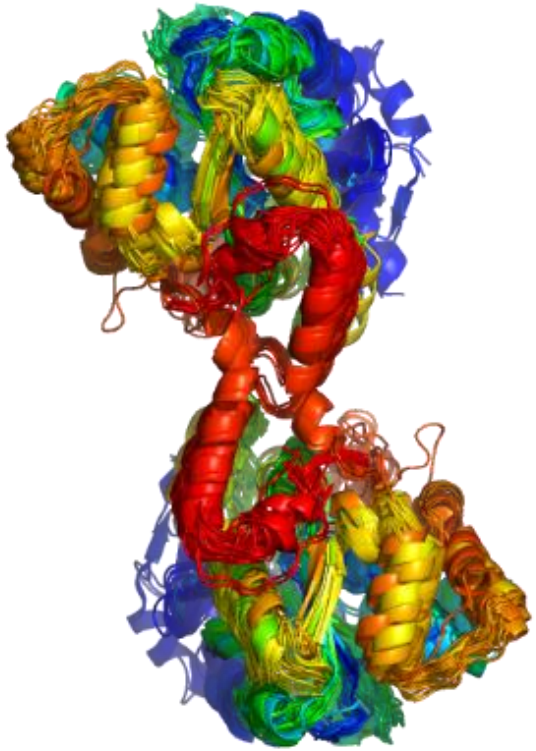
2H8K





Site-direced mutagenesis of residues at the interface and gel filtration analysis

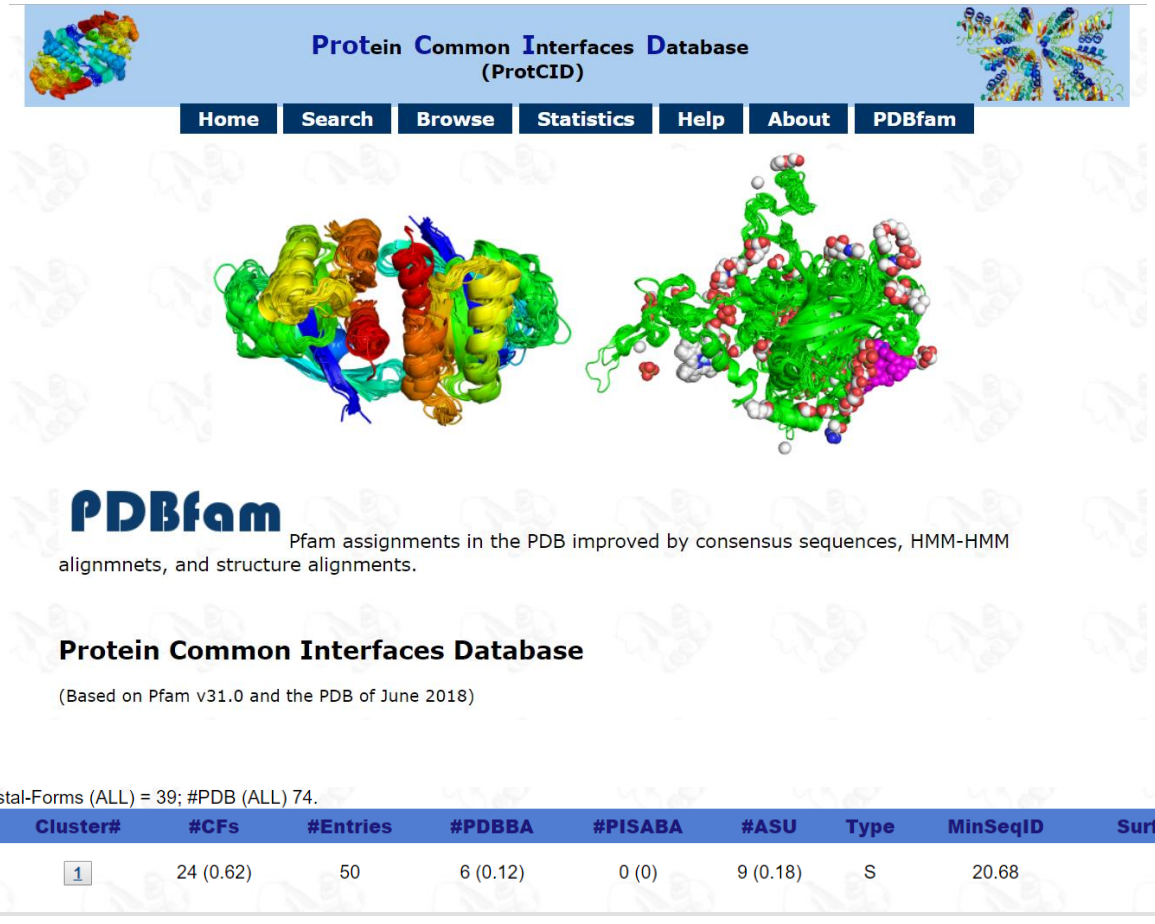
Further support from the **conservation of the interface across different crystal forms of compared crystal of homologous proteins**



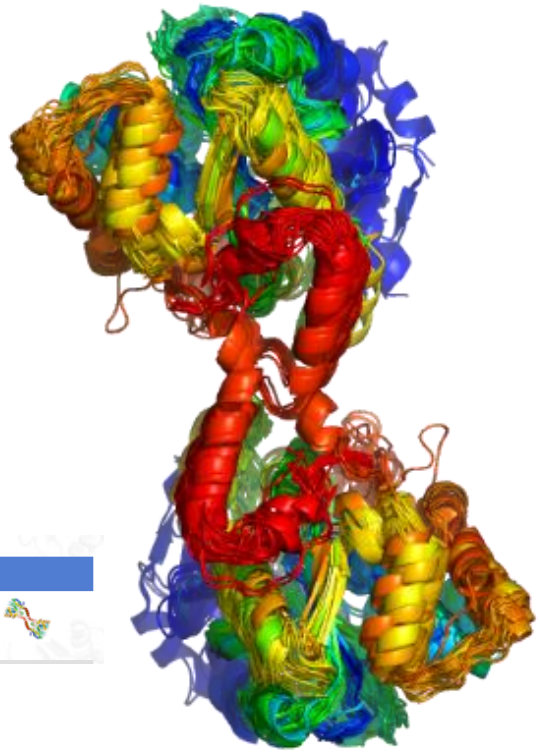
Petrotchenko EV et al, *FEBS Letters*, 2001

Difficult cases: small protein-protein interface

Human cytosolic sulfotransferases are classified as monomeric in the PDB structures. PISA and EPPIC failed to find a dimeric stable assembly. It has been confirmed experimentally that the biological unit is a dimer which is present in the crystal structure.



Further support from the **conservation of the interface across different crystal forms of compared crystal of homologous proteins**



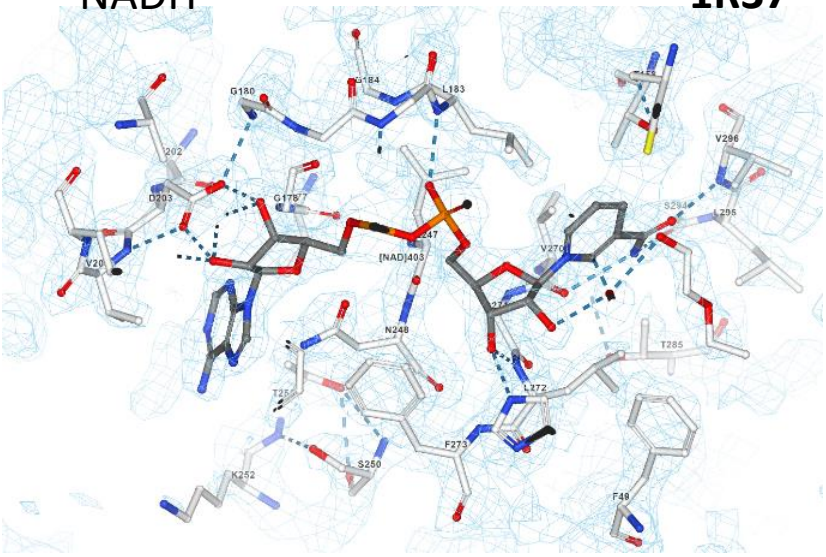
http://dunbrack2.fccc.edu/protcid/

Site binding sites (author and/or software annotation)

mmCIF **Category name** struct_site

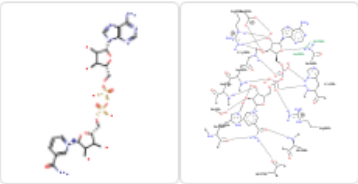
```
REMARK 800 SITE
REMARK 800 SITE_IDENTIFIER: ZNS
REMARK 800 EVIDENCE_CODE: AUTHOR
REMARK 800 SITE_DESCRIPTION: RESIDUES THAT BIND A STRUCTURAL ZINC ION (ZN
REMARK 800  A 400) IN THE CATALYTIC DOMAIN
...
SITE     1 ZNS  4 CYS A 112  GLU A  98  CYS A 101  CYS A 104

REMARK 800 SITE_IDENTIFIER: AC5
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE NAD A 403
...
SITE     1 AC5 33 CYS A  38  HIS A  39  SER A  40  HIS A  43
SITE     2 AC5 33 CYS A 154  THR A 158  GLY A 178  GLY A 180
SITE     3 AC5 33 GLY A 181  GLY A 182  LEU A 183  ASP A 203
SITE     4 AC5 33 VAL A 204  ARG A 205  LEU A 247  ASN A 248
SITE     5 AC5 33 VAL A 270  GLY A 271  LEU A 272  PHE A 273
SITE     6 AC5 33 SER A 294  LEU A 295  VAL A 296  LEU A 334
SITE     7 AC5 33 ARG A 342   ZN A 500  ETX A 600  HOH A 605
SITE     8 AC5 33 HOH A 613  HOH A 650  HOH A 664  HOH A 704
SITE     9 AC5 33 THR B 285
```

NADH **1R37**

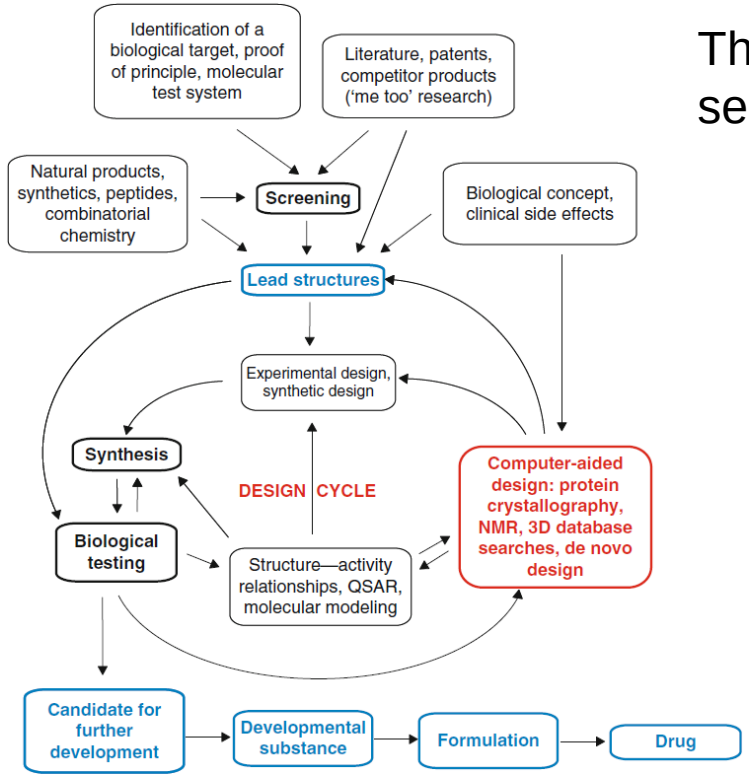


Small Molecules

Ligands 3 Unique

ID	Chains	Name / Formula / InChI Key	2D Diagram & Interactions	3D Interactions
NAD Query on NAD Download SDF File Download CCD File	A, B	**NICOTINAMIDE-ADENINE-DINUCLEOTIDE** $C_{21}H_{27}N_7O_{14}P_2$ BAWFJGJZGIEFAR-NNYOXOHSSA-N		Ligand Interaction

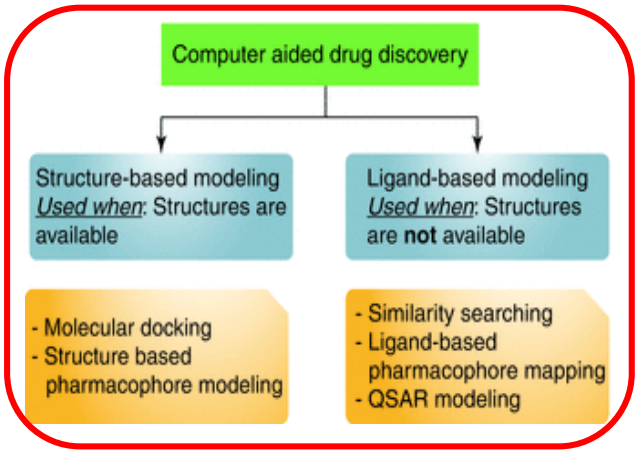
Drug Design



The starting point in the development of a new drug is the search for an appropriate **lead structure** for a target protein.

- ✓ High-throughput screening assay using natural and synthetic product libraries ("random approach")
- ✓ A "rational" approach involves the computer aided drug discovery.
  Structure-based route to drug discovery, where the decisive prerequisite is the knowledge about the **structure of the target protein**.

The 3D structure is the starting point for the initial design of a ligand, which is then to be synthesized and tested. In the case of good binding, an attempt is made to determine the 3D structure of the protein–ligand complex with the new compound. This is the starting point of the next design cycle.
→ Optimization of the lead structure (potency, selectivity, and metabolism)



- G. Klebe, Drug Design - Methodology, Concepts, and Mode-of-Action, 2013, Springer ed.
- Lu et al, Computer-Aided Drug Discovery in Accelerated Path to Cures, 2018, Springer ed.

Drug Design: Contribution of protein structures and PDB to New Drug approvals

US FDA approved 210 new molecular entities (NME; new drugs) in 2010-2016.

A) Small chemicals of MW<1000 Da (81.4%)
B) Proteins and peptide hormones with MW>=1000 (17.1%)
C) Nucleic acid drugs (antisense oligonucleotides that target mRNA) (1.5%)

- ✓ About **93%** of NME have a «relevant» structure in the PDB. 'Relevant': Structures that contain a known (>=95% seq id to the Uniprot reference sequence for the target protein) protein target of a NME or one of the NMEs.
- ✓ **95%** of relevant structure were determined using **X-ray** (median resolution limit, 2.08Å).
- ✓ For most of the known protein target of a given NME, the PDB contains more that one relevant structure (80%).
- ✓ A considerable fraction of the relevant structures (37%) include both the NME protein target and one or more partner proteins.

Westbrook & Burley, *Structure* 2019

Structure-based Drug Design

- ✓ In structure-based drug design, attempts are made to design small-molecule ligands by docking them directly into the binding pocket of a target protein.

- ✓ Structure-based design starts with a detailed **analysis of the binding pocket** to elucidate hot spots for putative interactions with the protein. Either experimental methods or computational tools can be used to perform an active site mapping with molecular probes or small solvent-like molecules.

- ✓ In an **iterative process** of structure determination, modeling of modified ligands, docking and screening, synthesis, and biological testing, the properties of small molecule ligands are improved to optimize binding to the target protein.

Molecular Docking

A plethora of software suites developed for the **automated docking of flexible small molecules into (mainly rigid) protein structures**:
Dock, GOLD, Autodock, RosettaLigand, SwissDock…

1) **Sampling procedure** (genetic algorithms-based optimization in the conformational space of the rotatable bonds or grid-based searches)
2) **Evaluation of the binding energy** (force-field based) considering also contribution for desolvation
3) **Multiple binding geometry solutions to be ranked** on the base of the estimated binding affinity – Scoring functions

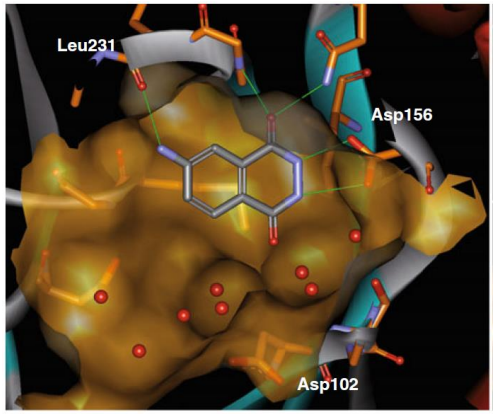
Crystal structures with newly found lead structures are important **to rationalize uncorrect binding predictions and to afford many new ideas for synthetic entry points to develop new inhibitors.**

→ Role of **local conformational changes** of protein target and of **interstitial water molecules**

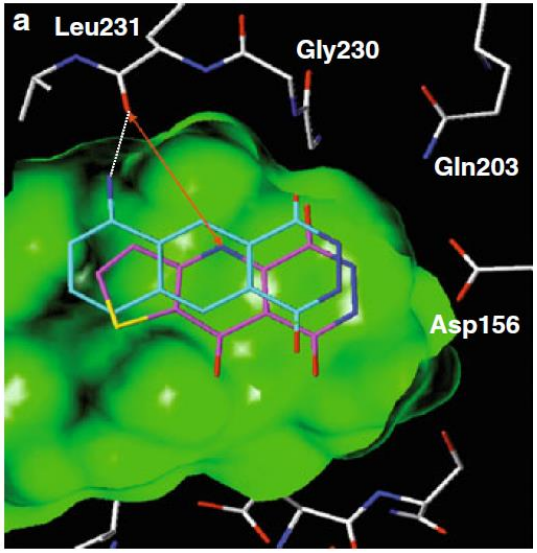
Molecular Docking

Structure-Based Inhibitor Design for tRNA-Guanine Transglycosylase (TGT)

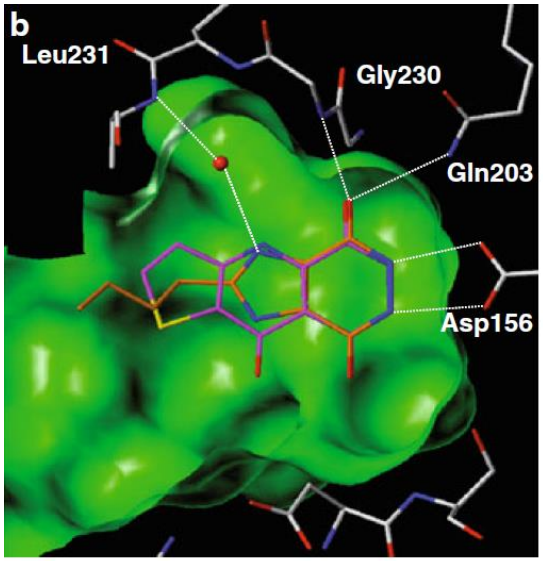
Crystal structure of TGT the first hit found



In the process of searching for an analogue, the compound in magenta is found, but in the docking the distance between the ring nitrogen atoms and the carbonyl group on Leu231 seems to be too large for an H-bond. Nonetheless, the compound binds to the protein with micromolar affinity!



X-Ray structure of a similar compound (orange)



The active site shows adaptations by flipping a peptide bond and mediating important interactions to the substrates through a water molecule

G. Klebe, Drug Design - Methodology, Concepts, and Mode-of-Action, 2013, Springer ed.

Molecular Docking

Challenges in protein-ligand docking :

- ✓ Receptor flexibility upon ligand binding → many binding sites cannot be represented by a single snapshot

- ✓ Treatment of water molecules (which are obligate in the binding site and which can be displaced by incoming ligand?) – importance of water in recognition

The water stability in the binding site can be determined based on the analysis of multiple crystal structures or…by running

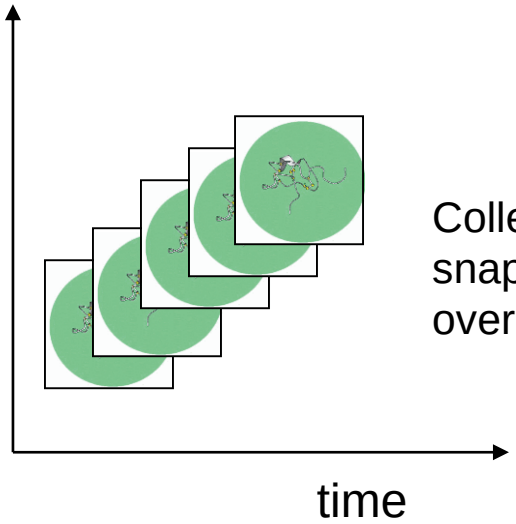
Molecular Dynamics (MD) simulations

Molecular Dynamics (MD) simulations

Computational technique that describes the time dependent behavior of a molecular system

Simulation of the temporal evolution of a system of N particles

MOLECULAR DYNAMICS is a way to generate an ensemble of conformations



Collection of snapshots over time

A trajectory (configurations as a function of time) of the molecular system is generated by simultaneous integration of Newton's equations of motion for all the atoms in the system

$$\vec{F}_i = m_i \vec{a}_i \qquad \vec{F}_i = -\, dV/d\,\vec{r}_i$$

Dynamics of the system

$$\vec{F}_i = m_i\, d\vec{v}_i/d\,t = m_i\, d^2\vec{r}_i/d\,t^2$$

V potential energy function / Force fied

Molecular Dynamics (MD) simulations

Hollingsworth & Dror, *Neuron* 2017

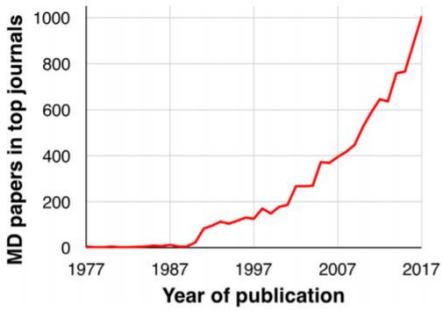
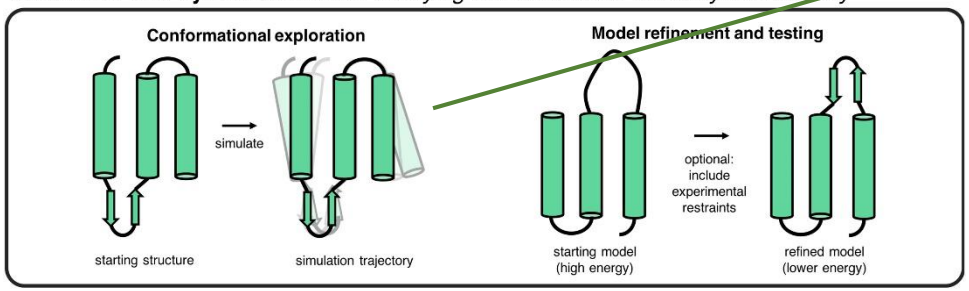
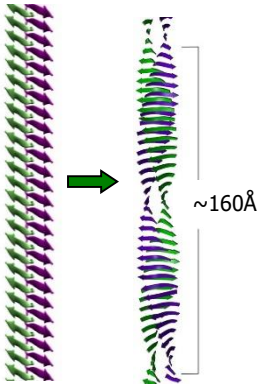


Figure 1. Growth of Molecular Dynamics Simulations in Structural Biology
For the top 250 journals by impact factor, we plotted the number of publications per year that include the term "molecular dynamics" in either the title, abstract, or keywords. The analysis was performed via Web of Science (https://www.webofknowledge.com/) in February 2018.

Structural and dynamic studies: *Studying conformational flexibility and stability*



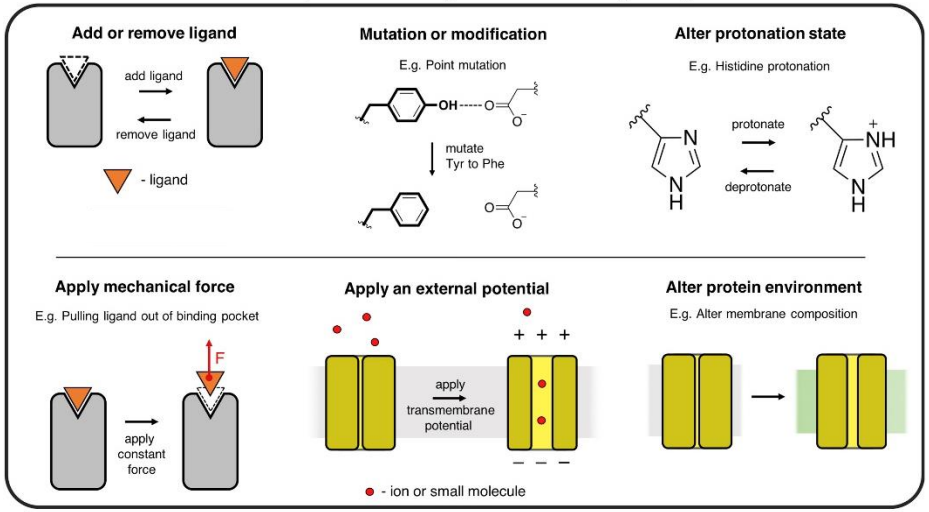
An amyloidogenic peptide
From crystal to solution



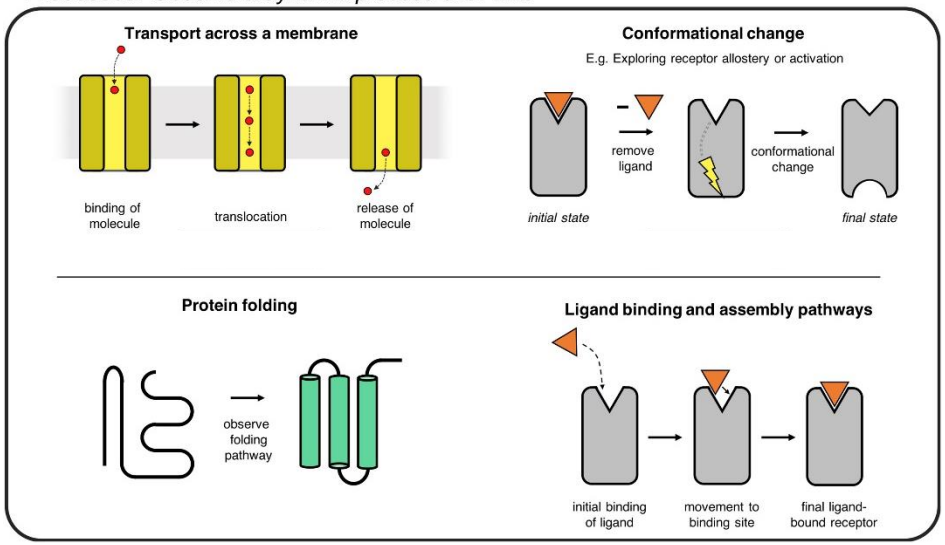
Esposito et al. *PNAS* 2006

'Computer alchemy'

Perturbations: *Observe response following controlled change to system*



Processes: *Observe a dynamic process over time*



Karplus & McCammon, *Nat. Struct. Mol. Biol.* 2002

Molecular Dynamics (MD) simulations

MD and Docking Data

@ Starting stages of docking

- MD for generating an ensemble of protein conformations (MD snapshots) that can be used in High-throughput Docking (thousands of molecules) / a single MD snapshot is selected by using a small set of inhibitors or fragments that all have a favourable binding (the inhibitor can also be unable to bind the crystal structure)

@ Final stages of docking

- MD validation of the top ranking docking poses (stability of the structure) / MD for ligand optimization (suggestion points of chemical derivatization.

MD and SAX Data
(protein size and shape)

Simulations can be used to select a minimal ensemble of structures that best fit the experimental SAXS data. Alternatively, the simulations can be used to generate a full ensemble of structures that can be directly validated by the experimental SAXS data.

MD and EM Data

MD helps in constructing atomic structural models into cryo-EM densities. MDFF Flexible Fitting of atomic structures: the MD simulation incorporates EM data through an external potential effectively biasing the system toward the region with the density distribution of the EM map.

Trabuco et al, *Structure* 2008

Sequence-based Function Predictions

"Protein A has function X, and protein B is a **homolog** (ortholog) of protein A; Hence B has function X"

Homology between two proteins means that they have **a common evolutionary origin**.

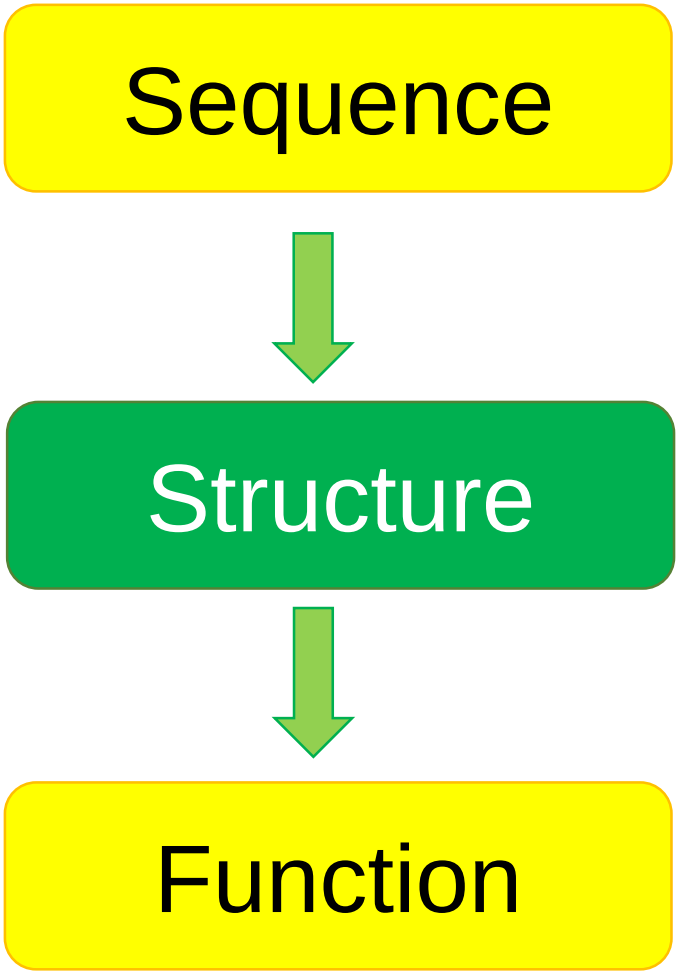
Homologous sequences are **orthologous** if they were separated by a **speciation** event

Homologous sequences are **paralogous** if they were separated by a gene **duplication** event:

In general, **function tends to be more conserved in othologs** than in paralogs

The most common way to infer homology is by detecting **sequence similarity**

Sequence similarity



Sequence-based Function Predictions

The higher the sequence similarity the better the chance that two proteins share the same function

However, one should keep in mind that sometimes very high sequence identity does not guarantee the same function.

An extreme case is represented by the so-called moonlighting proteins, i.e. proteins in which more than one physiologically relevant discrete function is performed by a single polypeptide chain

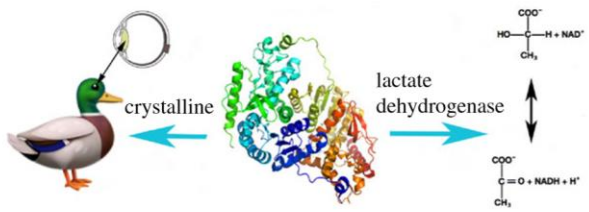


Figure 1. A moonlighting protein can have two very different functions in the same species. For example, in ducks, the epsilon crystalline found in the lens of the eye is the same protein as the ubiquitous enzyme lactate dehydrogenase, which catalyses the interconversion of pyruvate and lactate. (Online version in colour.)

Homologs of these proteins may retain only one of the original functions

MoonProt 2.0 (http://moonlightingproteins.org)

Sequence

↓

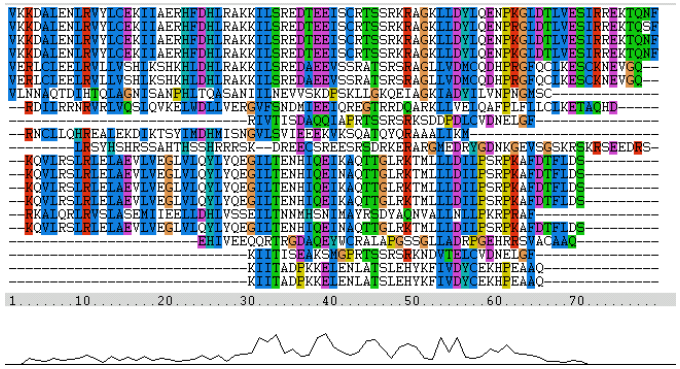
Structure

↓

Function

Jeffrey CJ, *Phil. Trans. R. Soc.* 2017
Chen et al., *NAR* 2018

Sequence similarity



Sequence alignment **methods** BLAST, FASTA, PSI-BLAST, HMMER…

- ✓ The **full-lenght sequence** is aligned against a **database** of annotated proteins such as UniProtKB/SWISS-PROT

- ✓ Search for **sequence signatures**: functional motif search InterProScan

Scan against PROSITE, PRINTS, PFam-A, TIGRFAM, PROFILES and PRODOM motifs

Small sequence signature may suffice to conserve the function of a protein even if the rest of the protein has changed considerably during the course of evolution

On the basis of the alignment also a **model structure** can be built for the protein of interest by using **Homology modeling procedures**.

Homology modeling workflow

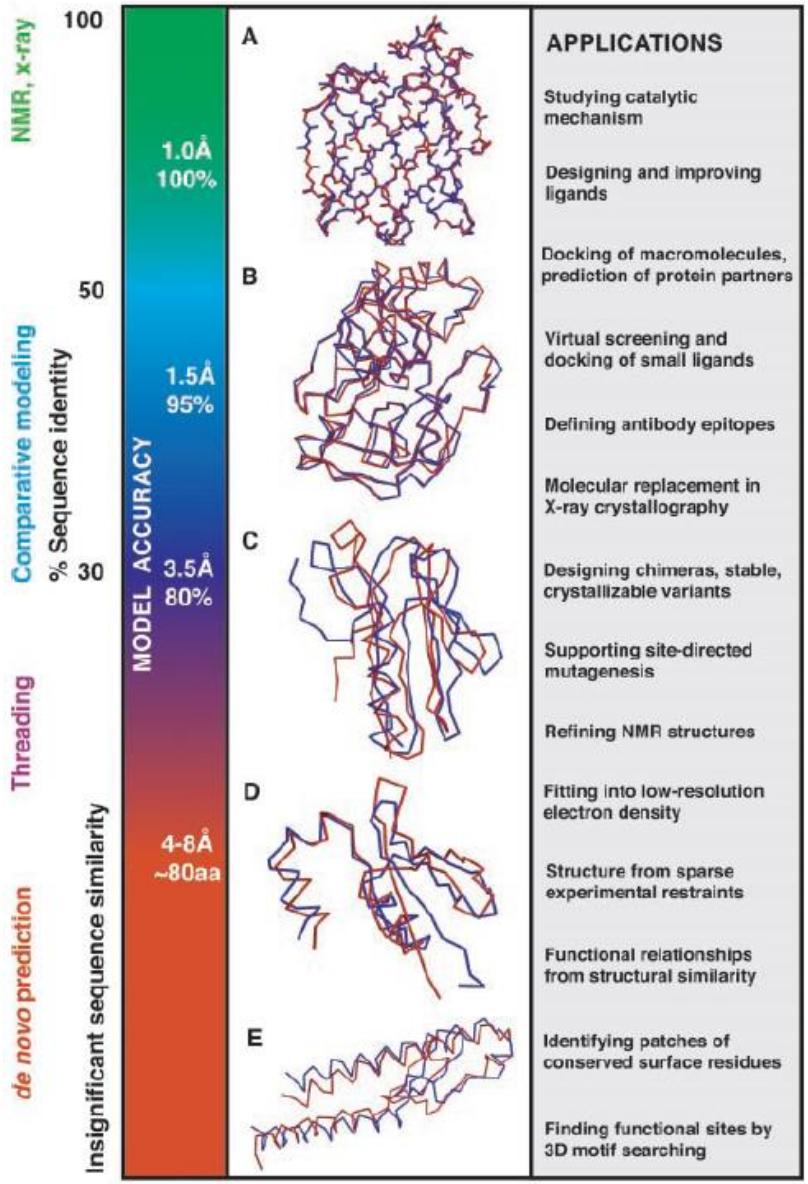
1. Identify a set of template proteins (with known structures) related to the target protein. This is based on sequence search similarity (i.e. BLAST, PSI-BLAST, Hhblits…).
2. Align the target sequence with the template proteins. This is based on multiple alignment (i.e. CLUSTALW). Identify conserved regions.
3. Build a model of the protein backbone, taking the backbone of the template structures (conserved regions) as a model.
4. Model the loops.
5. Add side chains to the model backbone.
6. Evaluate and optimize the entire structure.

Comparative modeling accuracy

High-accuracy comparative models are based on more than 50% sequence identity to their templates

Medium accuracy: 30-50%

Low accuracy: < 30%



Baker & Sali, *Science* 2000

The Protein Model Portal

https://www.proteinmodelportal.org/

PMP gives access to various models computed by comparative modeling methods provided by different partner sites, and provides access to various interactive services for model building, and quality assessment.

SWISS-MODEL was the first fully automated protein homology modelling. The pipeline relies on BLAST/HHblits and ProMod3. The modelling functionality has been recently extended to include **the modelling of homo- and heteromeric complexes**, given the amino acid sequences of the interacting partners as starting point.

https://swissmodel.expasy.org/

https://modbase.compbio.ucsf.edu/modbase-cgi/

MODBASE is a queryable database of annotated protein structure models. The models are derived by ModPipe, an automated modeling pipeline relying on the programs PSI-BLAST and MODELLER. The database also includes the fold assignments and alignments on which the models were based.

Structure-based Function Predictions

Function → Structure

- **The classical way**
  - A function is discovered and studied
  - The gene responsible for the function is identified
  - Product of this gene is isolated, crystallized and the structure solved
  - The structure is used to "rationalize" the function and provide molecular details

Structure → Function

- **Post-genomic**
  - A new, uncharacterized gene is found in a genome
  - Predictions or high-throughput methods select this gene for further studies
  - The protein is expressed and has to be studied in detail
  - The structure is solved and can be the first experimental information about the "hypothetical" protein whose function is unknown

The structure is more evolutionary conserved than sequence

Even when their sequences have changed considerably, protein structures can be well conserved

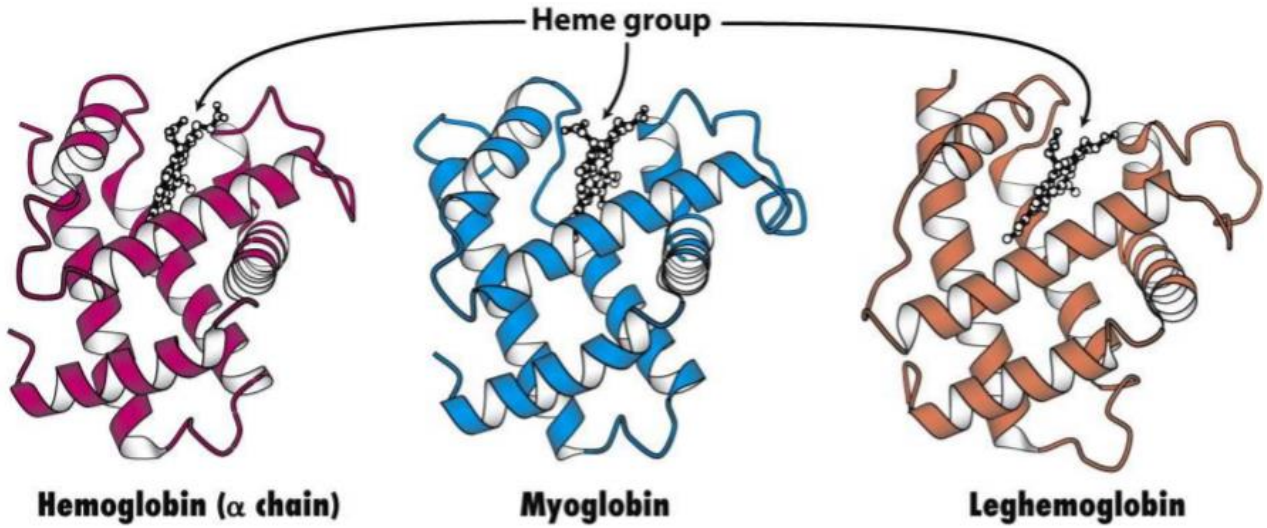


Figure 6-14
Biochemistry, Sixth Edition

Conservation of 3-D structure. The tertiary structures of human hemoglobin, human myoglobin, and lupine leghemoglobin are conserved. This structural similarity firmly establishes that the framework that binds the heme group has been conserved over a long evolutionary period.
On the other hand, sequence similarity between human myoglobin and lupine leghemoglobin is just barely detectable at sequence level and that between human hemoglobin and lupine leghemoglobin is not statistical significant.

Structure-based Function Predictions

Structure → Function

Structural similarity between two proteins, even in the absence of significant sequence similarity, possibly suggests **similar function**.

The structural similarity can be due to a common evolutionary origin or it may indicate evolutionary covergence caused by common functional constraints.

Global structural similarity

Local structural similarity

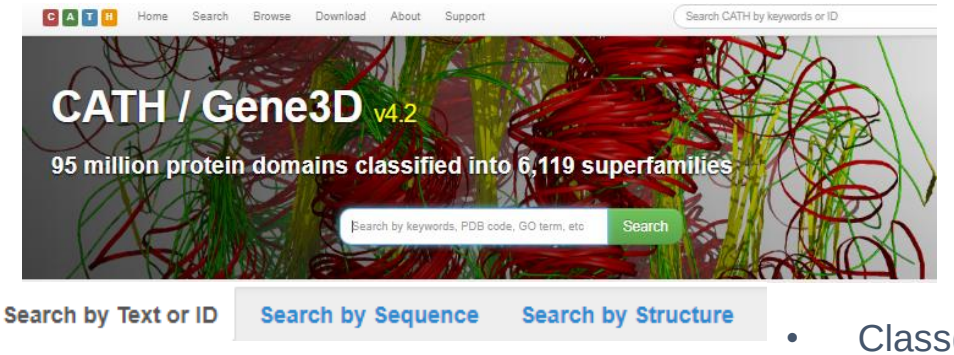
Globally similar structure: Protein fold comparison

Compare the studied structure to known structures to see if it has a known fold

Methods for structural comparison using the **PDB** or structure classification databases (es. **CATH, SCOP**) as a source.

Among the most used structural alignment methods are **DALI**, CATHEDRAL, SSM, VAST…

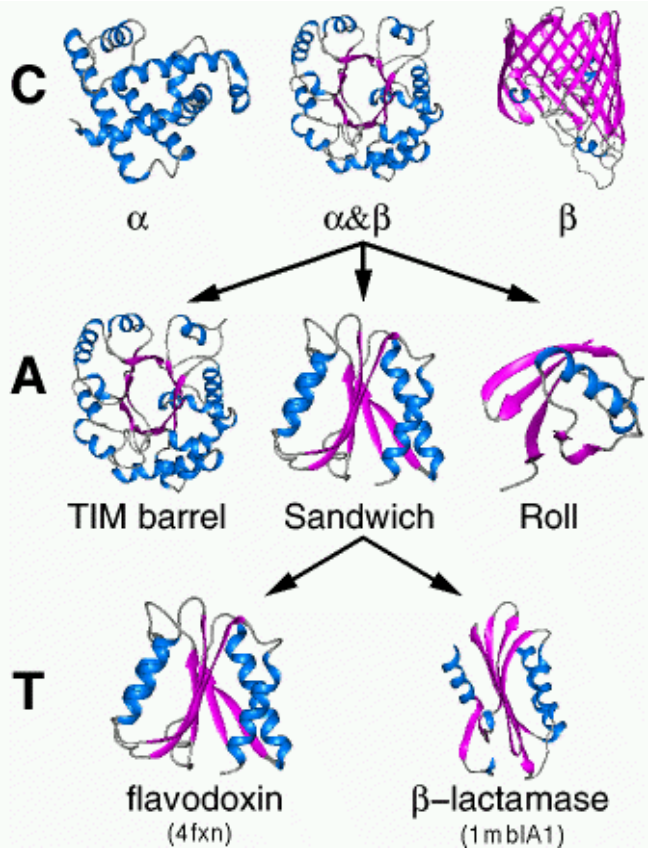
Protein Structure Classification Database



http://www.cathdb.info/

- UCL, Christine Orengo

Protein **domains** are identified within the polypeptide chains using a mixture of automatic methods and manual curation. The **domains** are then classified within the CATH structural hierarchy:



- Class(C)

Similar secondary structure content

All alpha; all beta; alpha-beta; etc

- Architecture(A) *overall shape of the domain structure*

  describes the gross orientation of secondary structures, independent of connectivity.

- Topology(T) (Fold)

  clusters structures according to their topological connections and numbers of secondary structures.

  Protein with the same Topology share the same overall shape and connectivity of the secondary structures in the domain core. Domains in the same fold group may have different structural decorations to the common core.

- Homologous superfamily (H) *groups together protein domains which are thought to share a common ancestor*

  Similarities are identified first by sequence comparisons and subsequently by structure comparison

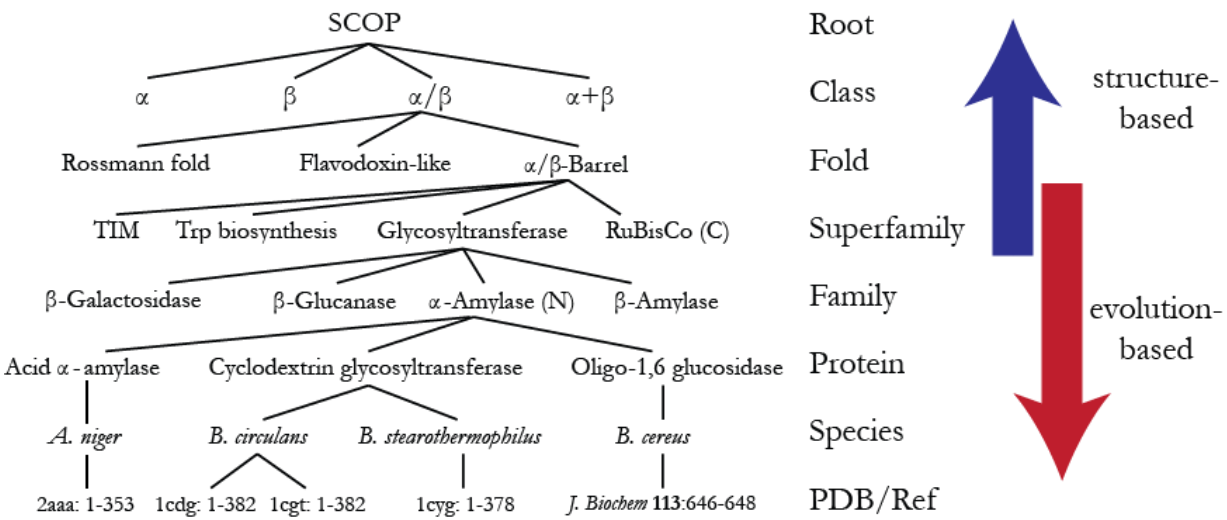
Protein Structure Classification Database

Proteins are organized according to their structural and **evolutionary** relationships

SCOPe | Browse | Stats & History | Downloads | Help

Welcome to **SCOPe**!

SCOPe (Structural Classification of Proteins — extended) is a database developed at the Berkeley Lab and UC Berkeley to extend the development and maintenance of SCOP. SCOP was conceived at the MRC Laboratory of Molecular Biology, and developed in collaboration with researchers in Berkeley. Work on SCOP (version 1) concluded in June 2009 with the release of SCOP 1.75.



- Berkeley

http://scop.berkeley.edu/

•Levels above *Superfamily* are classified based on structural features and similarity, and do not imply homology:
•*Folds* grouping structurally similar superfamilies.
•*Classes* based mainly on secondary structure content and organization.

•*Family* containing proteins with similar sequences but typically distinct functions
•*Superfamily* bridging together protein families with common functional and structural features inferred to be from a common evolutionary ancestor.

SCOP2 Structural Classification of Proteins 2
About | Browser | Graph | Download | Support
MRC Laboratory of Molecular Biology

prototype

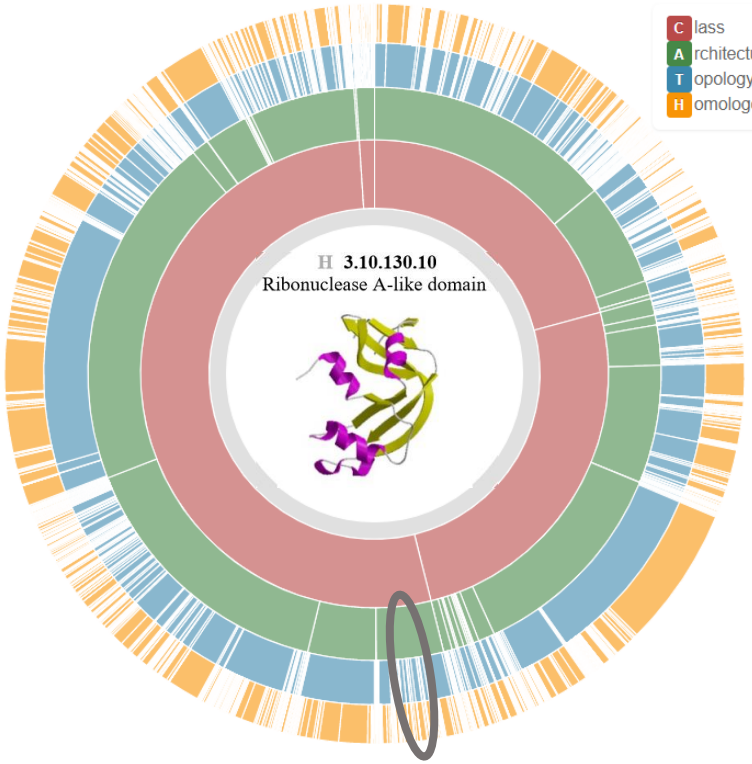
- MRC Cambridge

http://scop2.mrc-lmb.cam.ac.uk/

Superfamily in CATH

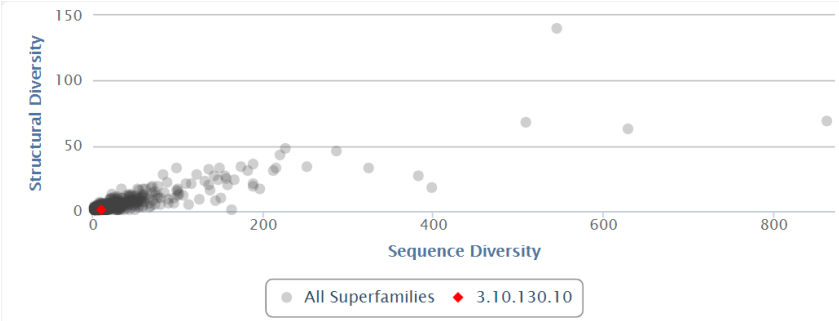
CATH Superfamily 3.10.130.10
Ribonuclease A-like domain

Search by PDB: 1DY5



C lass
A rchitecture
T opology
H omologous Super

Level	CATH Code	Description
C	3	Alpha Beta
A	3.10	Roll
T	3.10.130	P-30 Protein
H	3.10.130.10	Ribonuclease A-like domain



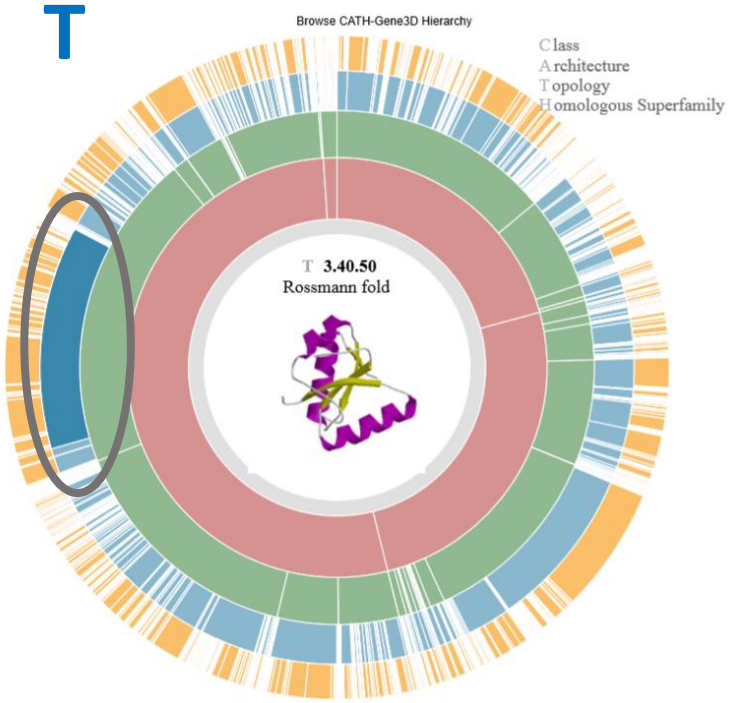
Structures	
Domains:	586
Domain clusters (>95% seq id):	42
Domain clusters (>35% seq id):	10
Unique PDBs:	411
Alignments	
Structural Clusters (5A):	1
Structural Clusters (9A):	1
FunFam Clusters:	22
Function	
Unique EC:	5
Unique GO:	85
Taxonomy	
Unique Species:	632

EC number: numerical classification scheme for enzyme based on the chemical reaction they catalyze.
Gene Ontology (GO): Vocabulary describing the function of a gene in any organism. 3 sets of vocabularies (or ontologies) that describe: molecular function of the gene product, the biological process in which it participates, and the cellular component where it can be found

CATH highly populated fold/superfamily

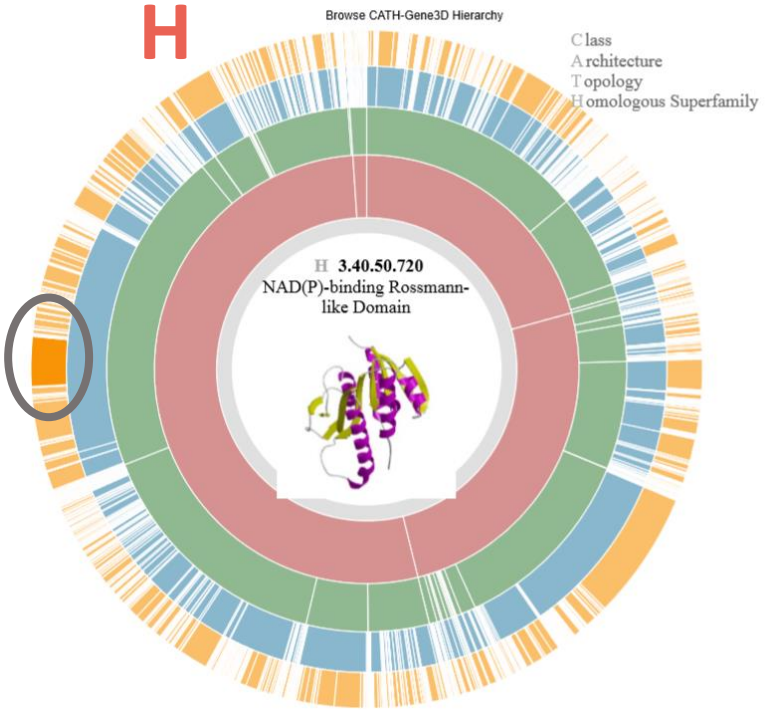
Rossmann fold

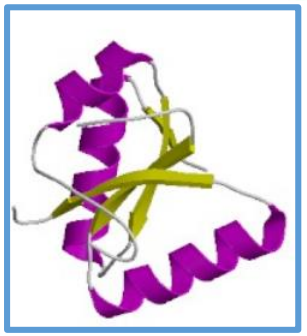
T

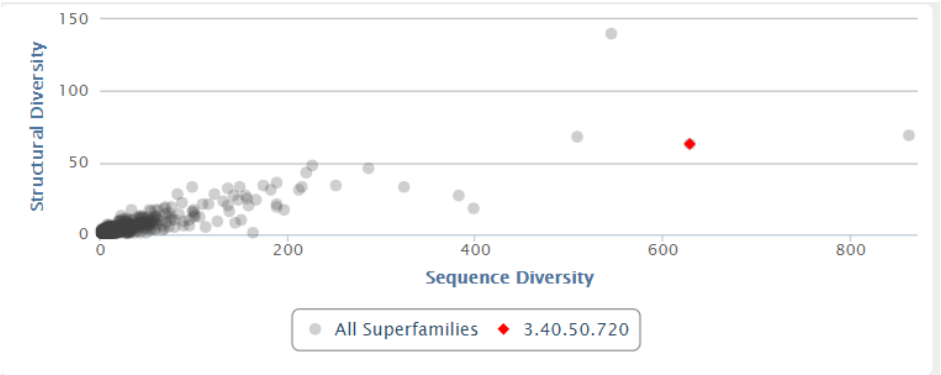


CATH Superfamily 3.40.50.720
NAD(P)-binding Rossmann-like Domain

H







Structures	
Domains:	10268
Domain clusters (>95% seq id):	1328
Domain clusters (>35% seq id):	629
Unique PDBs:	3303
Alignments	
Structural Clusters (5A):	63
Structural Clusters (9A):	29
FunFam Clusters:	1019
Function	
Unique EC:	1006
Unique GO:	2214

The ‘one structure many functions, many structures one function’ paradox

‘Many structures, one function’

Two unrelated proteins can resemble each other structurally. Indeed, two proteins evolving independently may have converged on a similar structure in order to perform a similar biochemical activity.

Convergent evolution: process by which very different evolutionary pathways lead to the same solution (starting from different origin points).

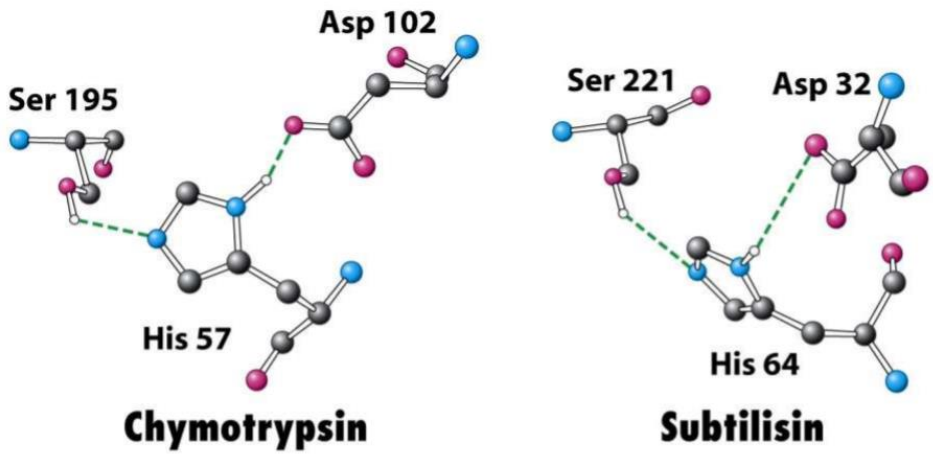


Figure 6-17
Biochemistry, Sixth Edition

One example of convergent evolution is the serine protease family, which cleaves peptide bonds by hydrolysis. The structure of the active sites at which the hydrolysis reaction takes place are remarkably similar.

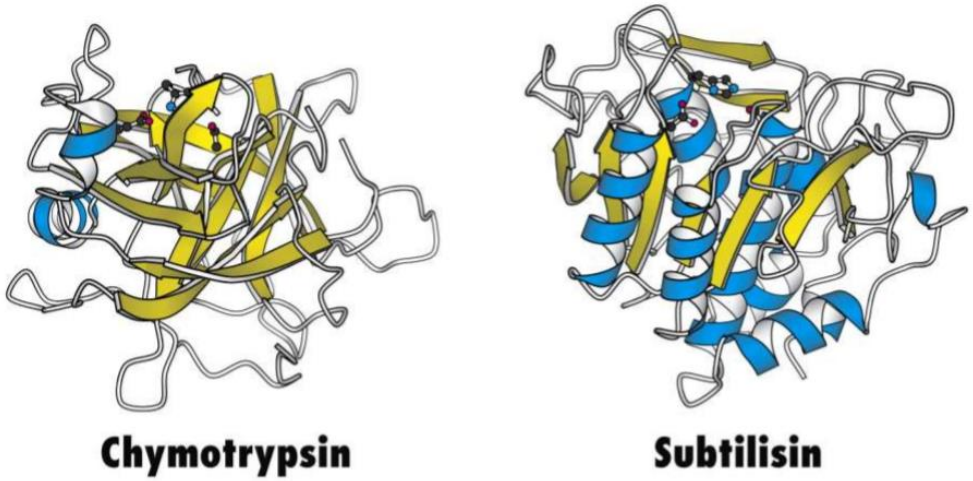


Figure 6-18
Biochemistry, Sixth Edition

The similarity might suggest that these proteins are homologous. However, striking differences in the overall structures of these proteins make an evolutionary relationship extremely unlikely.

The ‘one structure many functions, many structures one function’ paradox

‘One structure, many functions’

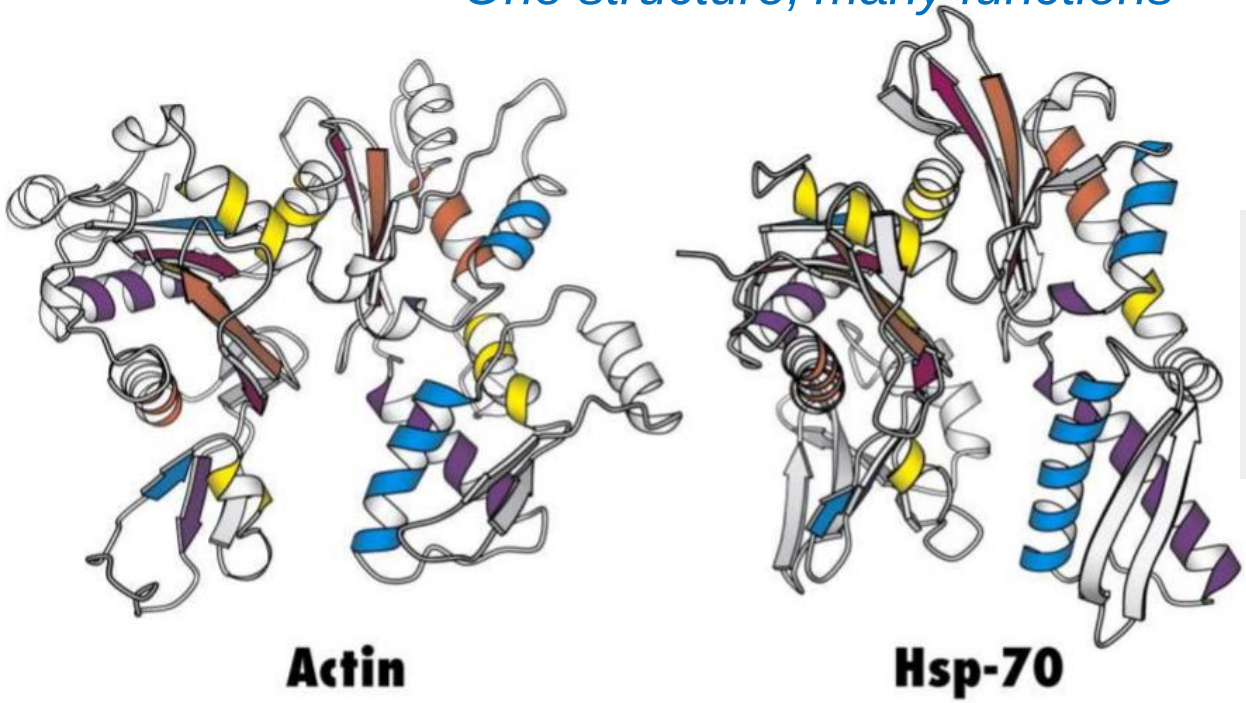


Figure 6-15
Biochemistry, Sixth Edition

Structures of Actin & Hsp-70. A comparison of the identically colored elements of secondary structure reveals the overall similarity in structure despite the difference in biochemical activities.

CATH Classification

Level	CATH Code	Description
C	3	Alpha Beta
A	3.30	2-Layer Sandwich
T	3.30.420	Nucleotidyltransferase; domain 5
H	3.30.420.40	

Superfamily	3.30.420.40
Functional Family	Actin, gamma-enteric smooth muscle

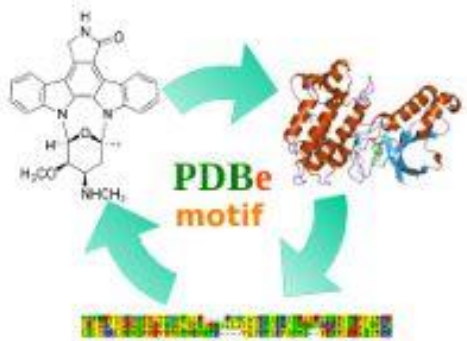
Superfamily	3.30.420.40
Functional Family	Heat shock cognate 70

Locally similar structure

Main chain motif

Supersecondary structure
- EF-hand (calcium binding)
- HTH (DNA binding)

Smaller motif
- Nest (ion-binding site)

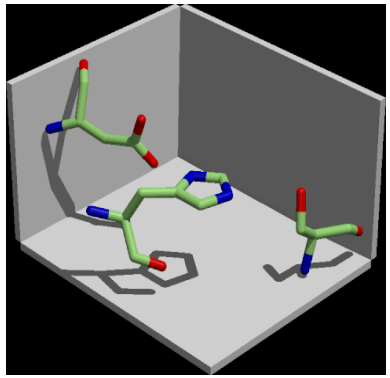


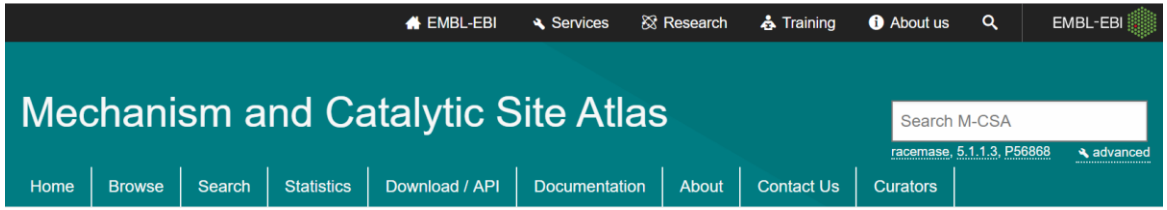
PDBeMotif

https://www.ebi.ac.uk/pdbe-site/pdbemotif/

Groupings of residues

- Catalytic residues
- Ligand binding
- Metal-binding





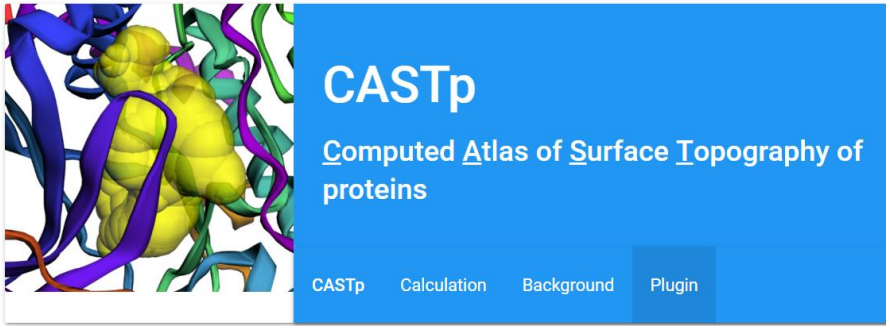
https://www.ebi.ac.uk/thornton-srv/m-csa/

Some methods target specific active-site residues (such as catalytic clusters and ligand-binding sites). These approaches utilize a variety of **template-based scans** to identify active sites and putative ligand-binding sites, the rationale being that **the 3D arrangement of enzyme active-site residues is often more conserved than the overall fold.**

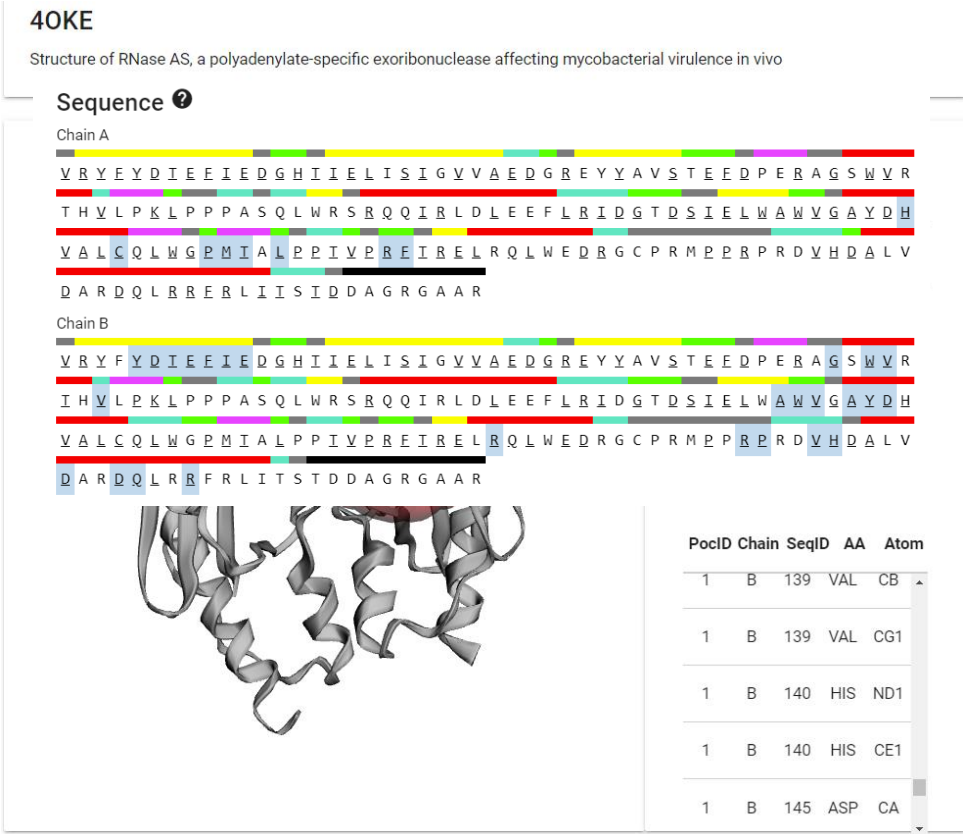
Locally similar structure: Surface based methods

Some methods focus on more localized regions that might be relevant to function, such as clefts, pockets and surfaces. As the ligand-binding site or active site is commonly situated in **the largest cleft** in the protein, the identification and comparison of such regions can suggest putative functions.

https://sts.bioe.uic.edu/castp/

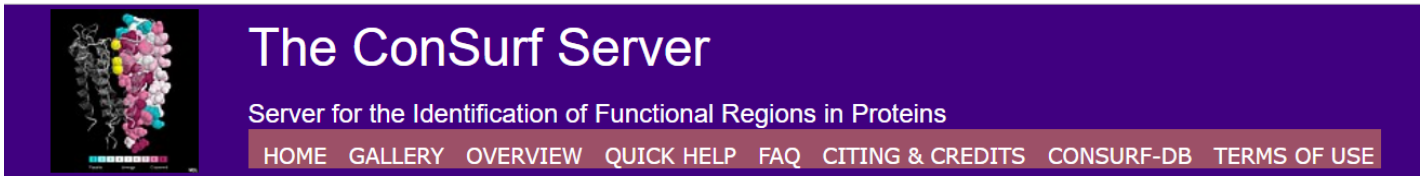


Tian et al., *Nucleic Acids Res.* 2018



CASTp: it detects pockets and cavities. Pockets are empty concavities on a protein surface into which solvent (probe sphere 1.4 A) can gain access; A cavity (or void) is an interior empty space that is not accessible to the solvent probe.

Locally similar structure: Surface based methods

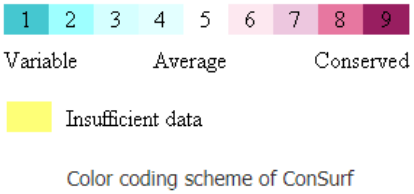


http://consurf.tau.ac.il/

Tian et al., Nucleic Acids Res. 2018

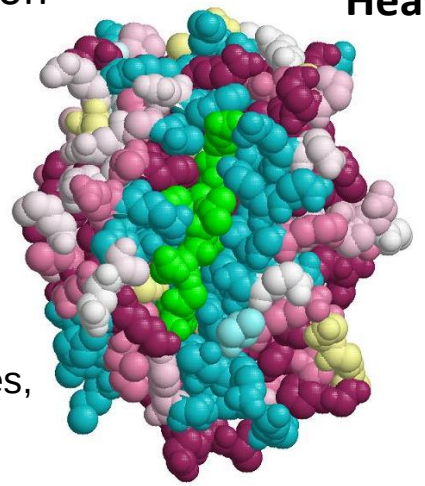
ConSurf: Server for the Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information / *it maps conserved residues in the structure*

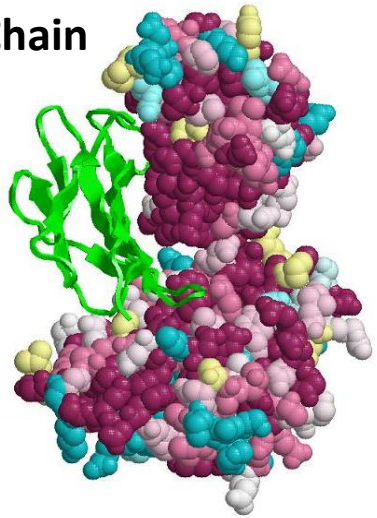
Variability and conservation reflect function



Functional variability is seen in the peptide-binding groove: diversity useful to present the maximal range of viral peptides to T lymphocytes, for defensive immune responses.

MHC Class I Heavy Chain



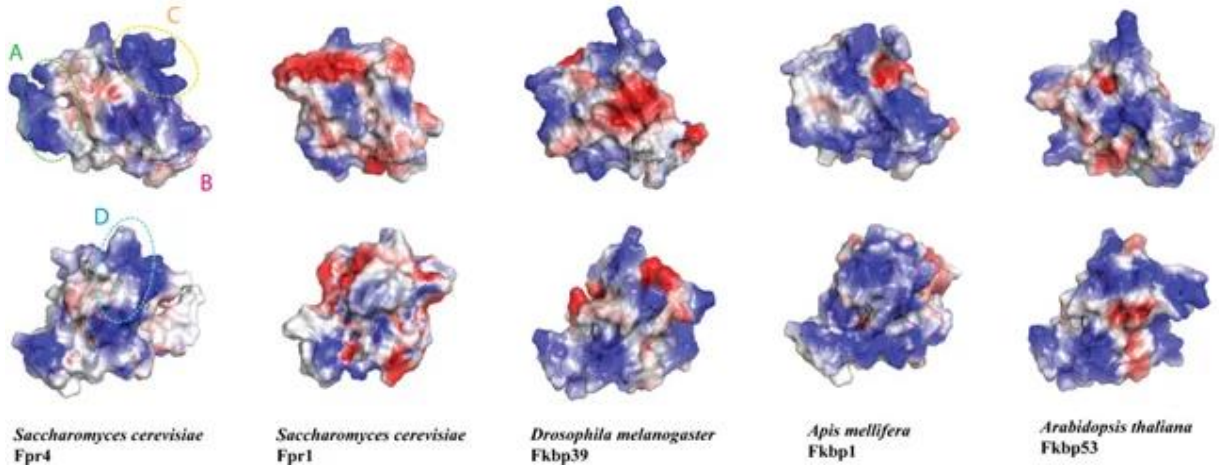


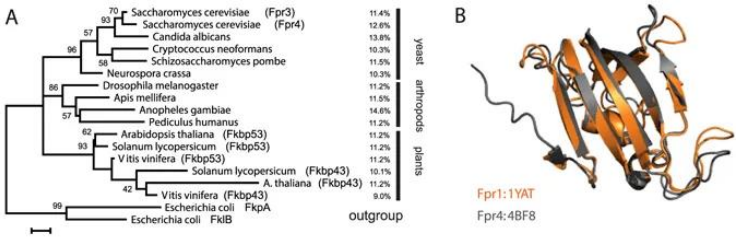
Conserved patch is evident in the interface binding chain B of Beta-2 Microglobulin (in green)

Peptide in green: Vesicular Stomatitis Virus Nucleoprotein (fragment 52-59, chain P)

Locally similar structure: Surface based methods

Electrostatic potential
- Possible binding/interaction sites





Leu et al., *SciRep* 2017

The chromatin binding ability of 'basic' FKBPs is shared amongst related orthologues. This ability is mediated by a collection of **basic patches** that enable the enzyme to stably associate with linker DNA

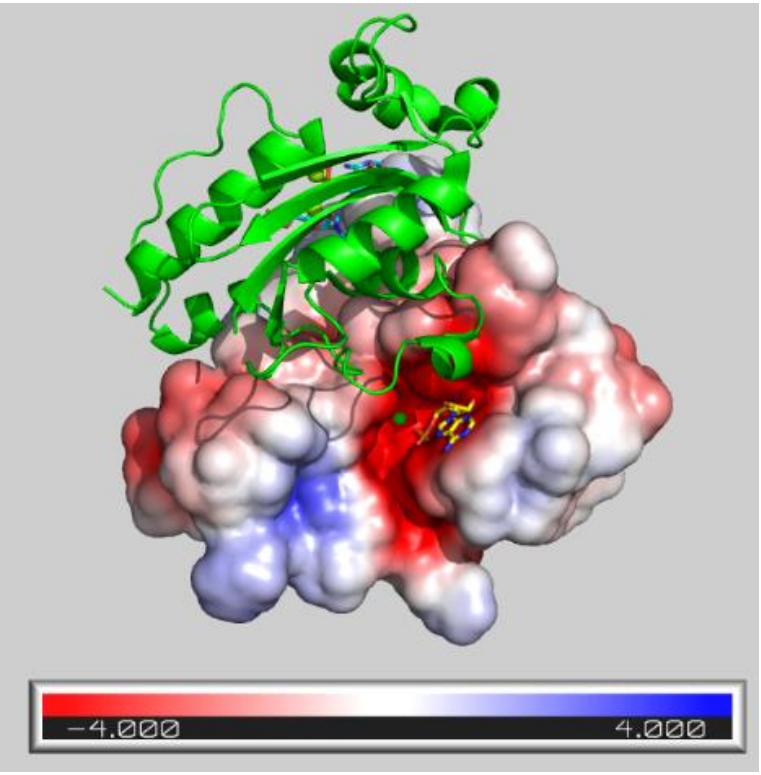
DelPhi web-server http://compbio.clemson.edu/sapp/delphi_webserver/
An online Poisson-Boltzmann solver for calculating **electrostatic energies** and potential in biological macromolecules. Originally developed in Dr. Barry Honig's lab and now maintained by Delphi Development team (Dr. Emil Alexov Group). Smith N et al. *Bioinformatics* 2012; Sarkar S et al. *Comm. Comp. Phys.* 2013

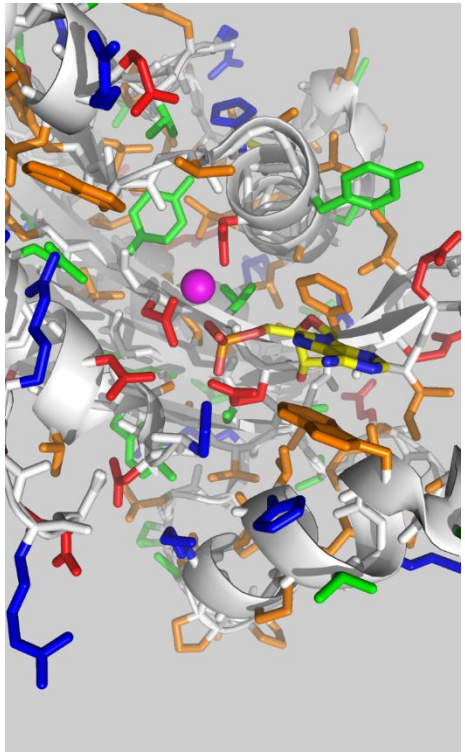
APBS (Adaptive Poisson-Boltzmann Solver) and **PDB2PQR** are software packages designed to analyze the **solvation properties** of small molecules as well as macro-molecules such as proteins, nucleic acids, and other complex systems. You can perform **electrostatics calculations** on your biomolecular structure of interest and easily visualize them in **Pymol**.
http://nbcr-222.ucsd.edu/pdb2pqr_2.1.1/

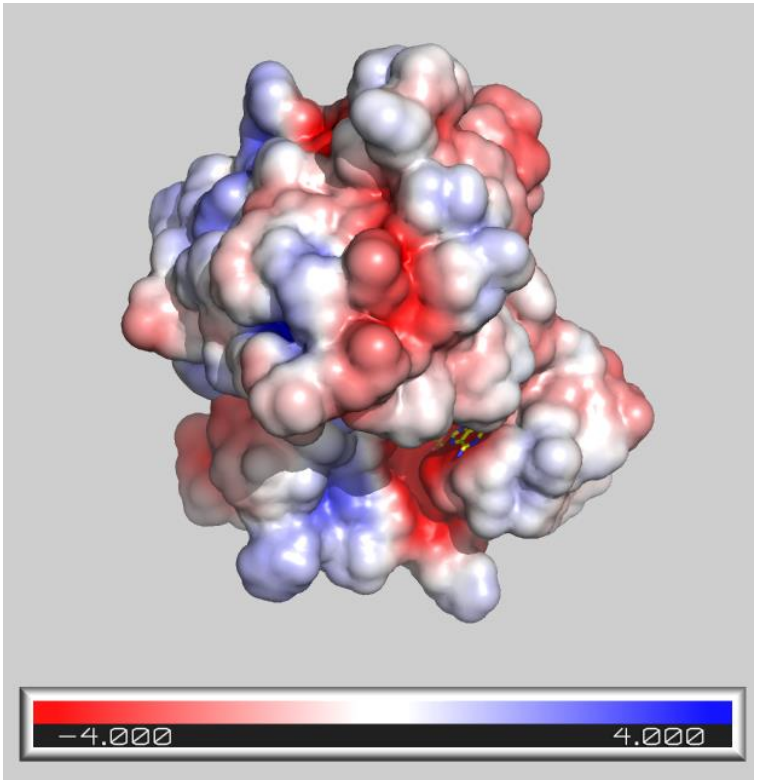
Locally similar structure: Surface based methods

4oke electrostatic potential visualized on the molecular surface by using APBS Tools in PyMol

Mg+



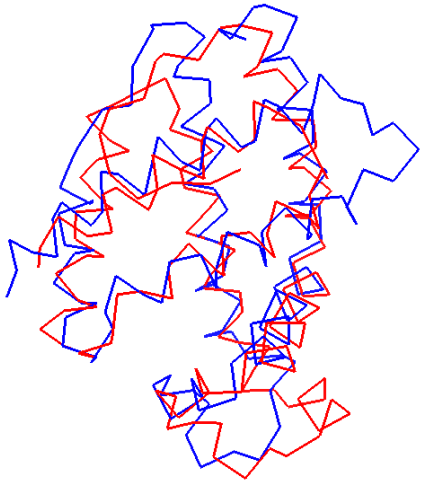




Structure Alignment

- In the same way that we align sequences, we wish to align structure
- To make it simple: How to score an alignment
  - Sequences: E.g. percentage of matching residues
  - Structures: **Rmsd** (root mean square deviation)

- **DALI** (**Distance-matrix-ALIgnment**) is one of the first tools for structural alignment



- How does it work?
  - **Atoms:**
    - Given two structures' atomic coordinates
  - **Compute two distance matrices:**
    - Compute for each structure all pairwise inter-atom distances.
  - **Align two distance matrices**:
    - Find small (e.g. 6x6) sub-matrices along diagonal that match
    - Extend these matches to form overall alignment

Structure Alignment

http://ekhidna2.biocenter.helsinki.fi/dali/

The Dali server is a network service for comparing protein structures in 3D.

You can perform four types of structure comparisons:

- Heuristic **PDB search** - compares one query structure against those in the Protein Data Bank
- Exhaustive **PDB25** search - compares one query structure against a representative subset of the Protein Data Bank
- **Pairwise** structure comparison - compares one query structure against those specified by the user
- **All against all** structure comparison - returns a structural similarity dendrogram for a set of structures specified by the user

Holm L, *Bioinformatics* 2019

Structure Metaserver: ProFunc

ProFunc

http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/

Sequence scans

- Sequence search *vs* PDB
- Sequence search *vs* Uniprot
- Sequence motifs (PROSITE, BLOCKS, SMART, Pfam, etc)
- Superfamily HMM library
- Gene neighbours

Fold and structural motifs

- SSM fold search
- Surface clefts
- Residue conservation
- DNA-binding HTH motifs
- Nest analysis

n-residue templates

- Enzyme active sites
- Ligand binding sites
- DNA binding sites
- Reverse templates

Laskowski RA, *Methods in Mol. Biol.* 2017

From Structure to Function: the case of a protein essential for *Mycobacterium tuberculosis* virulence

What we knew:

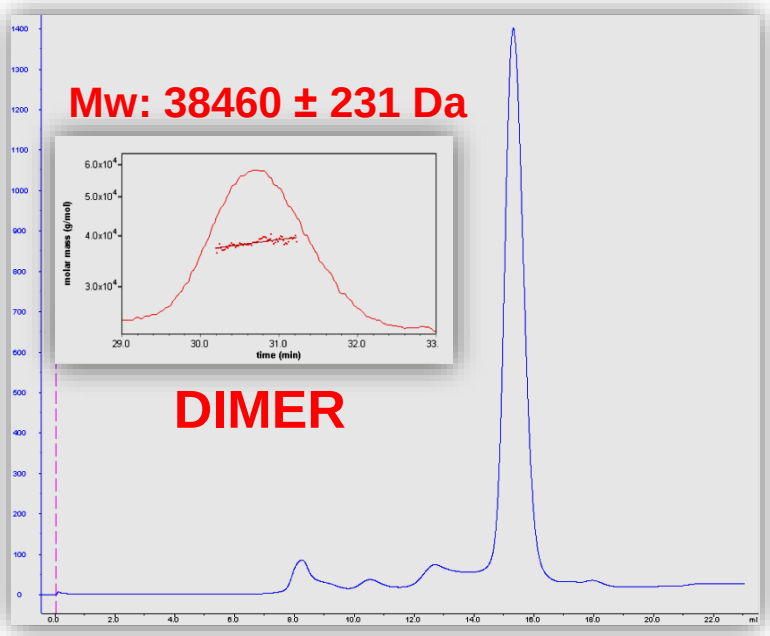
- Rv2179c (just a code, say **Mrs X**) aminoacid sequence

- **Mrs X** has effect on bacterial growth

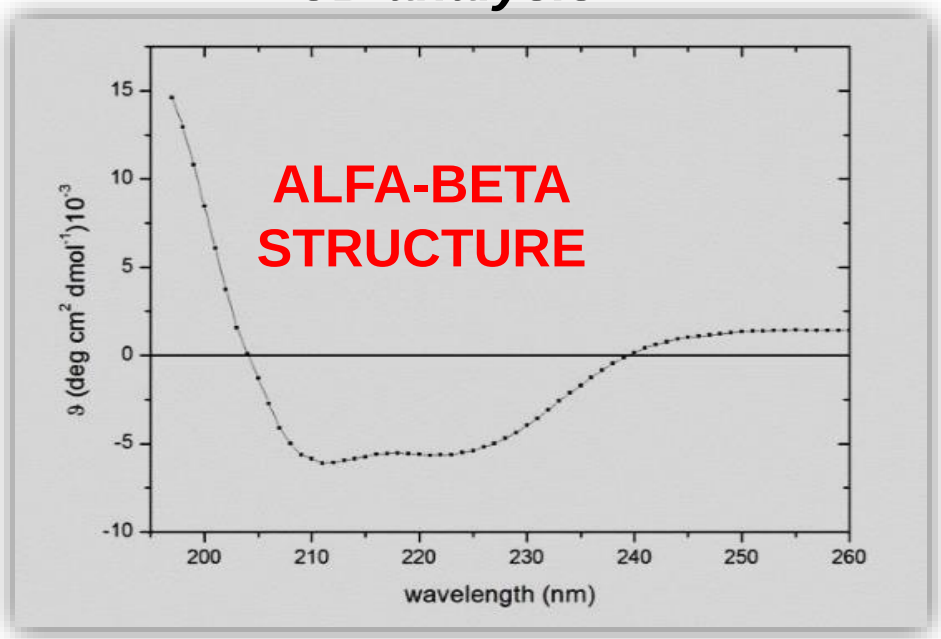
- **Mrs X** has a strong impact on bacterial virulence *in vivo.*

<u>Structural features in solution</u>

SEC-LS analysis

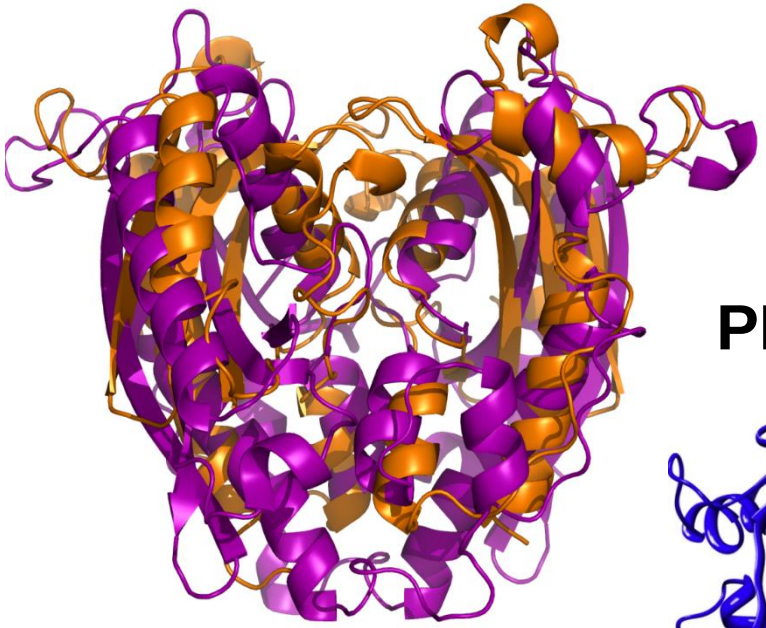


CD analysis



From Structure to Function: the case of a protein essential for *Mycobacterium tuberculosis* virulence

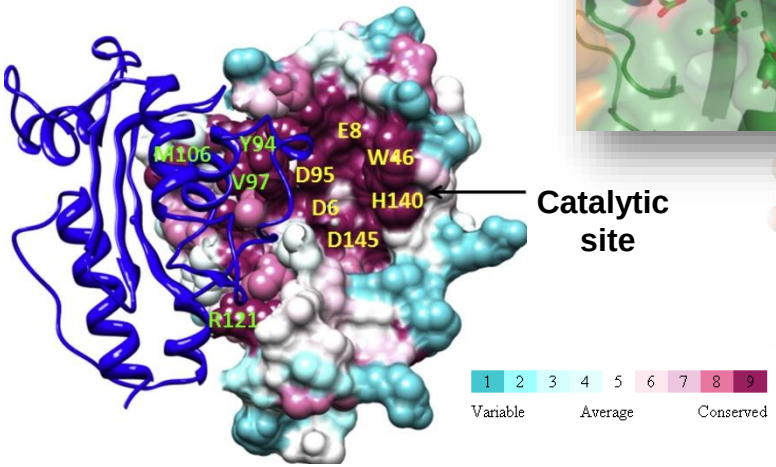
The structure was solved by single-wavelength anomalous dispersion (SAD) analysis of Europium-derivatized crystals and refined to a resolution of **2.1 Å**

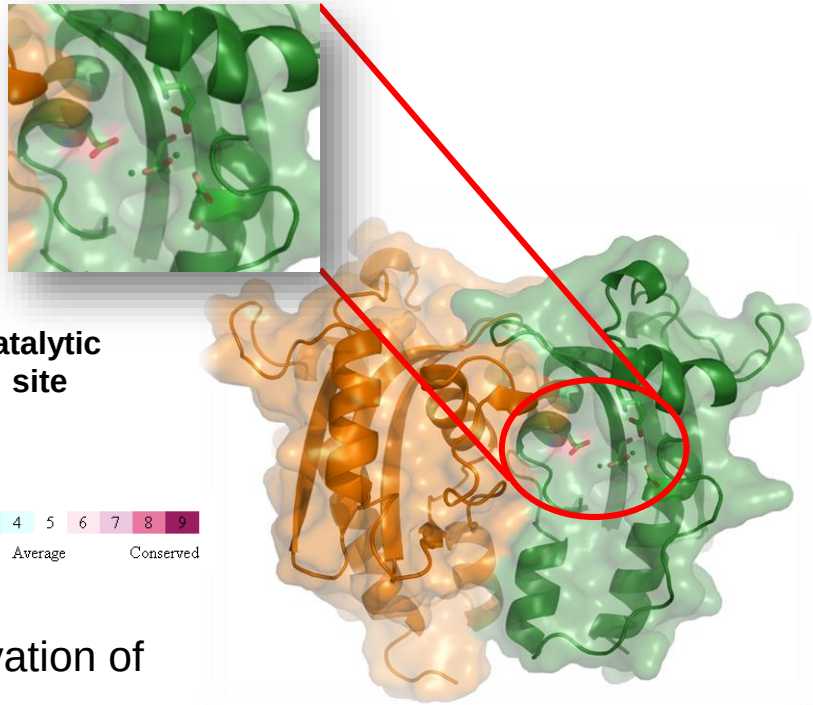


Structural comparisons: high structural similarity with E. coli RNase T

Only low sequence identity (13%)

PDBcode 4oke





Dali server: superposition

E. Coli RNase T

M. Tub. RNase AS

DEDD superfamily, a large family of 3′-5′ exonucleases

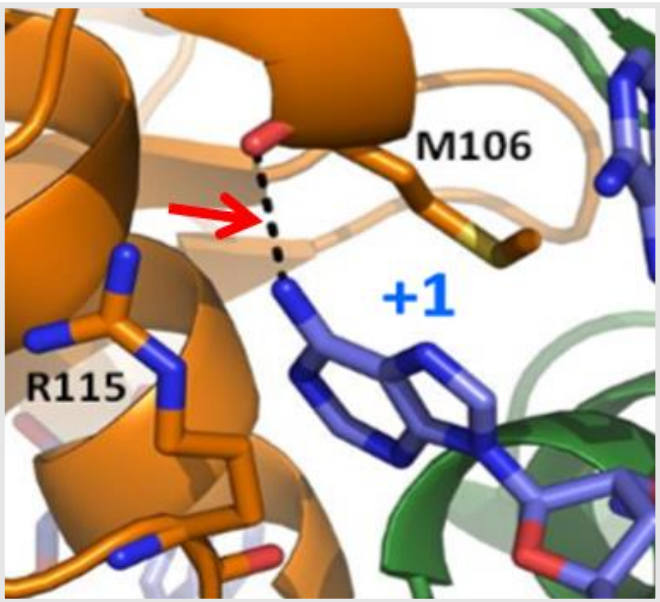
ConSurf: conservation of catalytic site

Romano M et al., *Structure* 2014

From Structure to Function: the case of a protein essential for *Mycobacterium tuberculosis* virulence

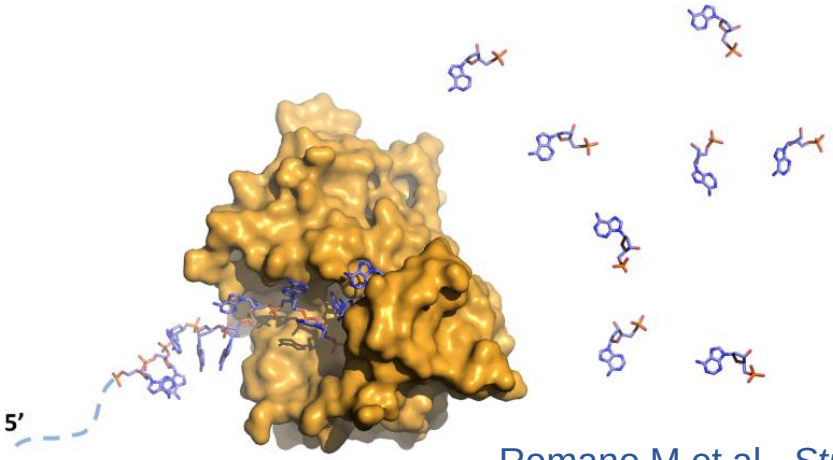
The protein specifically hydrolyses poly(A) → Mrs X= RNase AS
In vitro

The structure is useful to rationalize this behaviour: NH2 of adenine is a strict requirement as a hydrogen bond donor



Functional Hypothesis:
in the absence of RNase AS, tRNA is not de-adenylated → **Cell death** due to the blockage of protein biosynthesis

The addition of poly(A) tails promotes instability in prokaryotes

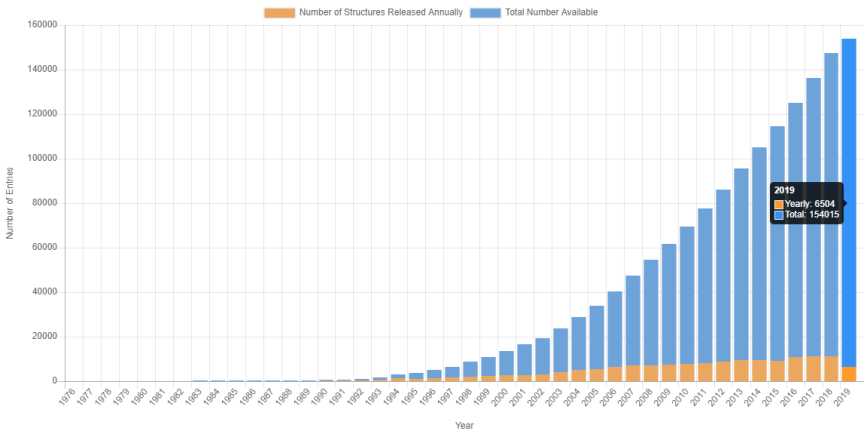


Romano M et al., *Structure* 2014

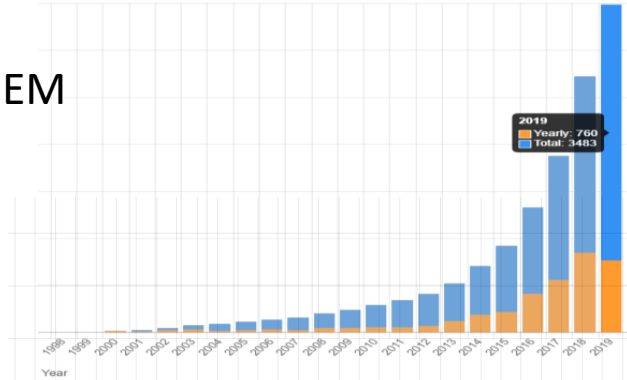
Corroboration: RNase AS Does Not Degrade Polyinosine, whose nucleobase holds all characteristics of AMP but has in its predominant keto form an oxygen in place of the adenine NH2 group

The Growth of PDB structures

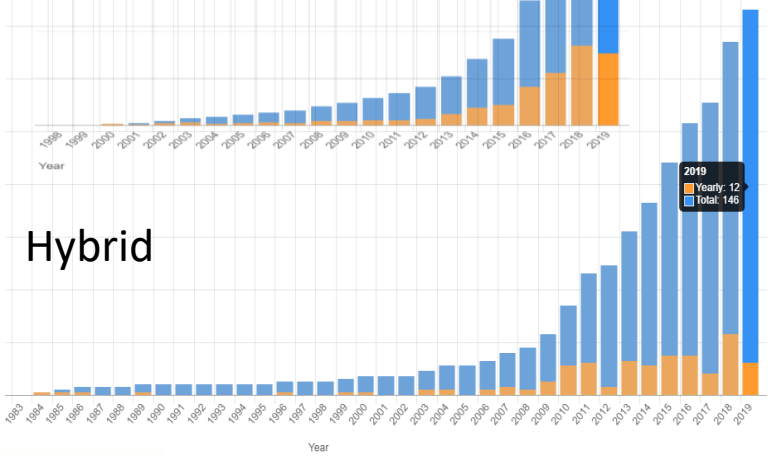
All: X-ray, NMR. EM, Hybrid



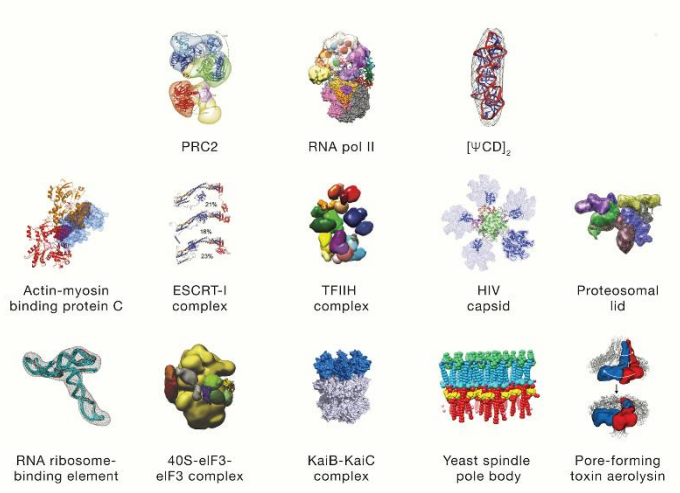
EM

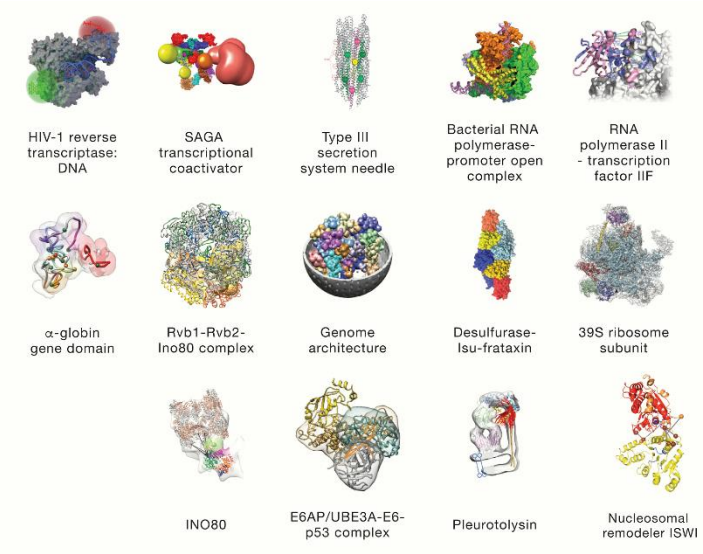


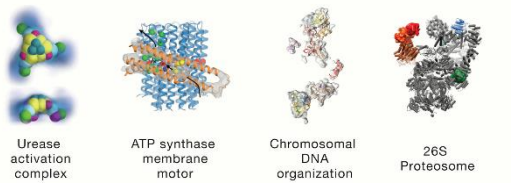
Hybrid



The Growth of complexity







Rout & Sali, *Cell* 2019

Integrative Structural Biology

Integrative structural biology combines different experimental techniques with computational modeling to build structural models of challenging macromolecular systems.

Motivation: Any system is described best by using all available information about it

Table 1. Example Methods that Are Informative about a Variety of Structural Aspects of Biomolecular Systems

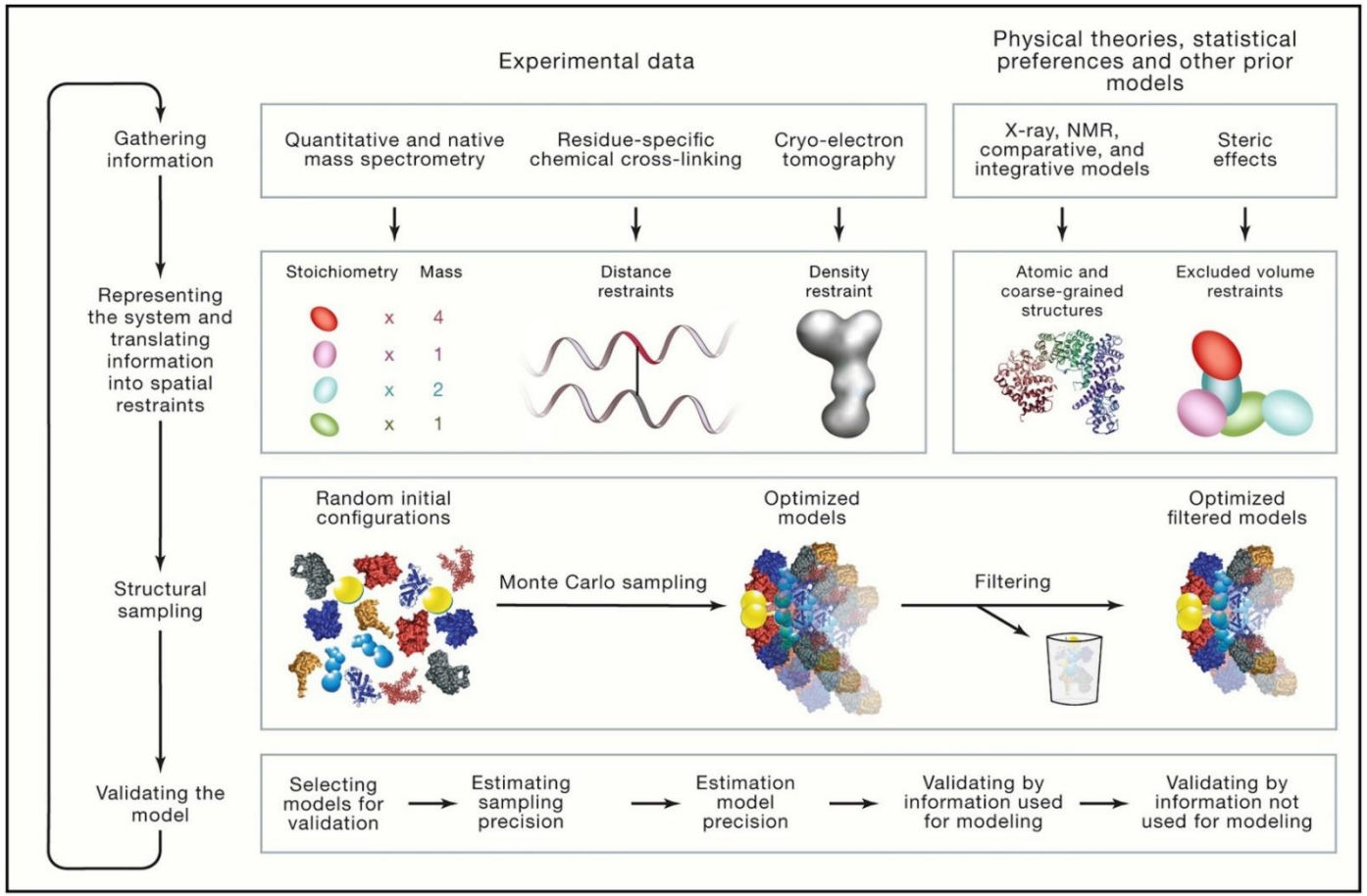
Structural information	Method
Stoichiometry	MS, quantitative fluorescence imaging
Atomic structures of parts of the studied system	X-ray and neutron crystallography, NMR spectroscopy, 3DEM, comparative modeling, and molecular docking
3D maps and 2D images	Electron microscopy and tomography
Atomic and protein distances	NMR, FRET, and other fluorescence techniques; DEER, EPR, and other spectroscopic techniques; and XL-MS and disulfide bonds detected by gel electrophoresis
Binding site mapping	NMR spectroscopy, mutagenesis, FRET, and XL-MS
Size, shape, and distributions of pairwise atomic distances	SAS
Shape and size	Atomic force microscopy, ion mobility mass spectrometry, fluorescence correlation spectroscopy, fluorescence anisotropy, and analytical ultracentrifugation
Component positions	Super-resolution optical microscopy, FRET imaging, and immuno-electron microscopy
Physical proximity	Co-purification, native mass spectrometry, XL-MS, molecular genetic methods, and gene/protein sequence covariance
Solvent accessibility	Footprinting methods, including HDex assessed by MS or NMR, and even functional consequences of point mutations
Proximity between different genome segments	chromosome conformation capture
Propensities for different interaction modes	Molecular mechanics force fields, potentials of mean force, statistical potentials, and sequence co-variation

Abbreviation are as follows: 3DEM, 3D electron microscopy; DEER, double electron-electron resonance; EPR, electron paramagnetic resonance; FRET, Foerster resonance energy transfer; HDex, hydrogen/deuterium exchange; NMR, nuclear magnetic resonance; SAS, small-angle scattering; XL-MS, cross-linking mass spectrometry.

- Rout MP & Sali A. Principles for Integrative Structural Biology Studies. *Cell* 2019.
- Sali A, Earnest T, Glaeser R, Baumeister W. From words to literature in structural proteomics. *Nature* 2003.
- Ward A, Sali A, Wilson I. Integrative structural biology. *Science* 2013.

Integrative Structural Modeling Workflow

- ✓ Uses multiple types of information (experiments, physical theory, statistical inference).
- ✓ Maximizes accuracy, resolution, completeness, and efficiency of the structure determination.
- ✓ Finds all models whose computed data match the experimental data within an acceptable threshold

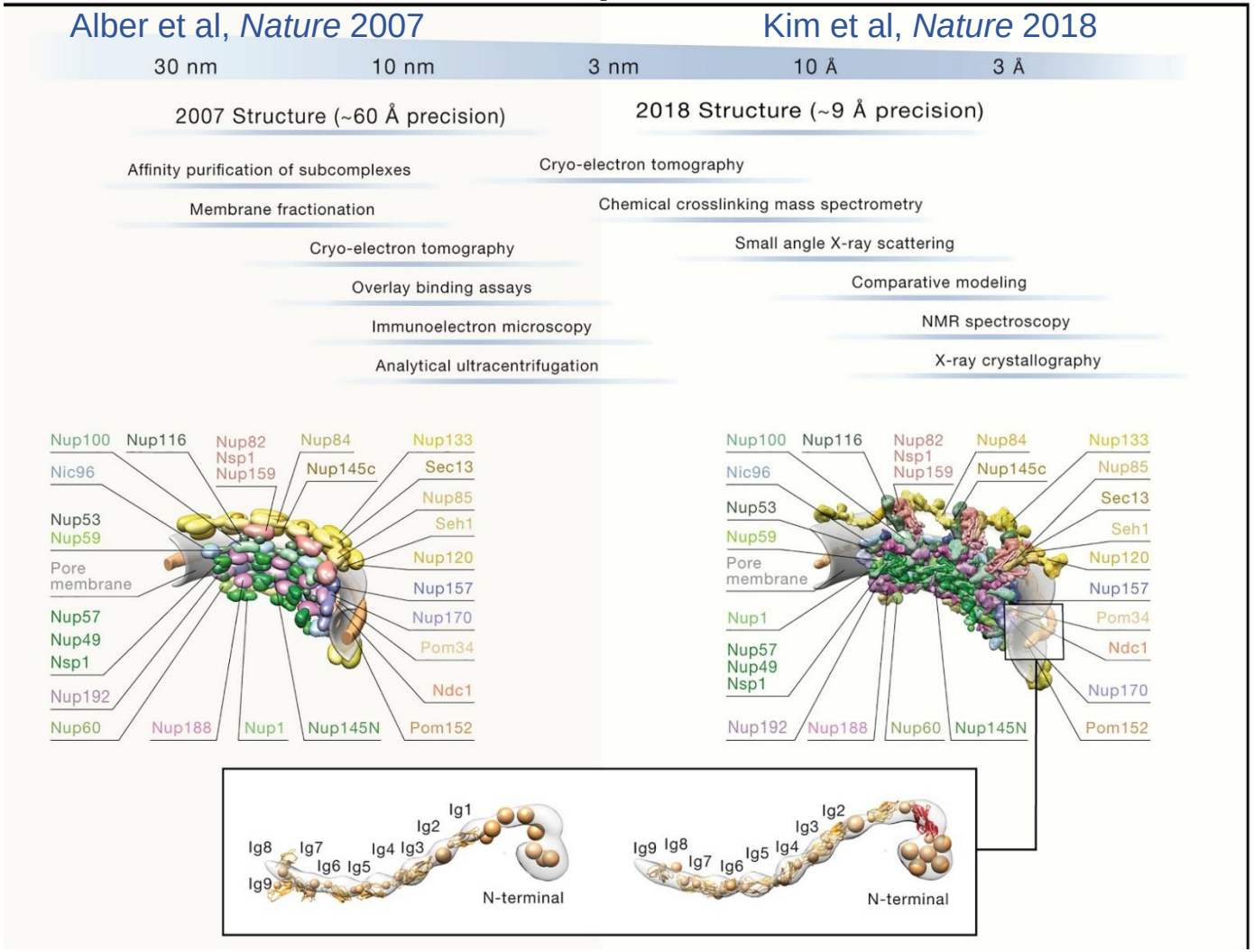


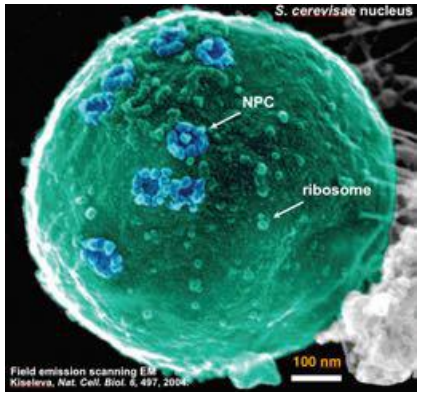
Rout & Sali, *Cell* 2019

The case of the Nuclear Pore Complex

NPC is a gatekeeper controlling the entry into and the exit from the nucleus of macromolecules

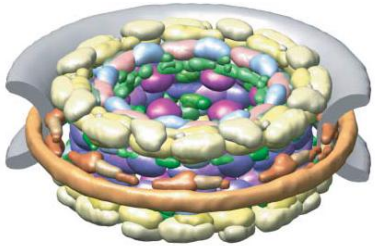
Iterative process

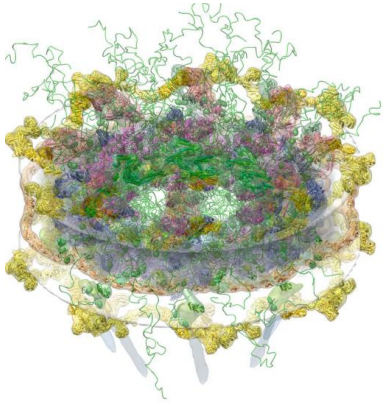




Consists of nucleoporins.
50 MDa complex: ~480 proteins of 30 different types.

2007 model





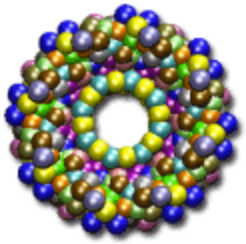
2018 model

Integrative Modeling Platform (IMP)

Software package

http://integrativemodeling.org

IMP, the Integrative Modeling Platform



IMP's broad goal is to contribute to a comprehensive structural characterization of biomolecules ranging in size and complexity from small peptides to large macromolecular assemblies, by integrating data from diverse biochemical and biophysical experiments. IMP provides an open source C++ and Python toolbox for solving complex modeling problems, and a number of applications for tackling some common problems in a user-friendly way. IMP can also be used from the Chimera molecular modeling system, or via one of several web applications.

D. Russel, et al, *PLoS Biol*, 2012.

Nascent wwPDB archive for integrative structures

PDB-Dev: A Prototype System for Depositing Integrative/Hybrid Structural Models

https://pdb-dev.wwpdb.org

Integrative structure and functional anatomy of a Nuclear Pore Complex

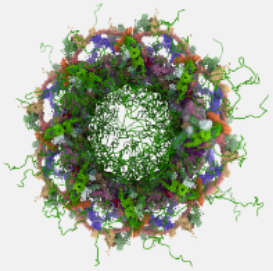
Structure of the 552-protein Nuclear Pore Complex (NPC) from yeast, determined by IMP using spatial restraints derived from cryo-electron tomography and chemical crosslinking experiments.

Publication: Kim SJ, Fernandez-Martinez J, Nudelman I, Shi Y, et al., Nature. 2018 Mar; 555(7697):475-82

Related resource: 10.5281/zenodo.1194547

Accession codes: PDBDEV_00000010, PDBDEV_00000011, PDBDEV_00000012

Download the structures: NPC single spoke, NPC three spokes, NPC eight spokes



Vallat et al, *Structure*, 2018.