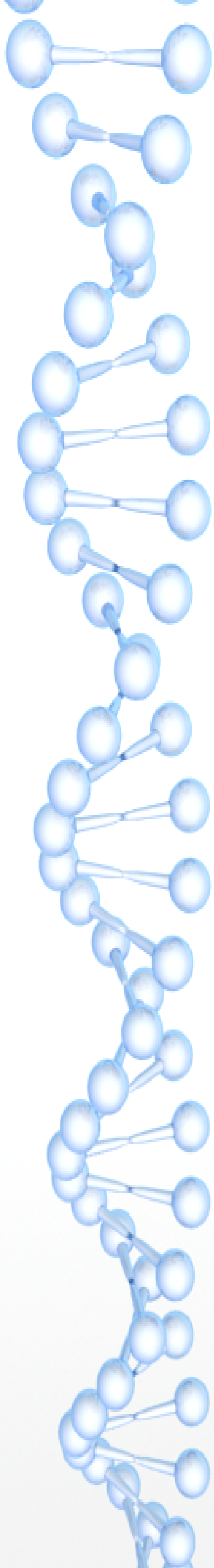


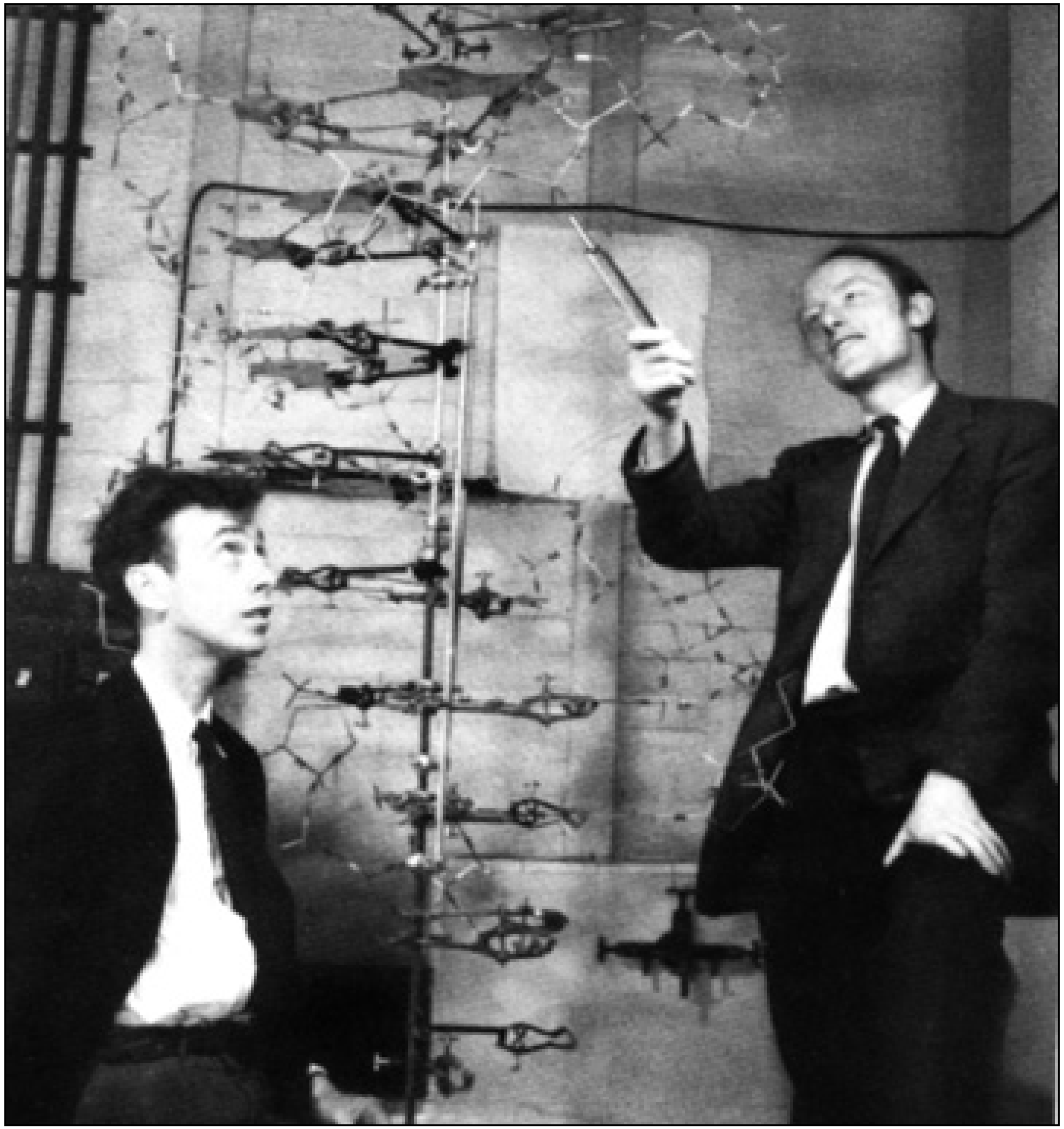
Structural bioinformatics

Dr. Zsuzsanna Dosztányi

MTA-ELTE Momentum Bioinformatics Group

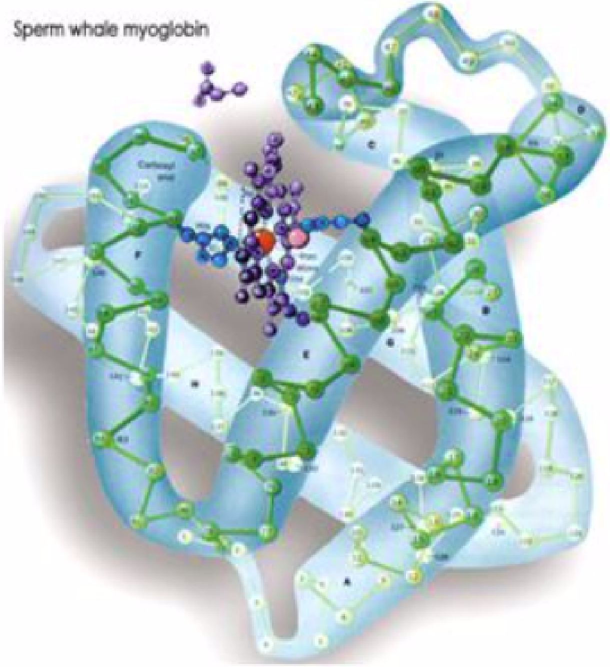
4. December 2017

- Basic features of protein structures
- Structure determination methods
- PDB database
- Visualization and analysis of structures
- Structure comparisons
- Structural classification
- Structure predictions



- **“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”**

- **1958**
  - John Kendrew et al., published the first structure of a globular protein, myoglobin.
  - " Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry"
- **1962**
  - Nobel prize in Chemistry was awarded to Max Perutz and John Kendrew.
- **Now**
  - ~80,000 structures in protein database (PDB)



MOLECULAR MACHINERY: A Tour of the Protein Data Bank

Living cells are filled with complex molecular machinery, a million times smaller than familiar machines like computers or automobiles. Cells use these tiny molecular machines to perform all of the jobs needed for life. Some are molecular scissors that cut food into cell-sized pieces. Some build new molecules when cells grow or when damaged tissues are repaired. Some are molecular beams and muscles that support cells and help them move and crawl. Some fight off attackers, defending against infection.

Researchers around the world are studying these molecules and determining their precise atomic structures. These structures are available on the Internet through the Protein Data Bank (http://www.pdb.org), the central storehouse of biomolecular structures. A few of the thousands of structures held in the Protein Data Bank are shown here. In these pictures, the molecules are all drawn at a magnification of 3,000,000 times, and each atom is shown as a small sphere. Many of these structures are composed of several subunits, which are indicated by different colors. An enormous range of sizes is shown here: the water molecule at the left has only three atoms and the rhinovirus shown below has hundreds of thousands.

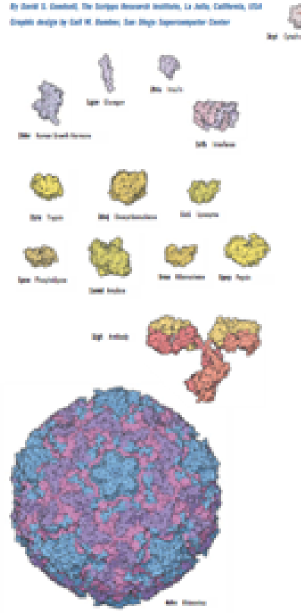
Water

Glucose

Adenosine Triphosphate (ATP)

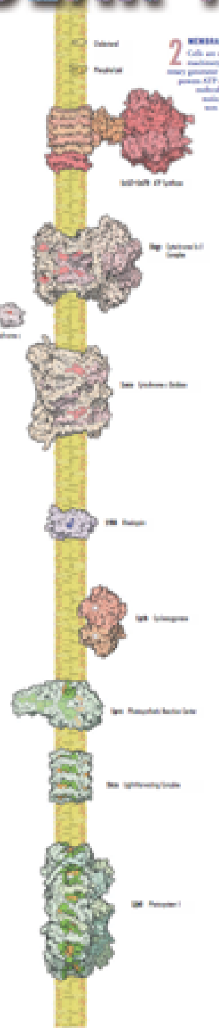
By David S. Goodsell, The Scripps Research Institute, La Jolla, California, USA

Graphic design by Gail W. Bamber, San Diego Supercomputer Center



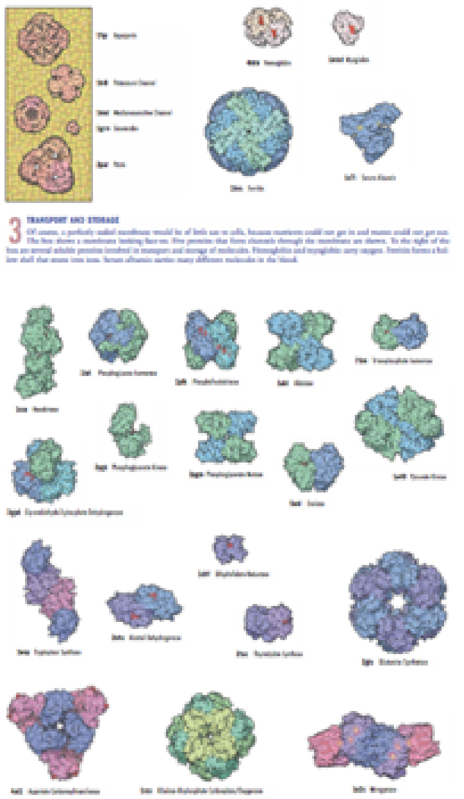
1 OUTSIDE THE CELL

Some molecular machines perform their jobs outside of cells. Many are compact, so that they can diffuse quickly to their site of action. This is true of the four hormones shown at the top: insulin and glucagon, which together regulate blood sugar levels; interferon, which carries signals in the immune system; and human growth hormone. The seven digestive enzymes (in yellow) are also small and very stable, so that they can survive the hostile environment in the digestive tract. Each of these enzymes has a small groove (oriented towards the top in each) that binds to a different target molecule and digests it. At the bottom is rhinovirus, the virus that causes the common cold, and an antibody, our major defense against viruses. Antibodies bind to viruses and prevent them from binding to cell surfaces, thus blocking infection.



2 MEMBRANES

Cells are surrounded by a membrane made of lipids, like the phospholipid and cholesterol molecules shown at the top. Membranes keep the cellular machinery inside and unwanted material out. Many proteins are embedded in this membrane, performing a variety of essential tasks. ATP synthase is a rotary generator that produces ATP (adenosine triphosphate), the small molecule used for powering cells. The two large complexes below it charge a battery that powers ATP synthase, and the tiny protein cytochrome c shuttles electrons between them. Rhodopsin is found in membranes in the retina. The small retinal molecule inside of it changes shape when illuminated, causing the surrounding protein to send a signal to the brain. Cyclooxygenase builds one of the molecules used to signal pain—the cyclooxygenase molecule here, however, is blocked by two molecules of aspirin, shown inside in white. At the bottom are three molecules involved in photosynthesis, which capture energy from light and use it to power the synthesis of sugar in plant cells.

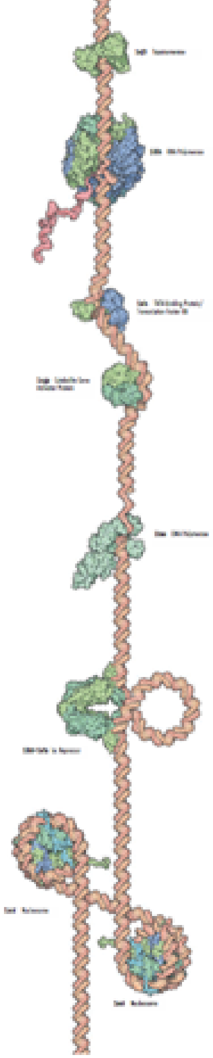


3 TRANSPORT AND STORAGE

Of course, a perfectly sealed membrane would be of little use to cells, because nutrients could not get in and wastes could not get out. The box shows a membrane looking face on. Five proteins that form channels through the membrane are shown. To the right of the box are several soluble proteins involved in transport and storage of molecules. Hemoglobin and myoglobin carry oxygen. Ferritin forms a hollow shell that stores iron ions. Serum albumin carries many different molecules in the blood.

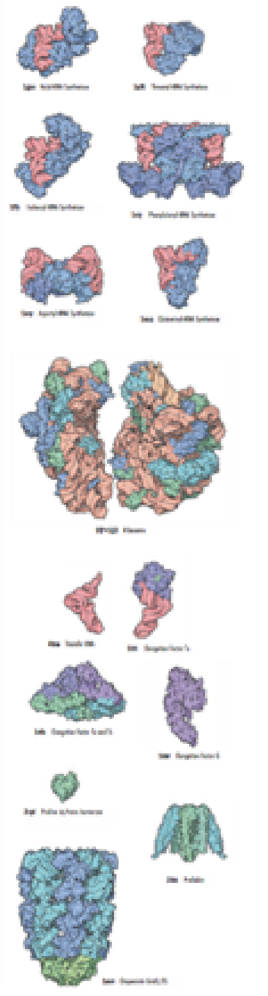
4 CHEMICAL FACTORIES

Cells build a bewildering variety of enzymes—proteins that perform chemical reactions. At the top are the ten enzymes that perform glycolysis, the breakdown of sugar to form ATP. Below that are several enzymes that perform different housekeeping reactions. Dihydrofolate reductase activates a key cofactor molecule and alcohol dehydrogenase breaks down alcohol. Ribulose bisphosphate carboxylase/oxygenase is the most common enzyme on the Earth, and performs a key step in the capture of carbon dioxide by plants to form sugar. The three synthases and the transferase make different building blocks for creating new molecules. Nitrogenase performs an essential role in the ecosystem by converting nitrogen gas into a form that living cells can use.



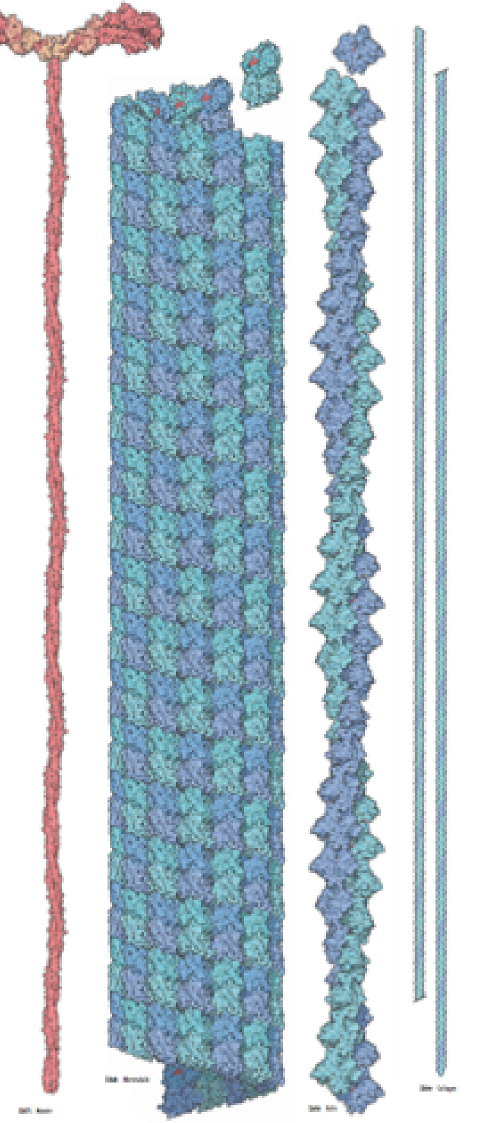
5 DNA

Genetic information is stored in the DNA double helix, seen running from top to bottom here. Many proteins are used to copy, read, and store this information. RNA polymerase copies the information into a strand of RNA that will be used to direct the construction of new proteins. It is assisted by topoisomerase, which releases tension when the helix is wound and unwound, and guided to appropriate starting points by the two protein complexes below it. DNA polymerase replicates DNA strands—here the polymerase is filling a gap in the double helix. Some proteins, like the lac repressor, grab DNA and bend it sharply, or even wrap it all the way around themselves, like the two nucleosomes at the bottom.



6 BUILDING NEW PROTEINS

New proteins are built by ribosomes—complex molecular factories that read the genetic code and use it to direct construction. Many molecular machines are needed to assist the process. Twenty different aminoacyl-tRNA synthetases (six are shown here) load the building blocks onto tRNA, ready to be added to a growing protein chain. Several protein factors, shown below the ribosome, guide each tRNA into the proper spot. The three chaperone proteins shown at the bottom help each new protein fold into its proper shape.



7 BEAMS AND GIRDERS

Cells are braced and supported by a complex infrastructure. This cytoskeleton is formed of sturdy filaments like actin and microtubules, composed of many subunits stacked like bricks. Myosin is a molecular motor that climbs along actin filaments, allowing the cell to move. Collagen, broken into two pieces here, is actually found outside of cells, where it forms connective tissue between cells.

PDB
PROTEIN DATA BANK

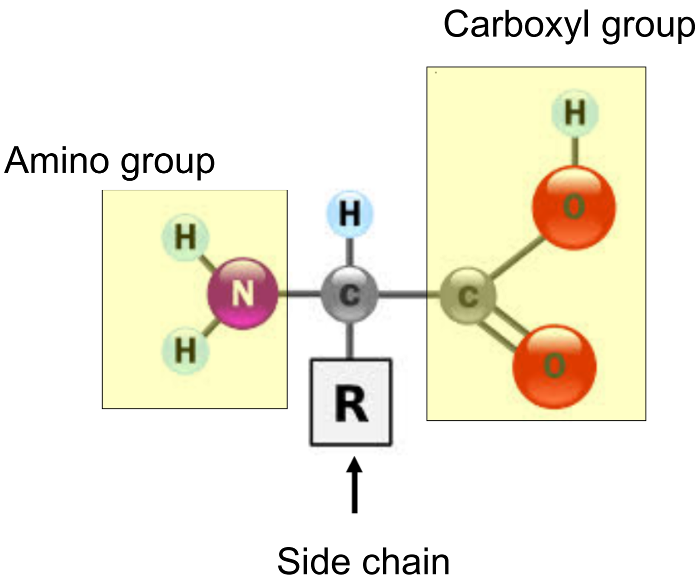
http://www.pdb.org/ • info@rcsb.org

Research Collaboratory for
Structural Bioinformatics

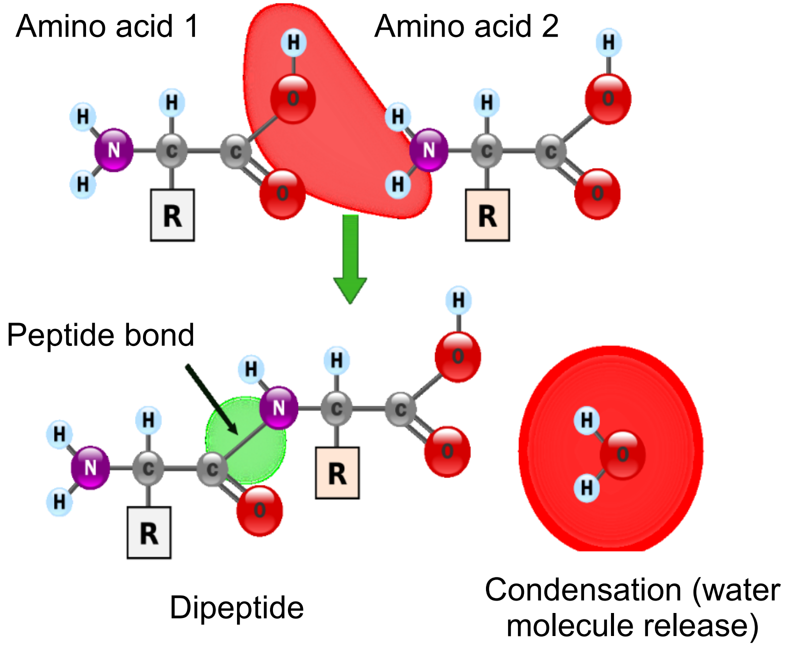
RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY
SAN DIEGO SUPERCOMPUTER CENTER
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

[illegible]

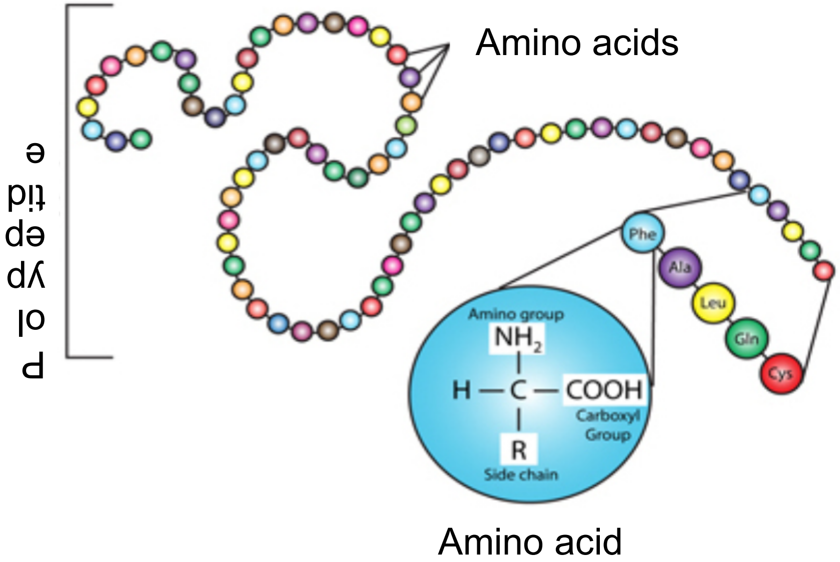
Amino acid



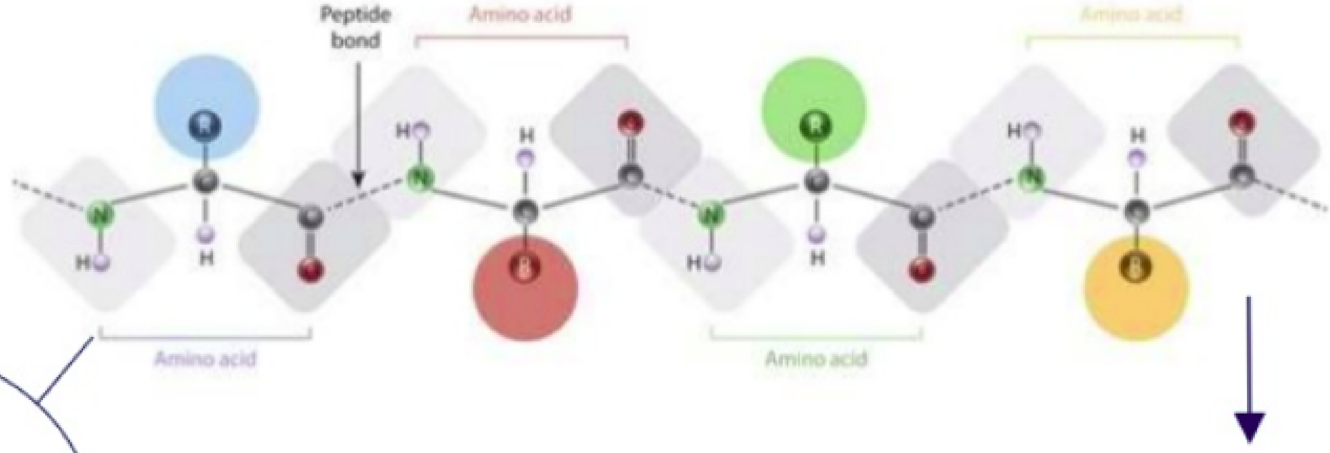
Peptide bond

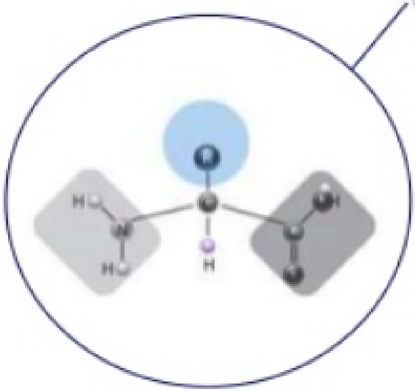


Polypeptide chain



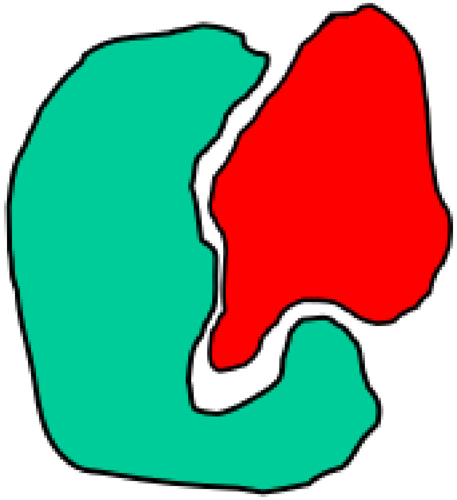
Globular proteins





They adopt a well-defined compact structure

Why are protein structures interesting?



Function is heavily dependent on the shape of the protein

- Atomic-level understanding of biological processes (DNA, RNA, enzymes, hormones, receptors)
- Understanding the molecular basis of diseases
- Drug design, protein-drug interactions

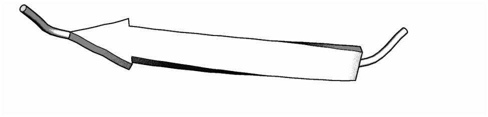
✖ No information on e.g. binding strength

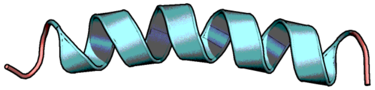
Levels of structure for globular proteins

Primary structure =
Amino acid sequence

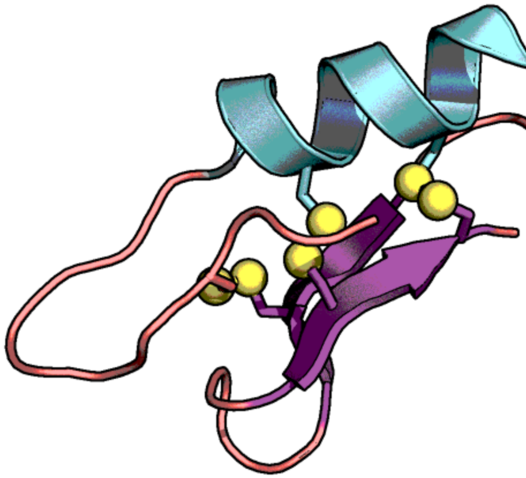
MSSVLLGHIKKLEMGHS…

Secondary structure =
alpha helix, beta sheets/strands, turns (based on main chain H-bonds)





Tertiary structures =
Relative positions of secondary structure elements within the chain



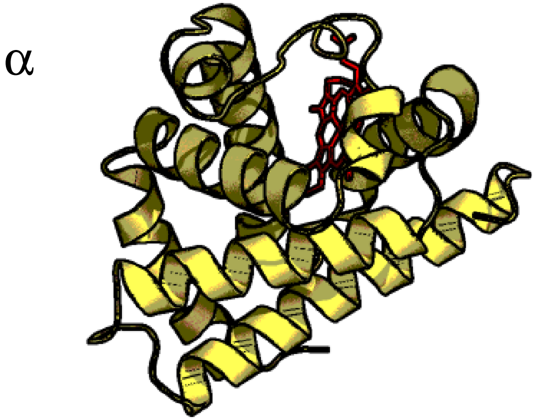
Quaternary structure

Monomer

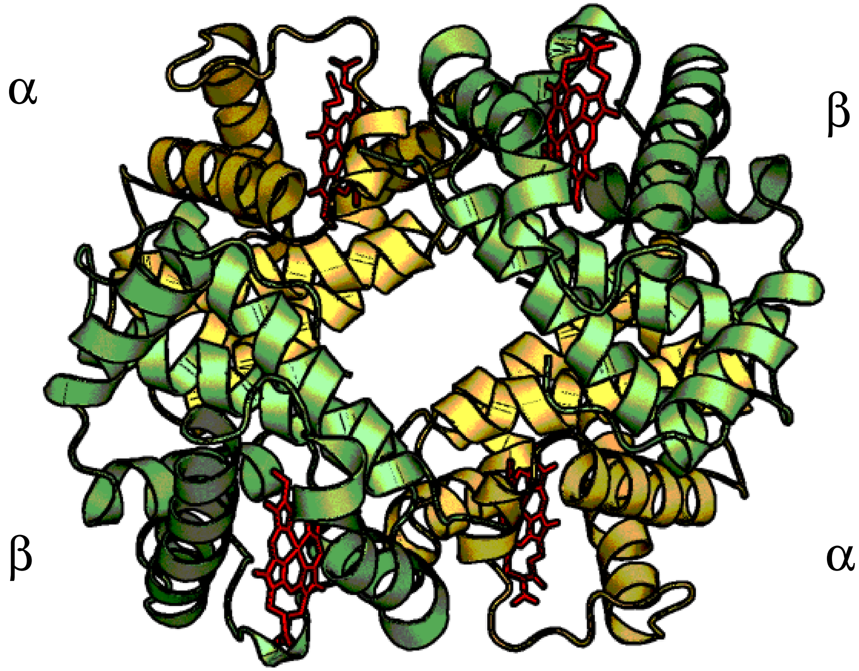
Multi subunit protein complexes

- Homo and hetero oligomers

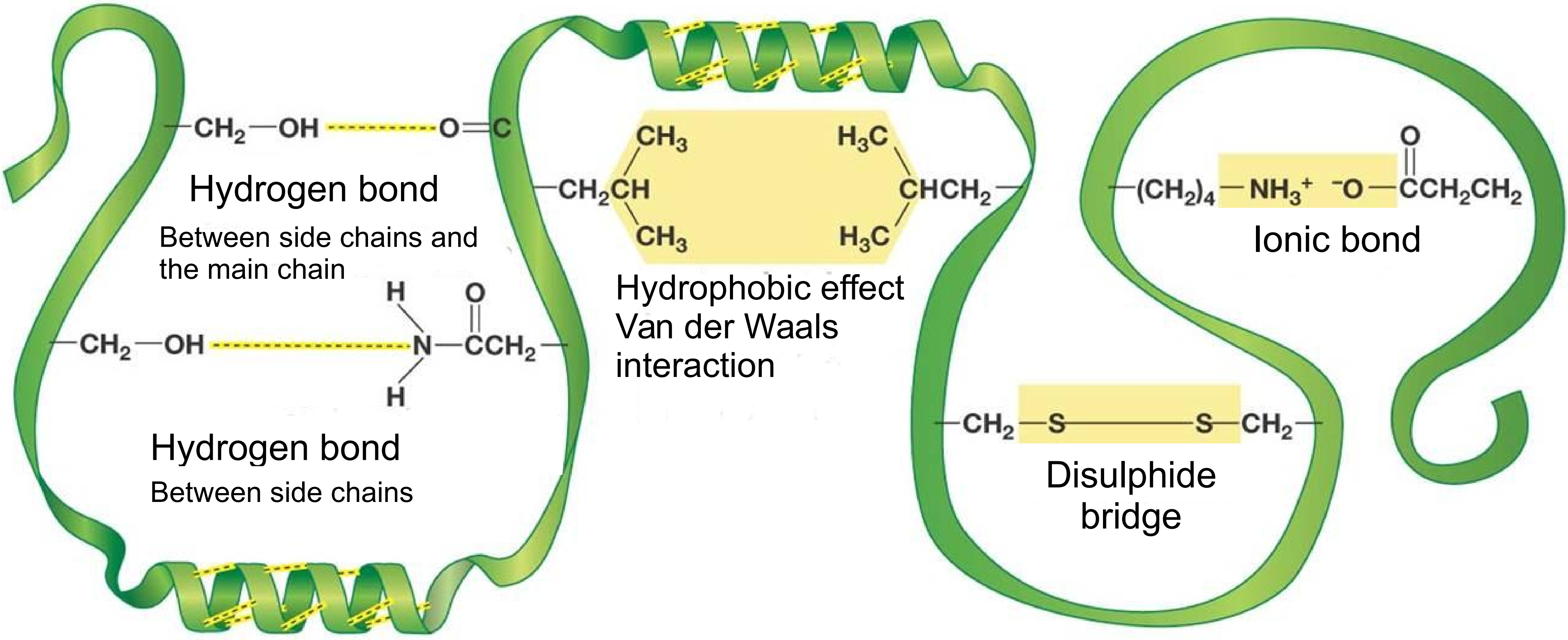
Myoglobin



Hemoglobin



Interactions stabilizing proteins


–CH2—OH ------------ O=C
–CH2CH(CH3)(CH3)   H3C, H3C CHCH2–
–(CH2)4—NH3+ –O—C(=O)CH2CH2
H
–CH2—OH ------------ N—C(=O)CH2–
H
–CH2—S——S—CH2–

Hydrogen bond

Between side chains and the main chain

Hydrogen bond

Between side chains

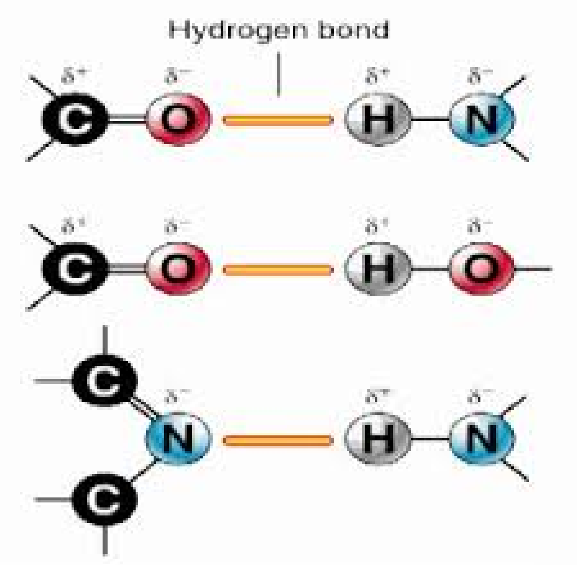
Hydrophobic effect
Van der Waals
interaction

Ionic bond

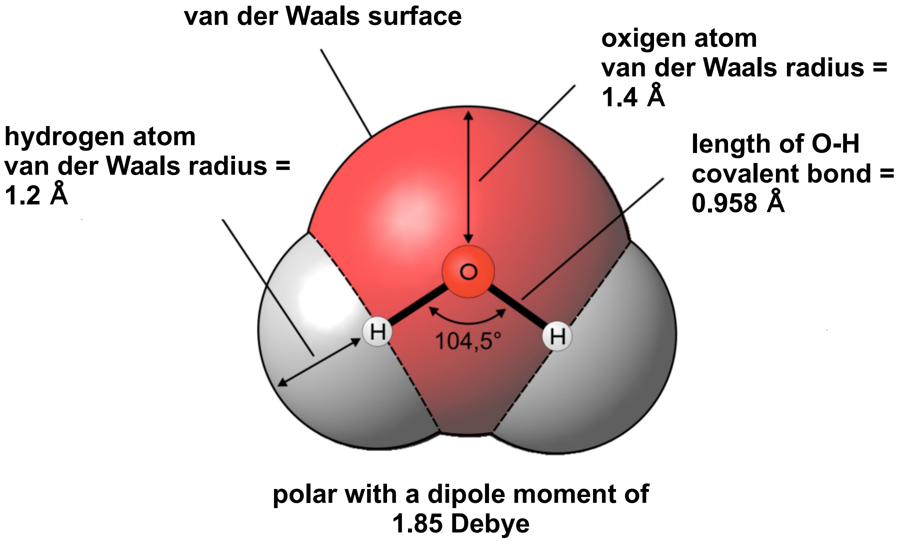
Disulphide
bridge

Hydrogen bond

Hydrogen bonds are formed by a H-atom bound in the structure with a **high electronegativity atom** (F, N, O) from a different functional group, i.e. a hydrogen atom establishes a bond between two other atoms.



Water molecule



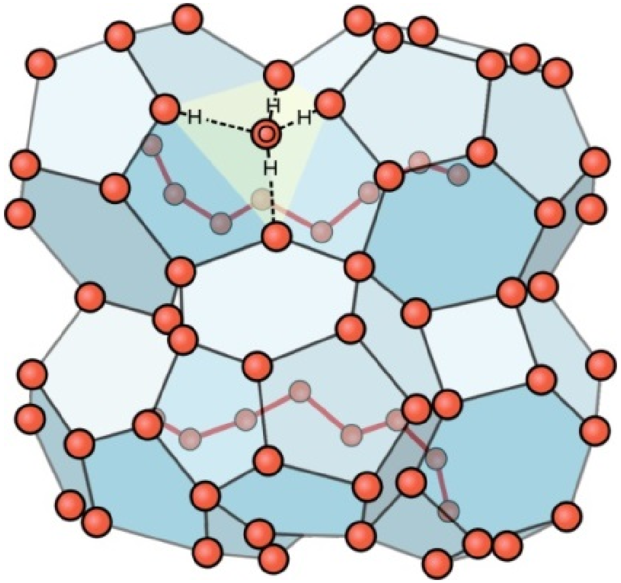
van der Waals surface

oxigen atom
van der Waals radius =
1.4 Å

hydrogen atom
van der Waals radius =
1.2 Å

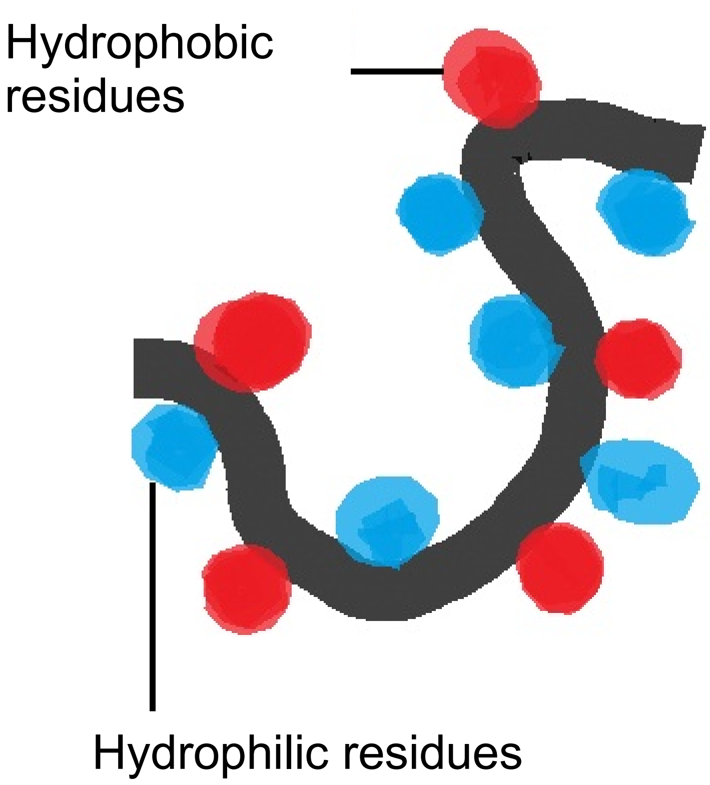
length of O-H
covalent bond =
0.958 Å

polar with a dipole moment of
1.85 Debye

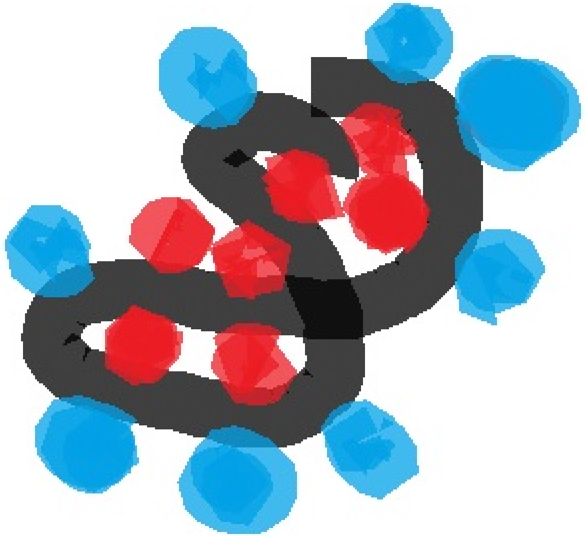


Hydrophobic effect: dominated by entropic terms

Hydrophobicity

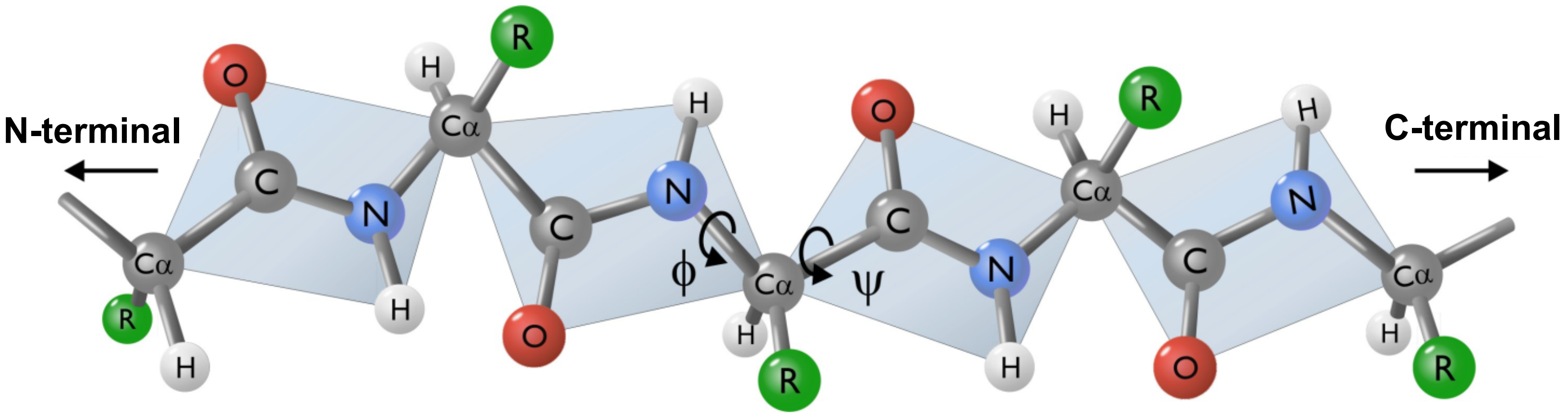


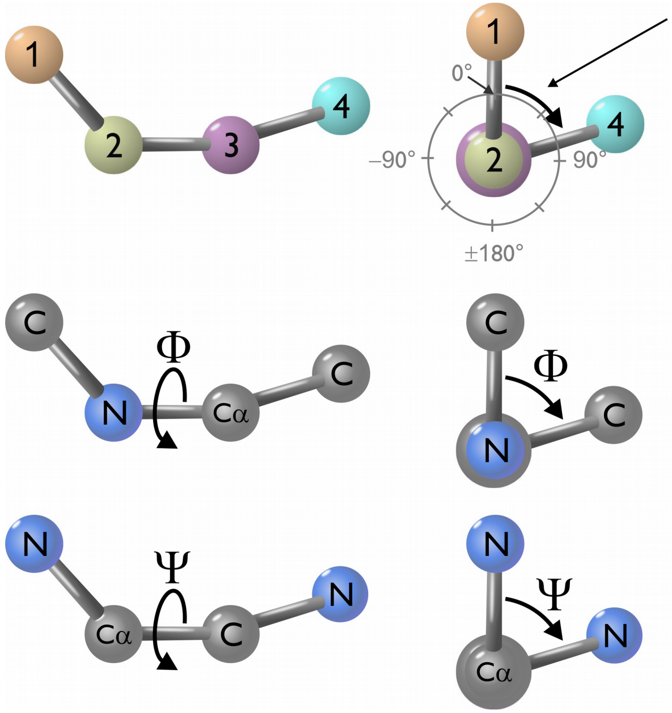
Protein in isolation



Protein in aquaneous environment

Main chain conformation





Torsion angle
clockwise (+)
counter-clockwise (-)

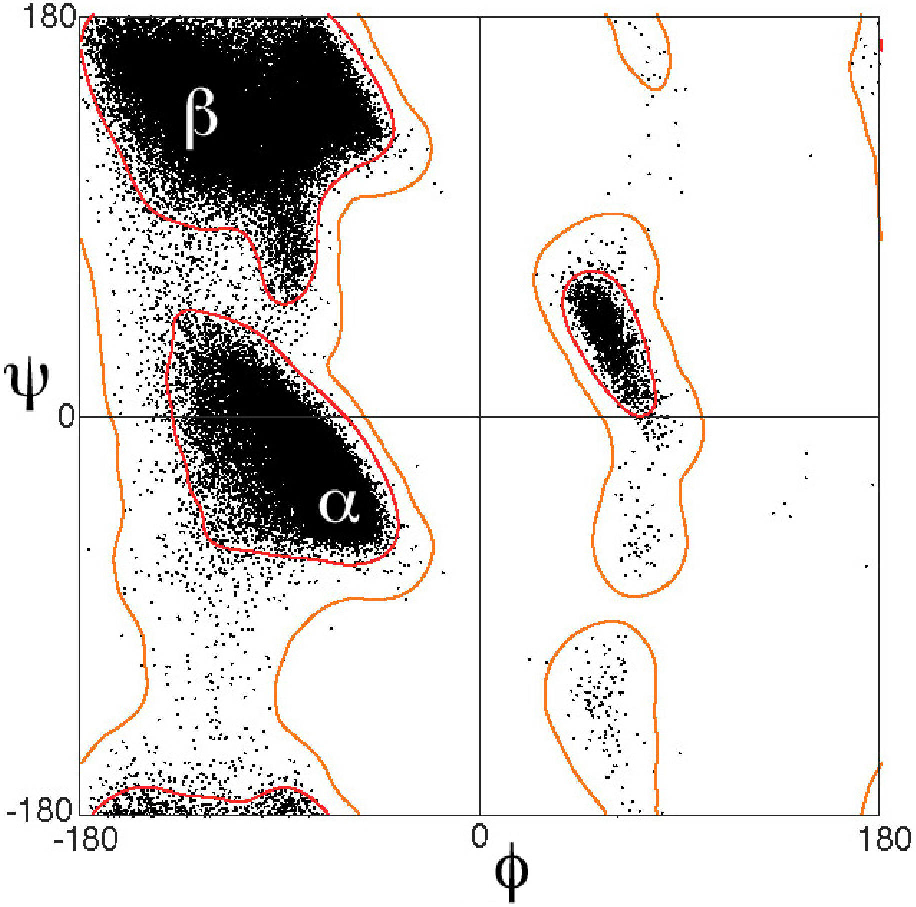
Rotation around N-Cα bond: phi

Rotation around Cα-C' bond: psi

The main chain φ and ψ torsion angles of a protein cannot take arbitrary values, there are preferred conformations.

Ramachandran plot

We can plot the angle-pairs of all residues in a coordinate system using the two torsion angles as X and Y-coordinates.



Glycines and prolines are typically left out from the plot as they have unique conformational preferences. As expected, most residues fall into regions corresponding to α-helices and β-strands, but most residues from bends and turns are also within the allowed regions.

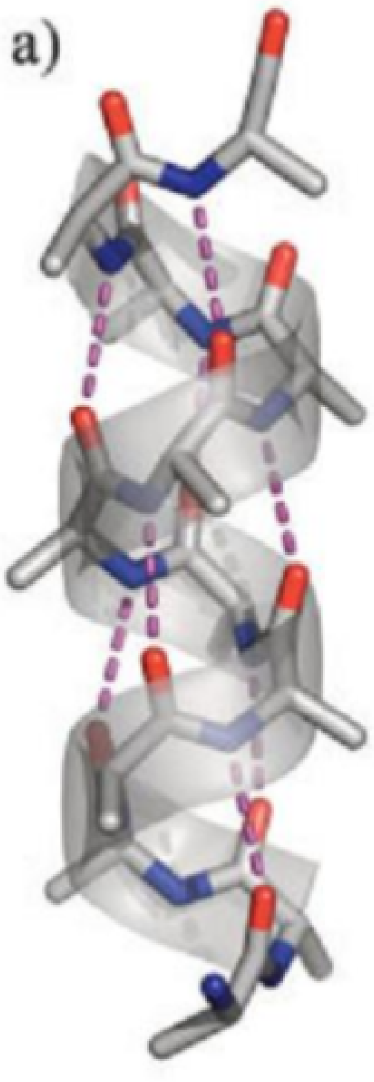
The (right-handed) α helix

Approx. 30% of globular proteins

5-40 residues in length (10 on average)

Individual H-bonds are relatively weak, they have a significant contribution to helix stability

The helix-forming propensity of a peptide segment depends on its sequence



α helix

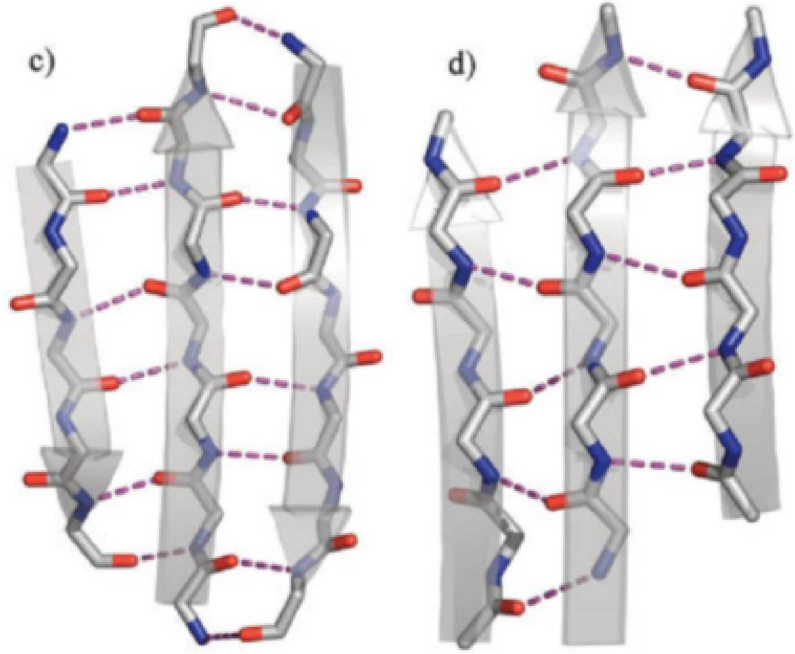
β sheet conformation

Approx. 30% of globular proteins

Strands of 5-10 residues run in parallel

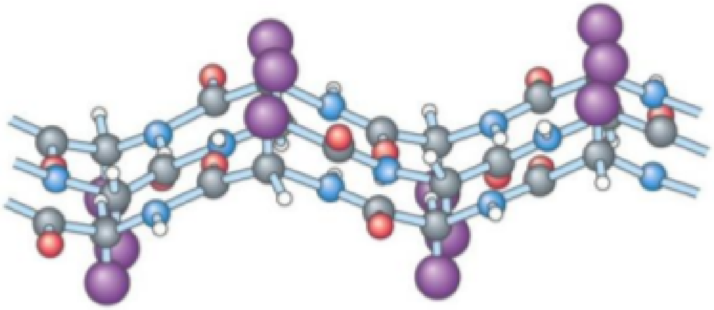
Strands are held together by H-bonds

Different β-sheet forming propensity for various residues



antiparallel β-sheet

parallel β-sheet



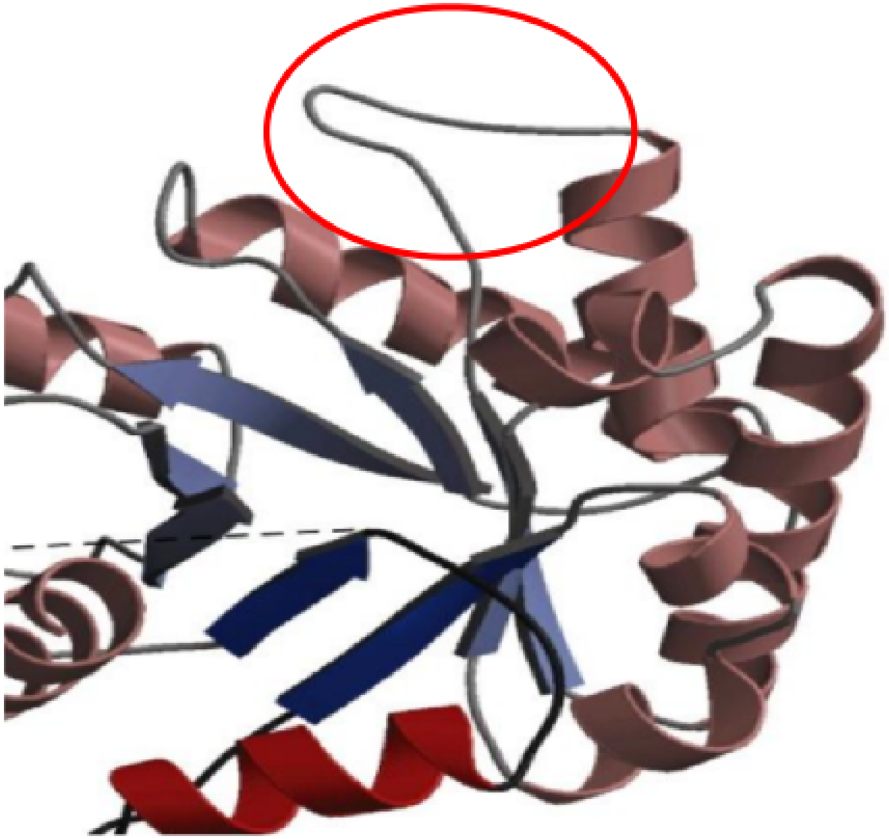
Side view

Loops and turns

Typically have hydrophilic characters. Occur on the outer regions of the protein, form H-bonds with water and other molecules

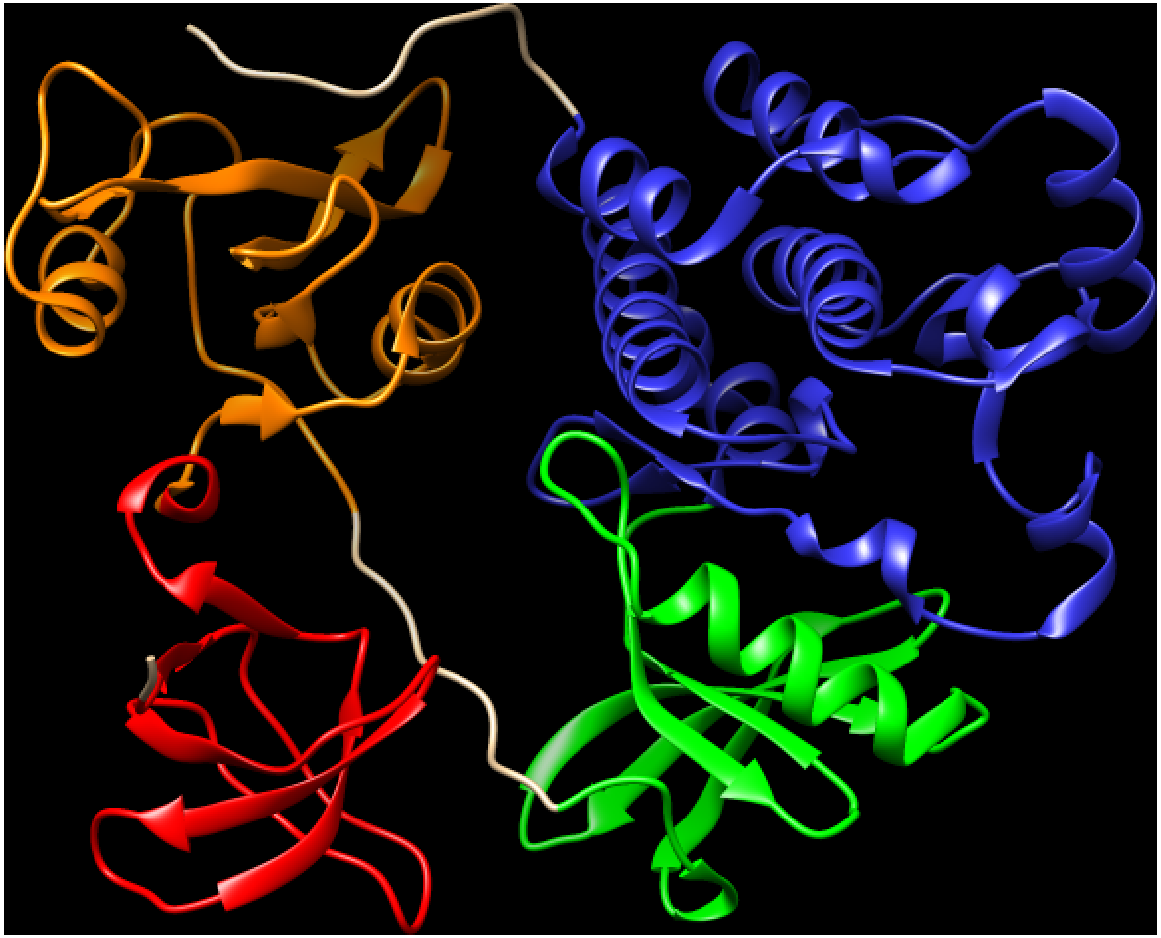
Often form binding regions and active sites in enzymes and receptors

Different loop-forming propensity for various residues



Domains

Many proteins feature distinct compact structural units



Domains

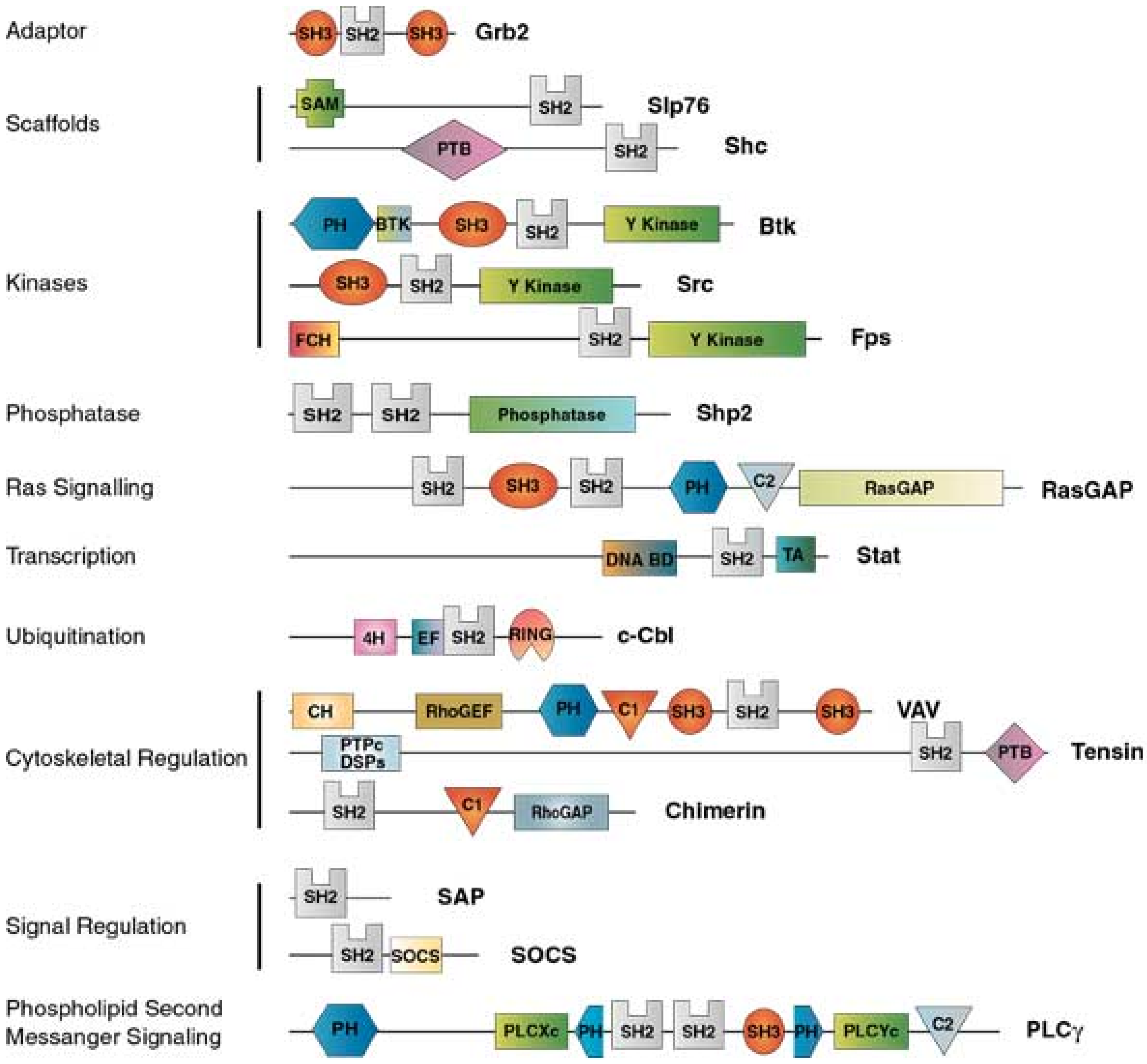
Compact units with globular-like structures

Domains are basic building blocks of proteins

Typically fulfill a well-specified function

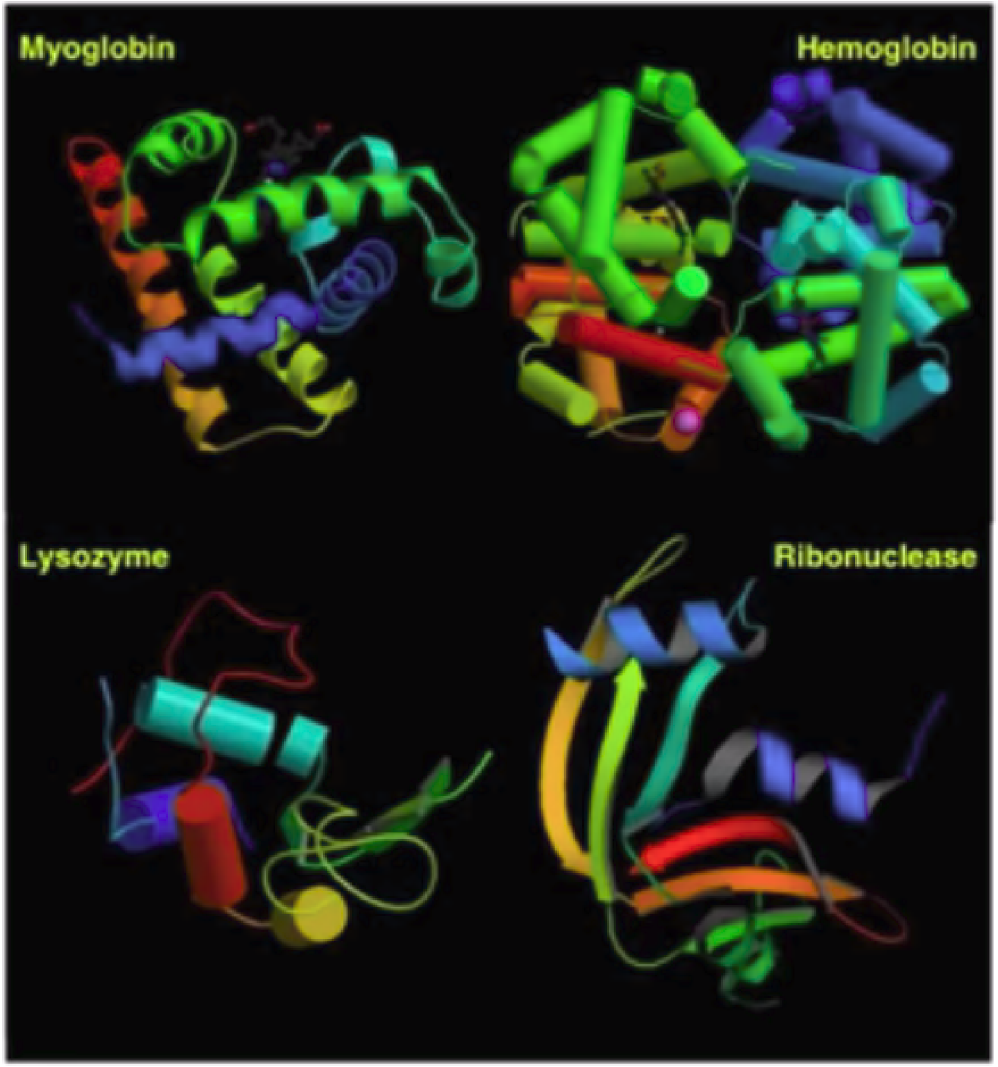
Can appear in various biological contexts

SH2 domain


Adaptor
SH3
SH2
SH3
Grb2
Scaffolds
SAM
SH2
Slp76
PTB
SH2
Shc
Kinases
PH
BTK
SH3
SH2
Y Kinase
Btk
SH3
SH2
Y Kinase
Src
FCH
SH2
Y Kinase
Fps
Phosphatase
SH2
SH2
Phosphatase
Shp2
Ras Signalling
SH2
SH3
SH2
PH
C2
RasGAP
RasGAP
Transcription
DNA BD
SH2
TA
Stat
Ubiquitination
4H
EF
SH2
RING
c-Cbl
Cytoskeletal Regulation
CH
RhoGEF
PH
C1
SH3
SH2
SH3
VAV
PTPc DSPs
SH2
PTB
Tensin
SH2
C1
RhoGAP
Chimerin
Signal Regulation
SH2
SAP
SH2
SOCS
SOCS
Phospholipid Second Messanger Signaling
PH
PLCXc
PH
SH2
SH2
SH3
PH
PLCYc
C2
PLCγ

Protein Data Bank

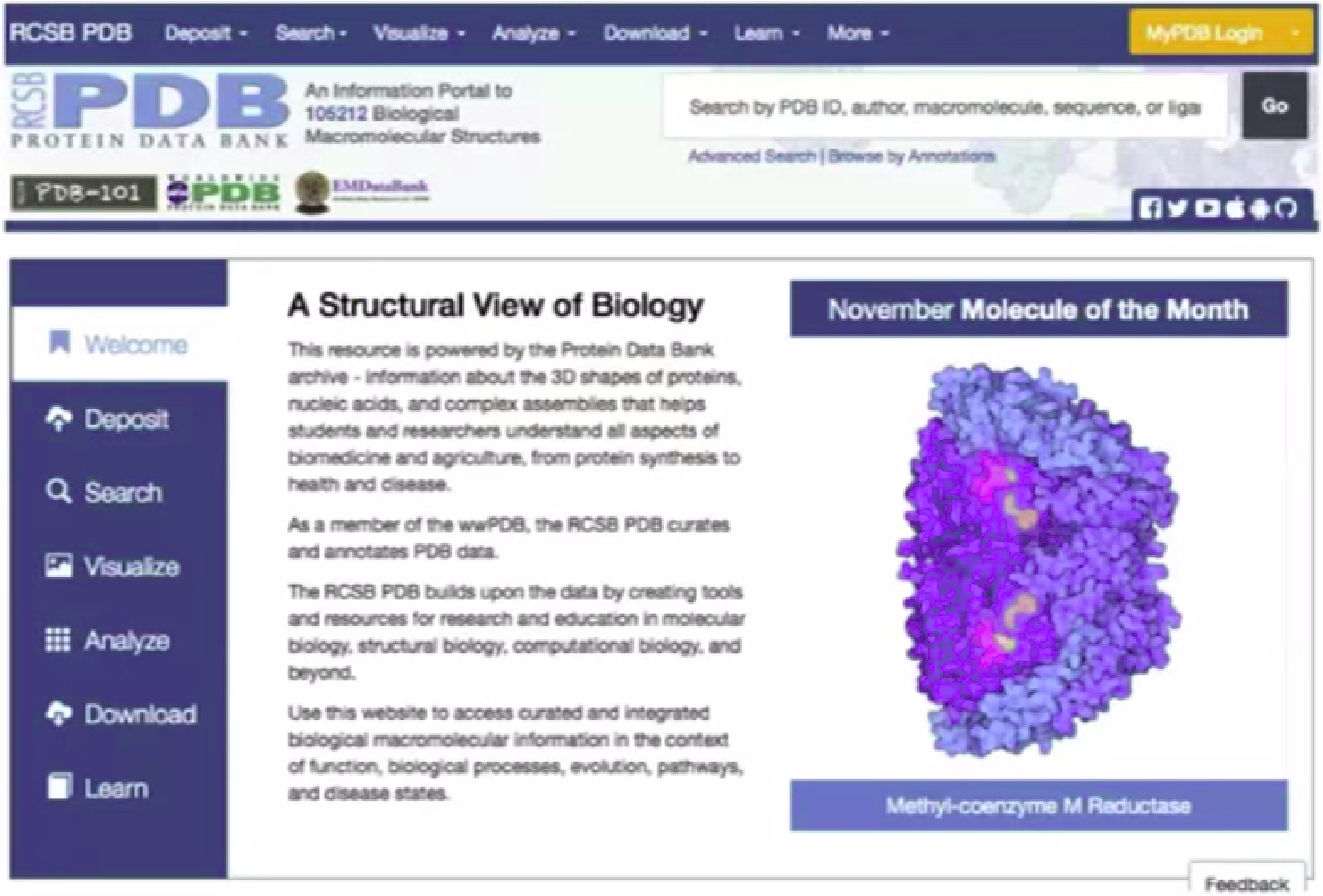
- First open access digital resource in biology (est. 1971 with 7 entries)
- Single global archive of 3-D macromolecular structures (contains >100,000 entries)
- US PDB = RCSB PDB
  - Headquartered at Rutgers/UCSD (NSF, NIH, DOE)
  - Part of Worldwide PDB (with EU and Japan)
  - Makes PDB data freely available to all *via* www.rcsb.org



Some of the first few structures in the PDB

RCSB PDB Portal rcsb.org

- Searching
- Visualizing
- Comparing
- Accessing external data
- Reporting
- PDB-101 resources for education



An Information Portal to 105212 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or liga

A Structural View of Biology

This resource is powered by the Protein Data Bank archive - information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

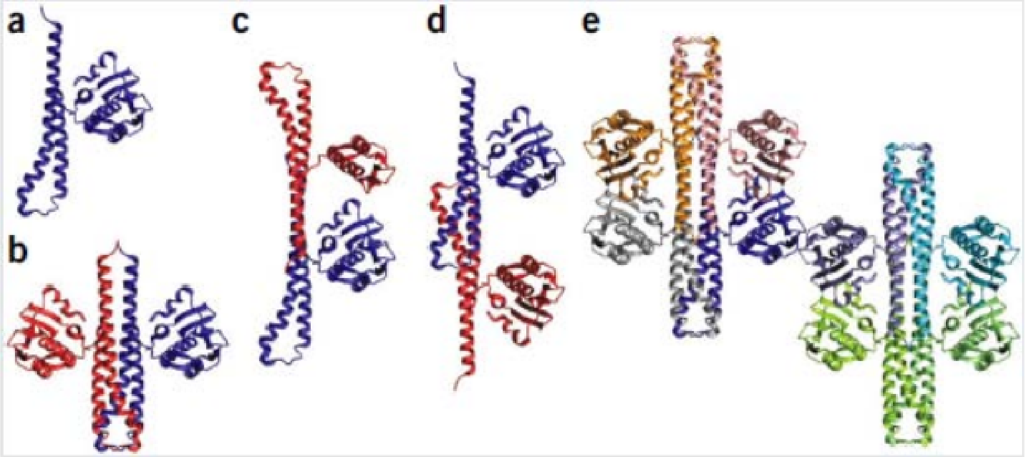
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Use this website to access curated and integrated biological macromolecular information in the context of function, biological processes, evolution, pathways, and disease states.

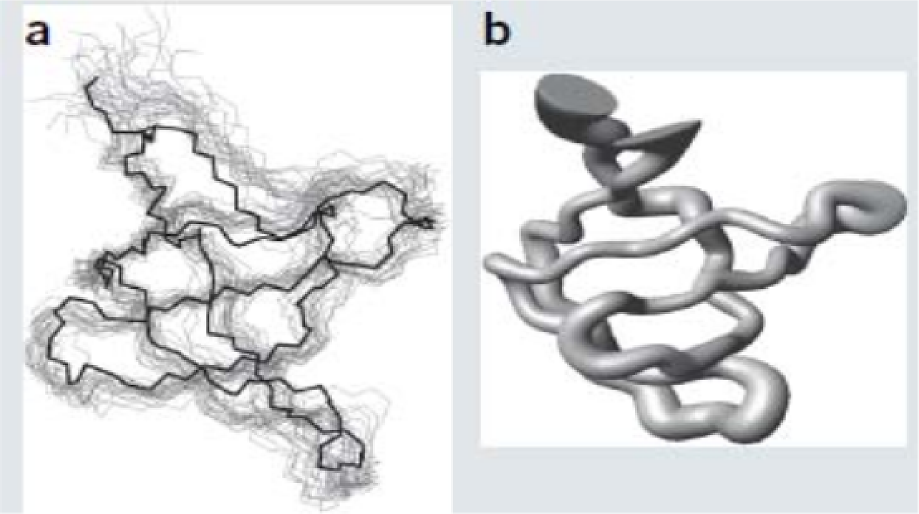
November Molecule of the Month

Methyl-coenzyme M Reductase

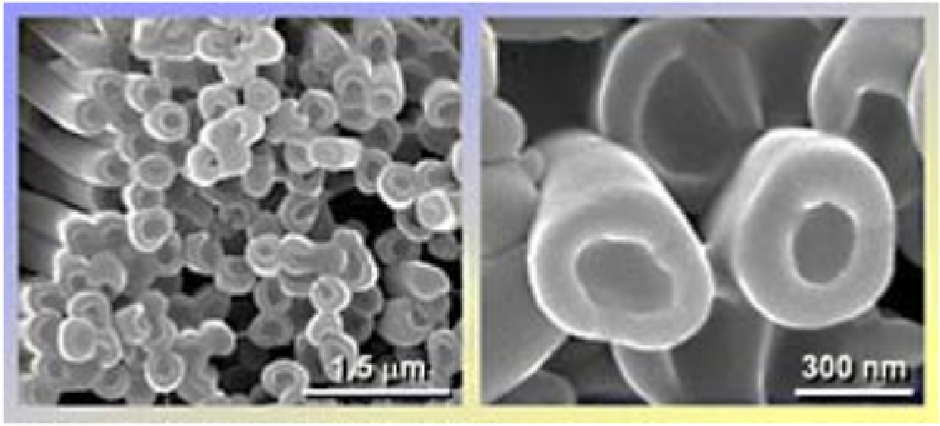
X-ray crystallography



NMR



Electron microscopy



X-ray crystallography

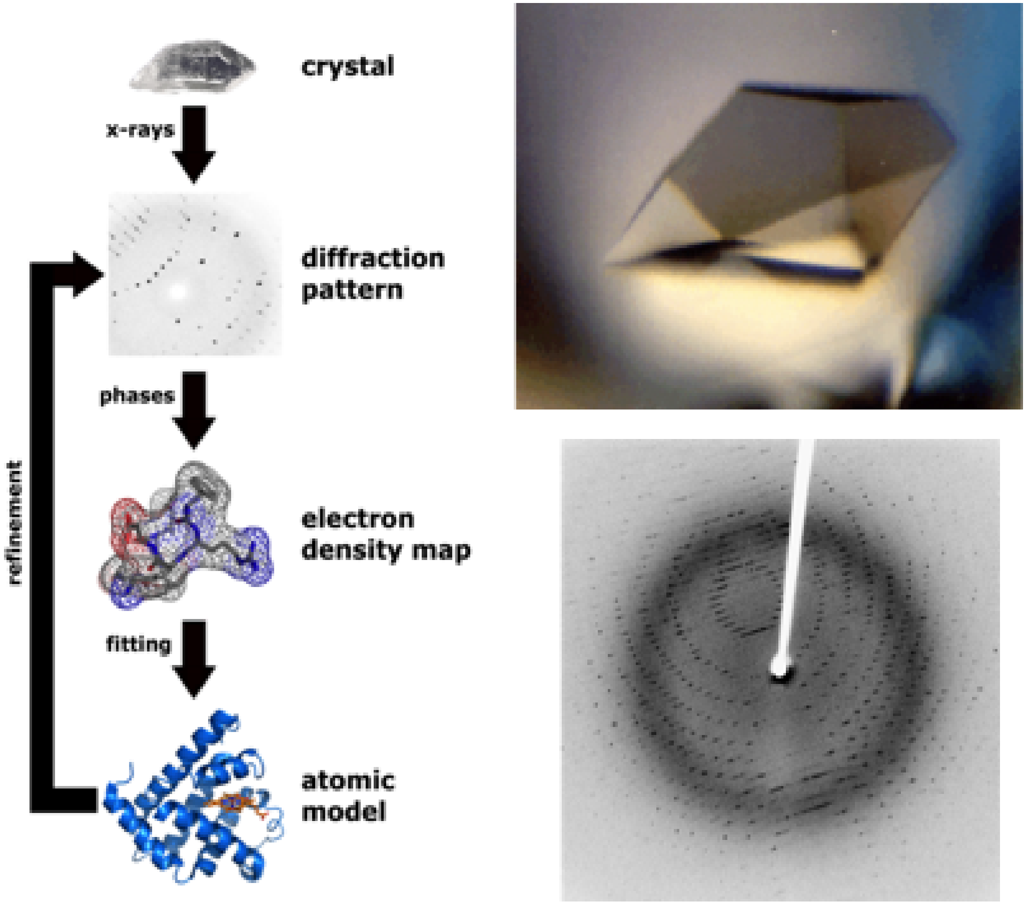


Figure 4: Left: Structure determination by X-ray crystallography. Work by Bragg and others connected spots on diffraction pattern with arrangement of atoms in the crystal to solve simple structures like salts. Work by Perutz, Rossmann, and Blow allowed automated processing of crystal data to solve complex structures. Right (top): typical protein crystal, less than one millimeter in size; Right (bottom): protein crystal diffraction pattern.

X-ray:

- X-rays have short wave lengths (approx. 1.5 Å) – needed to measure the typical atom-atom distances

- gives information about electron density, the model has to be fit into that

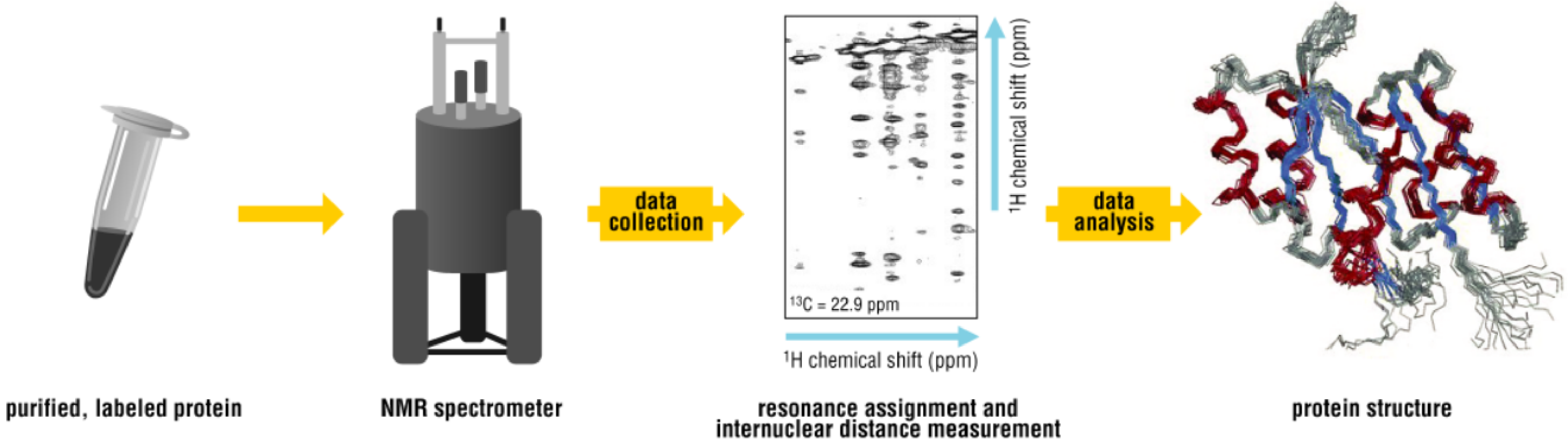
- crystallization artefacts

- non-physiological environment

- no information on hydrogens

NMR

^{1}H-NMR (Proton Nuclear Magnetic Resonance)



purified, labeled protein

NMR spectrometer

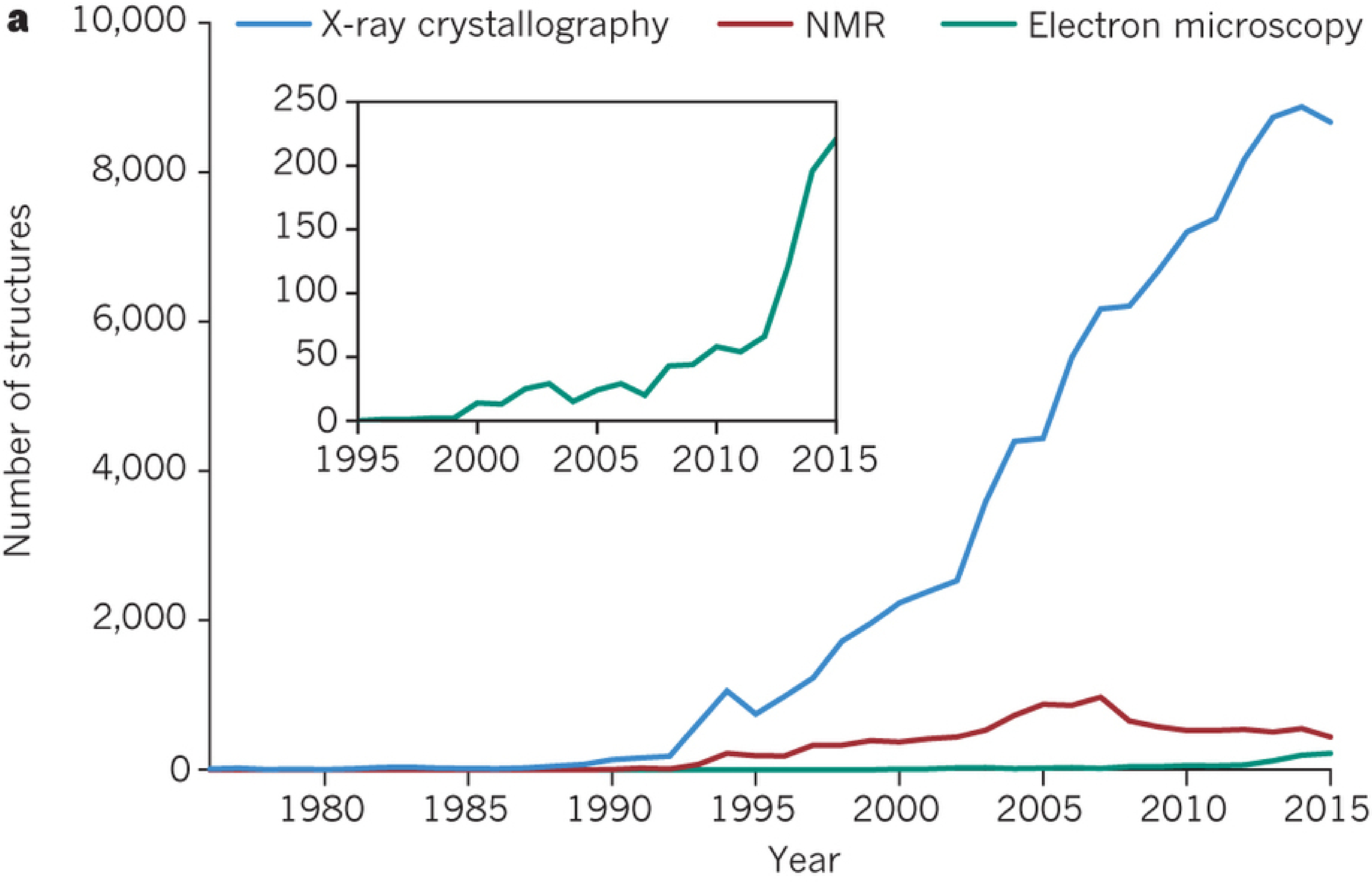
data collection

resonance assignment and internuclear distance measurement

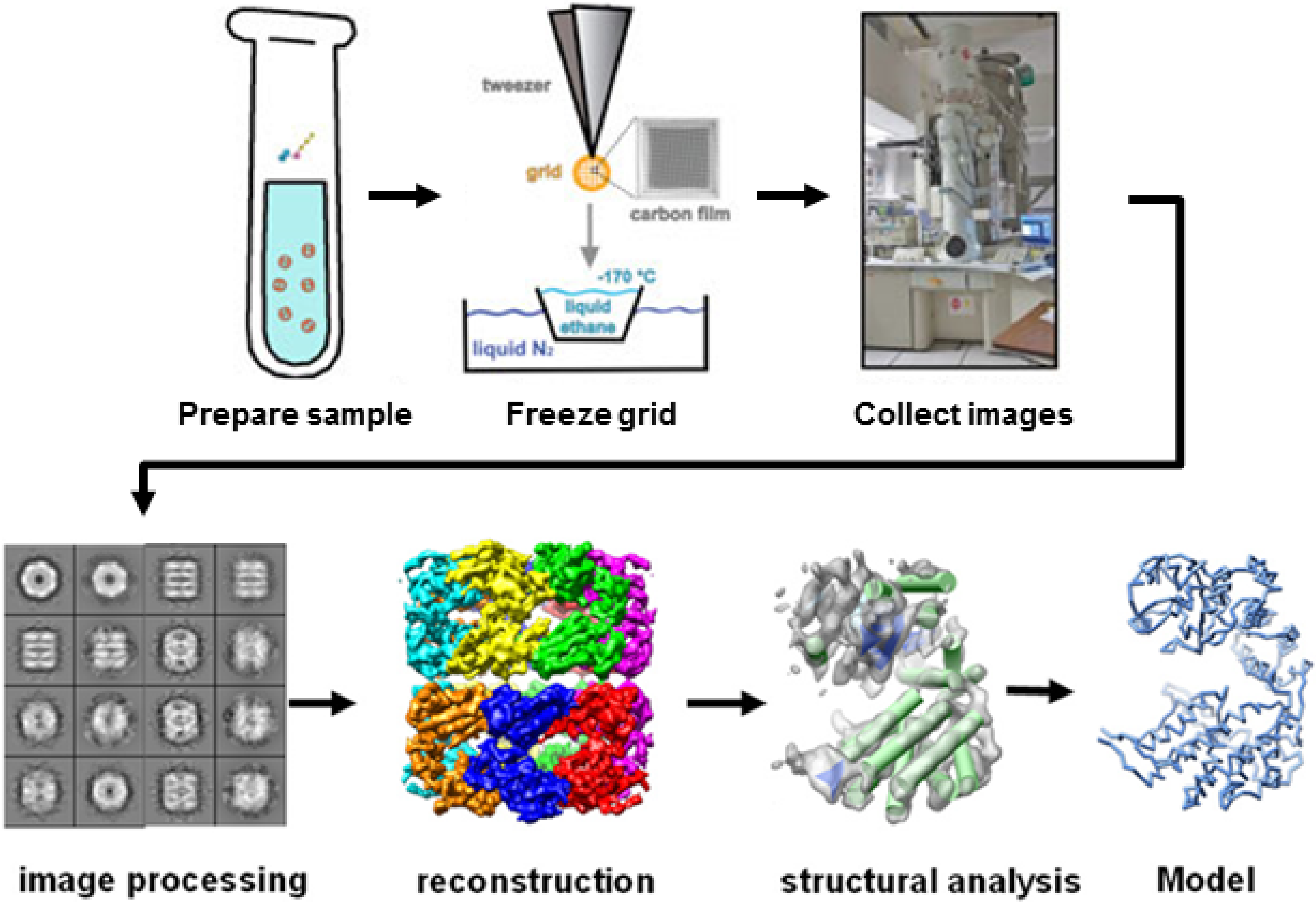
data analysis

protein structure

- in solution
- usually yields a structural ensemble that fulfills the distance constraints
- only small proteins
- less precise model
- usable for flexible proteins as well

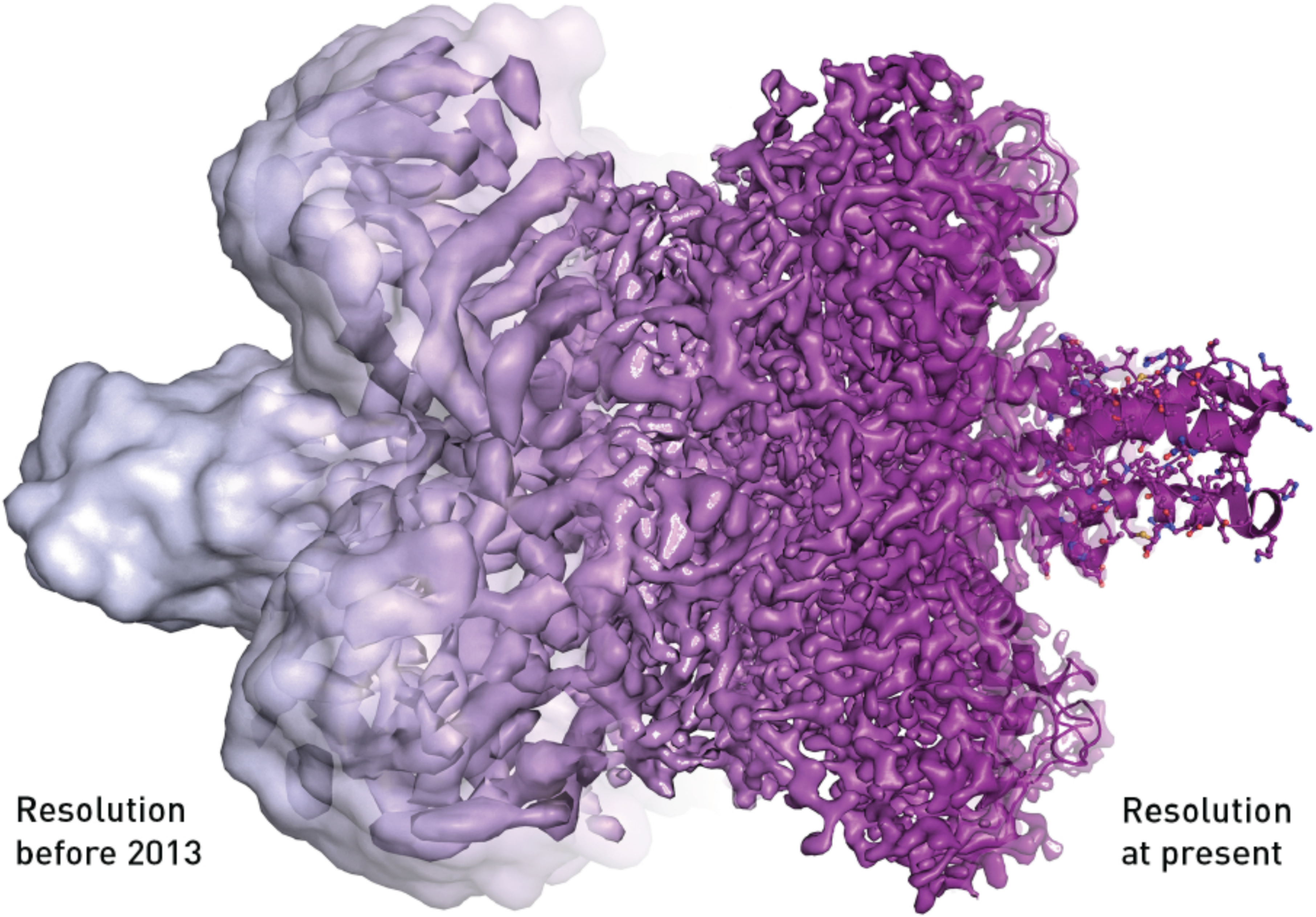

a
X-ray crystallography
NMR
Electron microscopy
10,000
8,000
6,000
4,000
2,000
0
Number of structures
1980
1985
1990
1995
2000
2005
2010
2015
Year
250
200
150
100
50
0
1995
2000
2005
2010
2015

Cryo-EM (atomic resolution)

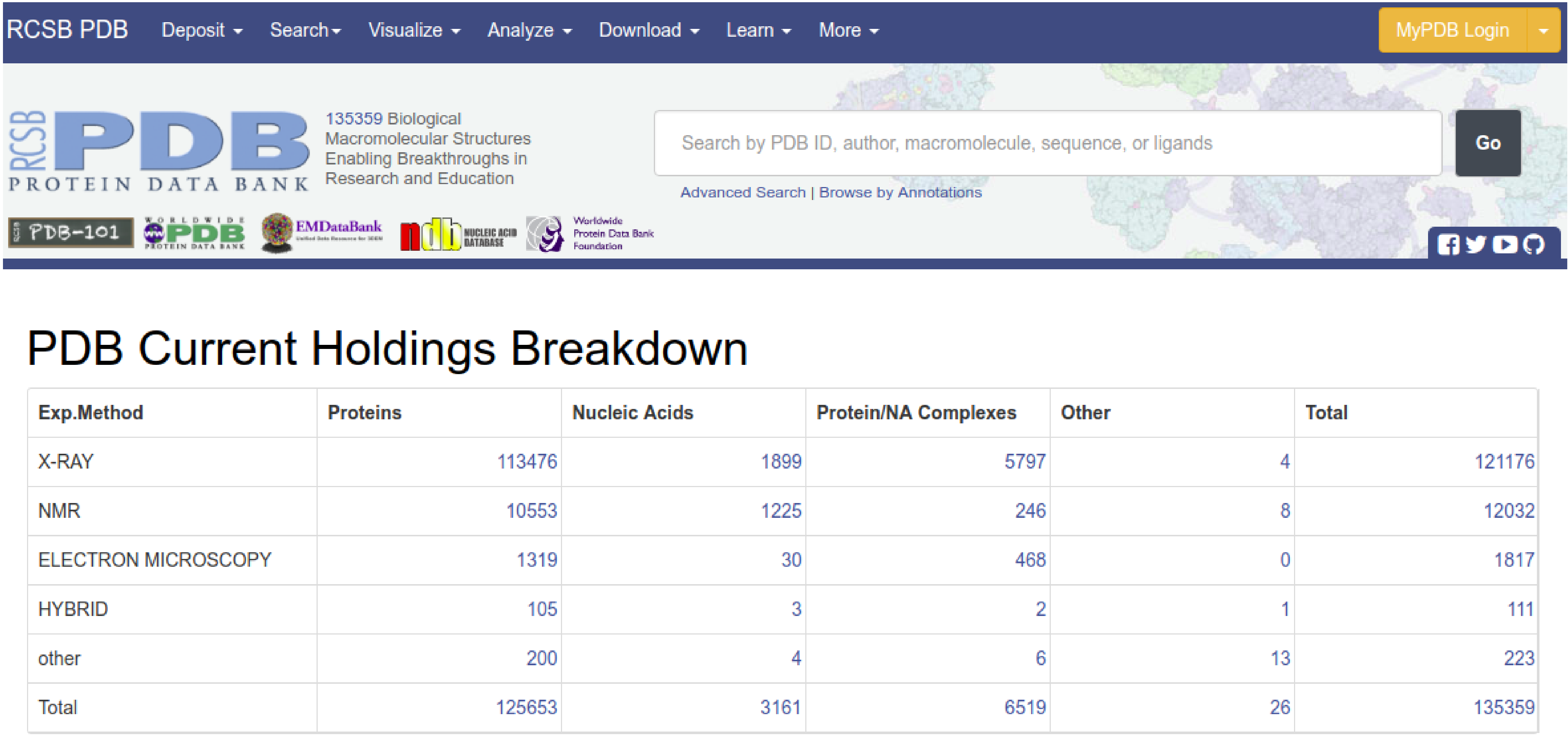


Prepare sample → Freeze grid → Collect images → image processing → reconstruction → structural analysis → Model

Nobel Prize in Chemistry 2017



PDB statistics

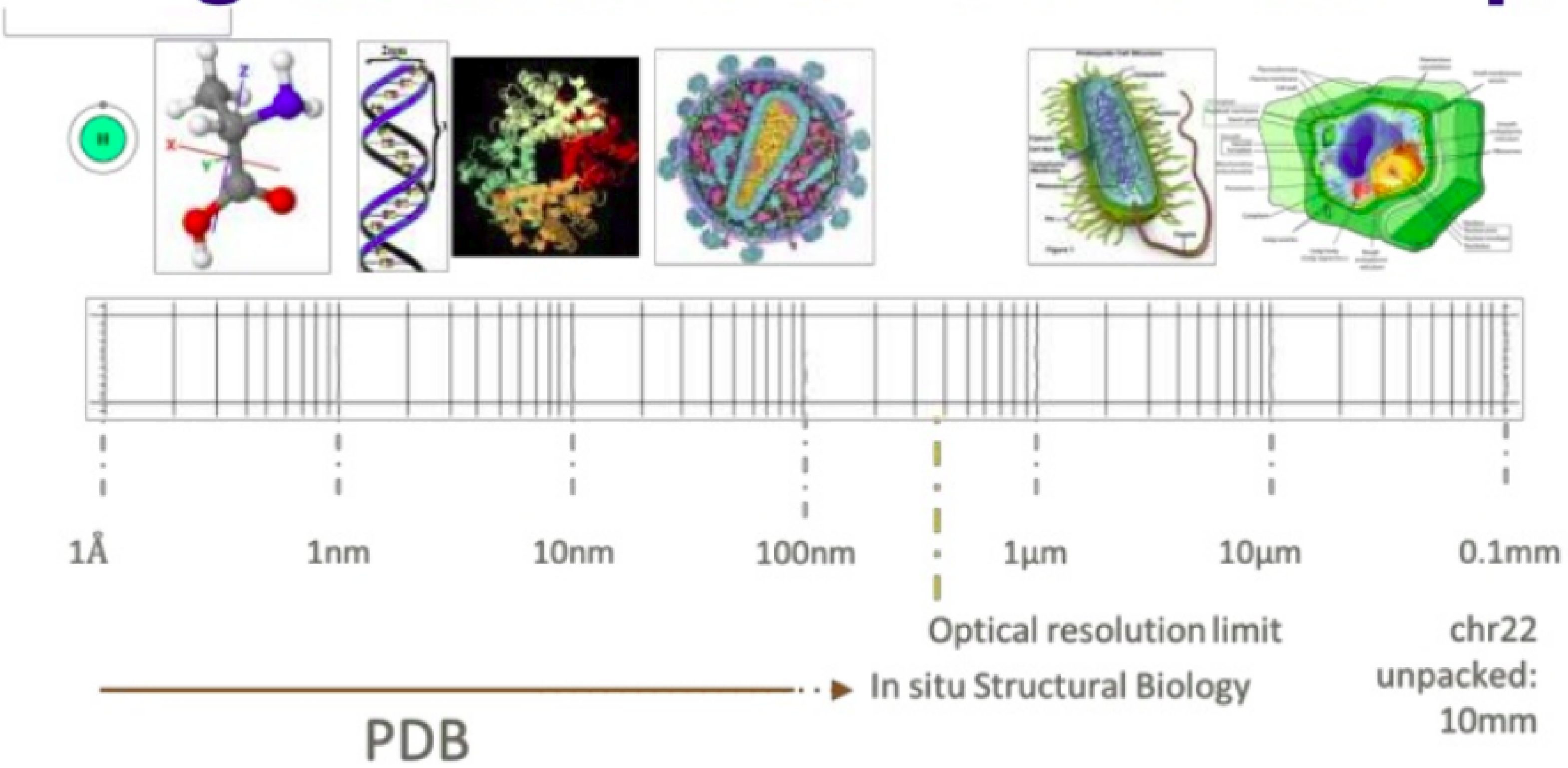

RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
RCSB PDB
PROTEIN DATA BANK
Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataBank
ndb NUCLEIC ACID DATABASE
Worldwide Protein Data Bank Foundation

135359 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB Current Holdings Breakdown

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	113476	1899	5797	4	121176
NMR	10553	1225	246	8	12032
ELECTRON MICROSCOPY	1319	30	468	0	1817
HYBRID	105	3	2	1	111
other	200	4	6	13	223
Total	125653	3161	6519	26	135359

Growing Structure Size and Complexity



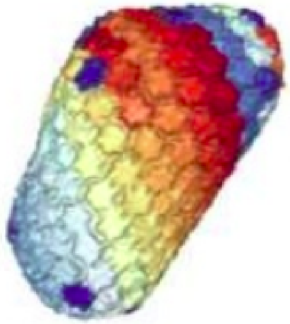
Optical resolution limit

chr22
unpacked:
10mm

In situ Structural Biology

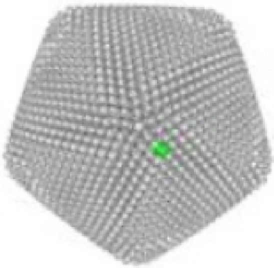
PDB

Largest asymmetric structure in PDB



HIV-1 capsid: PDB ID 3J3Q
~2.4M unique atoms

Largest symmetric structure in PDB



Faustovirus major capsid: PDB ID 5J7V
~40M overall atoms

The .pdb file format

The PDB (Protein Data Bank) file format is a text format describing the structure of macromolecules incorporated in the database.

- Description and annotation of the structures of proteins and nucleic acids, including atomic coordinates, side-chain rotamers, secondary structure elements, and atomic connectivity.

- Structures often contain other molecules as well, such as water, ions, ligands, etc. These are also described in the pdb format.

Format description:

http://www.wwpdb.org/documentation/format33/v3.3.html

PDB ID: unique identifier

Each atomic coordinate file in the Protein Data Bank has a unique identifier composed of exactly 4 characters. The first one is always a number, the rest can be either a number or a letter.

There are over 400,000 possible 4-digit PDB IDs (419,904 or 466,560 if "0" can also be the first character). Currently there are approx. 120,000 entries.

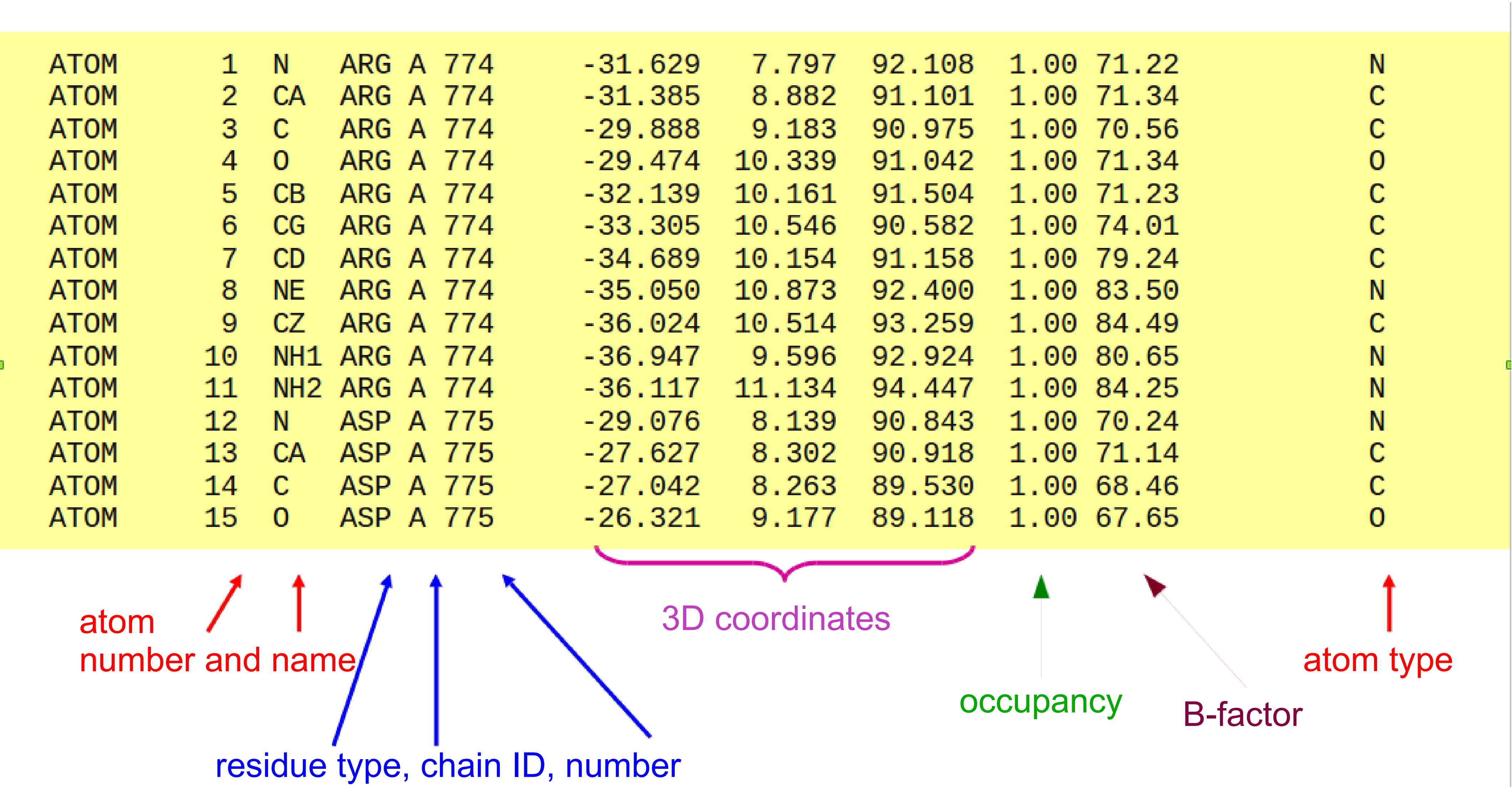
Examples:

- 1mbn - 1973, the first protein structure model, **myoglobin**
- 1tna - 1975, the first RNA structure, **yeast phenylalanine transfer RNA**
- 1bna - 1980, the first **B-DNA** double helix structure (determined using X-ray 27 years after the 1953 theoretically determined structural model of Watson & Crick)
- 2hhd - human **hemoglobin,** (deoxy form)
- 9ins - **insulin**

The .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
...
```

The .pdb file format



```
ATOM      1  N   ARG A 774     -31.629   7.797  92.108  1.00 71.22           N
ATOM      2  CA  ARG A 774     -31.385   8.882  91.101  1.00 71.34           C
ATOM      3  C   ARG A 774     -29.888   9.183  90.975  1.00 70.56           C
ATOM      4  O   ARG A 774     -29.474  10.339  91.042  1.00 71.34           O
ATOM      5  CB  ARG A 774     -32.139  10.161  91.504  1.00 71.23           C
ATOM      6  CG  ARG A 774     -33.305  10.546  90.582  1.00 74.01           C
ATOM      7  CD  ARG A 774     -34.689  10.154  91.158  1.00 79.24           C
ATOM      8  NE  ARG A 774     -35.050  10.873  92.400  1.00 83.50           N
ATOM      9  CZ  ARG A 774     -36.024  10.514  93.259  1.00 84.49           C
ATOM     10  NH1 ARG A 774     -36.947   9.596  92.924  1.00 80.65           N
ATOM     11  NH2 ARG A 774     -36.117  11.134  94.447  1.00 84.25           N
ATOM     12  N   ASP A 775     -29.076   8.139  90.843  1.00 70.24           N
ATOM     13  CA  ASP A 775     -27.627   8.302  90.918  1.00 71.14           C
ATOM     14  C   ASP A 775     -27.042   8.263  89.530  1.00 68.46           C
ATOM     15  O   ASP A 775     -26.321   9.177  89.118  1.00 67.65           O
```

atom
number and name

residue type, chain ID, number

3D coordinates

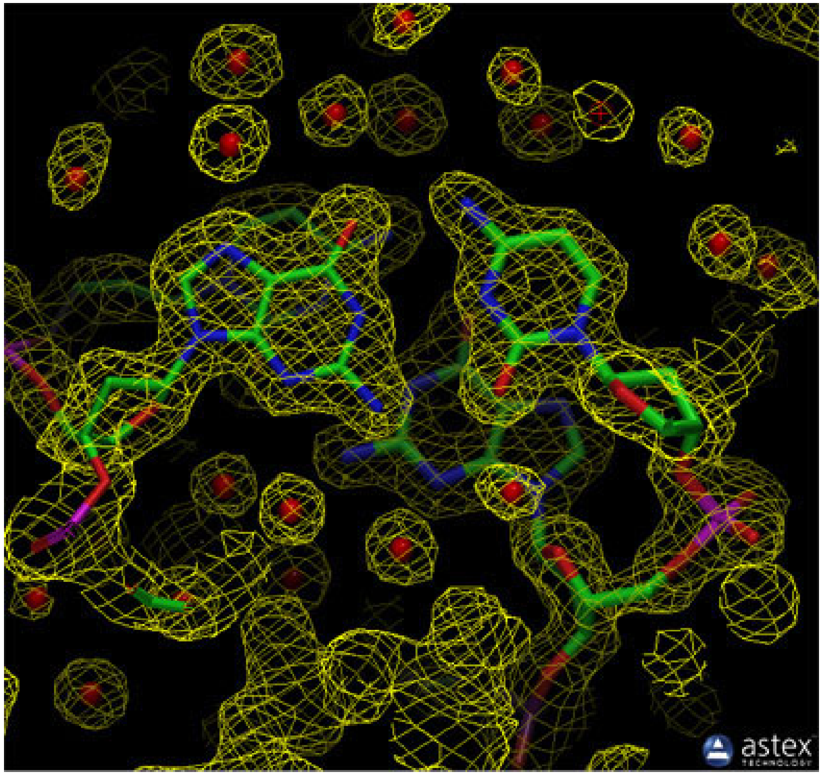
occupancy

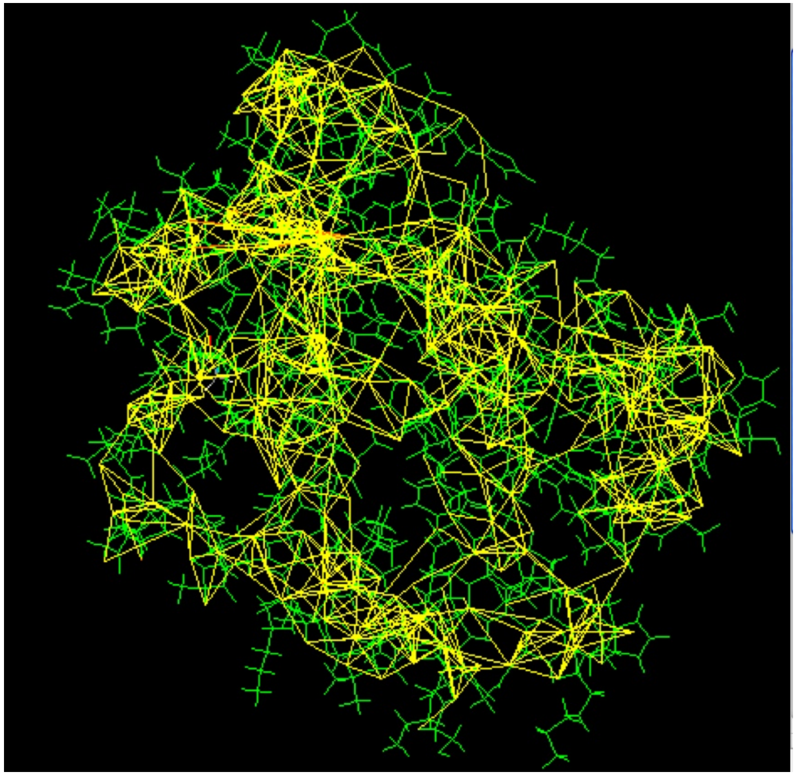
B-factor

atom type

Model

All protein structures are models! Structures are not directly measured, but are generated as models that best fit the collected experimental data.





Resolution (X-ray)

- Describes the reliability of determined atomic coordinates

Very low:>4Å

Individual coordinates cannot be interpreted

Low: 3.0-4.0Å

The fold is recognizable

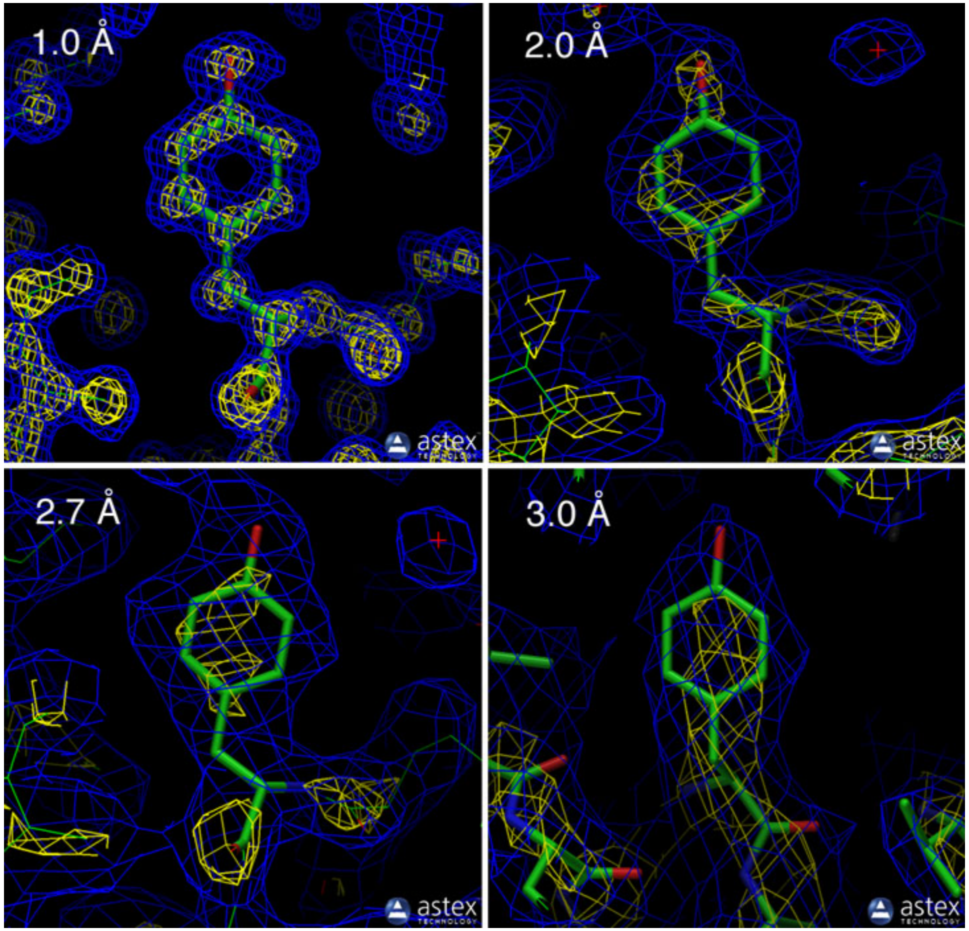
Average: 1.8-3.0Å

The majority of the structure is correct, with incorrect rotamers and unreliable surface loop conformations

Good: 1.0 – 1.8Å

Atomic level: <1.0A

Resolution can change for each position!



Describing structure quality

Expected distribution:

Based on small molecules

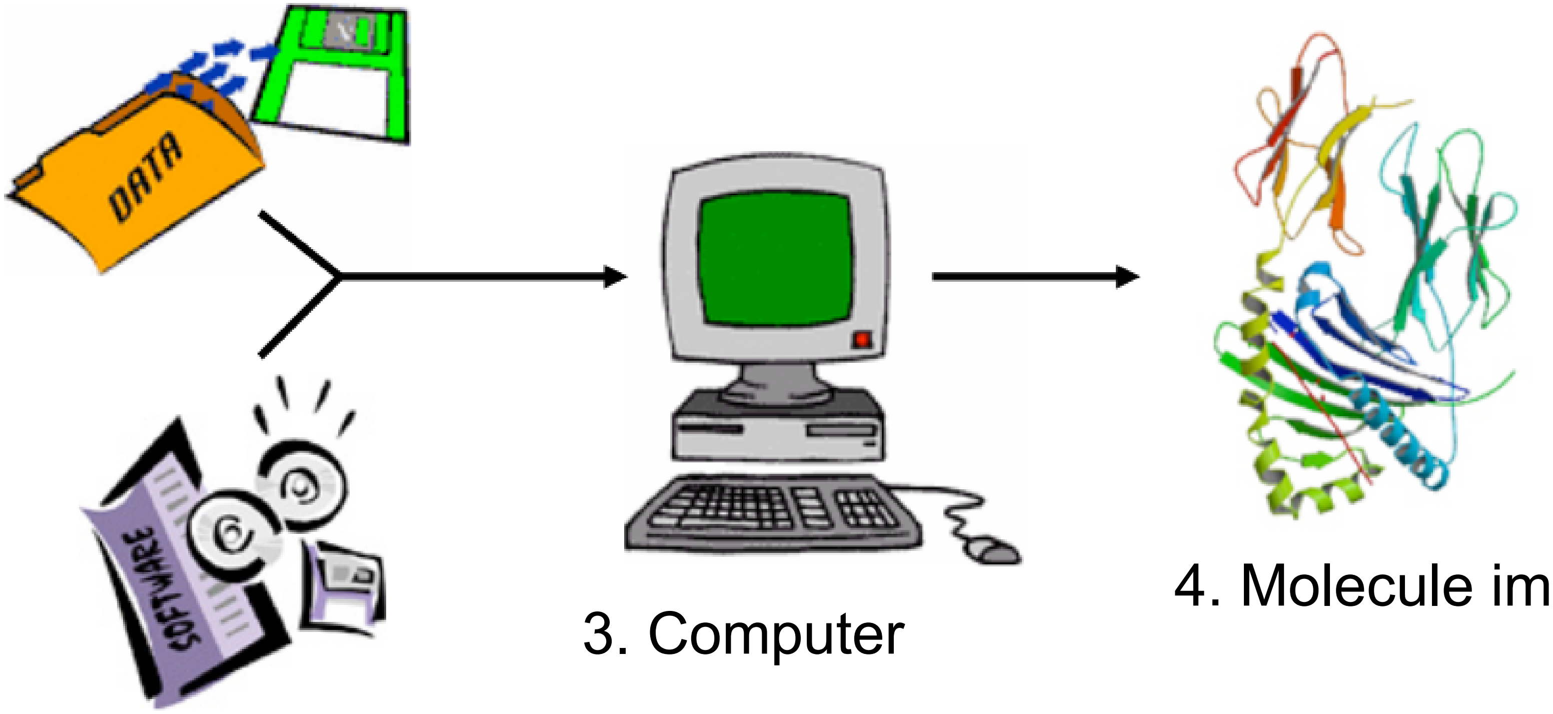
Based on known, high quality structures

Possible parameters

- Correct bond lengths and bond angles
- No atom-atom clashes
- Most buried amid groups form H-bonds
- Based on main chain conformational properties

Visualizing protein structures

1. PDB coordinate file

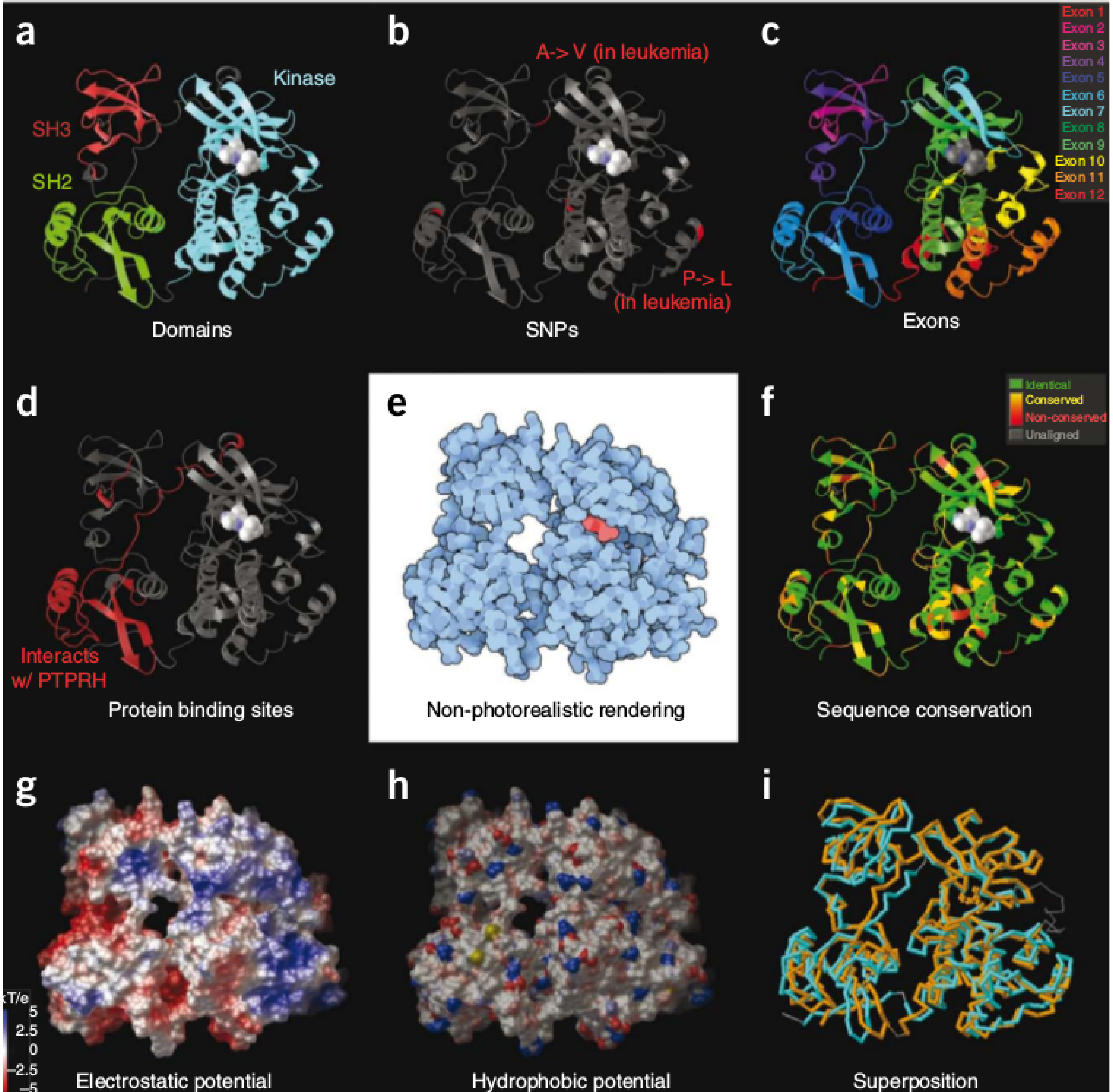


3. Computer

4. Molecule image

2. Visualization program

Eg.: Rasmol, Pymol, Chimera,
VMD, Jmol, Swiss PDB viewer



Domains

SNPs

Exons

Protein binding sites

Non-photorealistic rendering

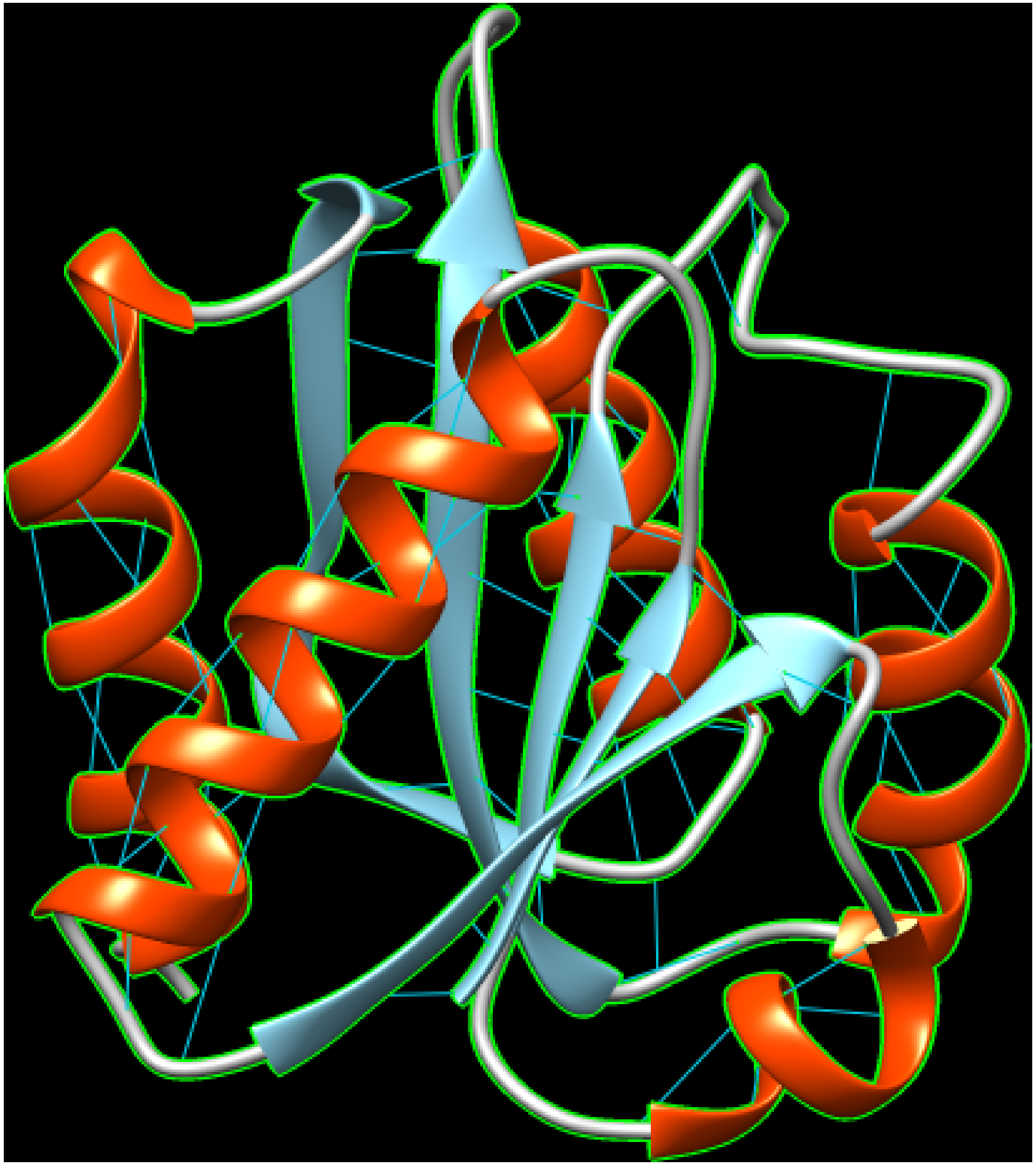
Sequence conservation

Electrostatic potential

Hydrophobic potential

Superposition

Secondary structures are stabilized by H-bonds



Secondary structure determination

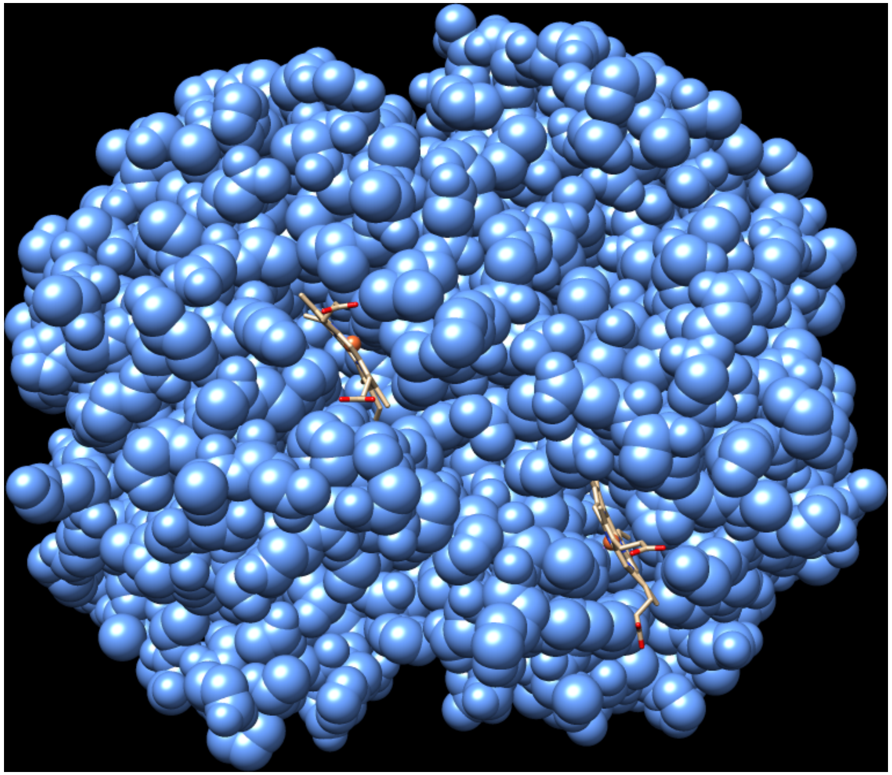
Can be based on:
- H-bond patterns
- Dihedral angles

Automatic determination using algorithms
- DSSP
- STRIDE

3 (alpha, beta, coil)
or more categories (e.g. turn, other helix types)

Do not agree 100%

The inside of the protein is tightly packed

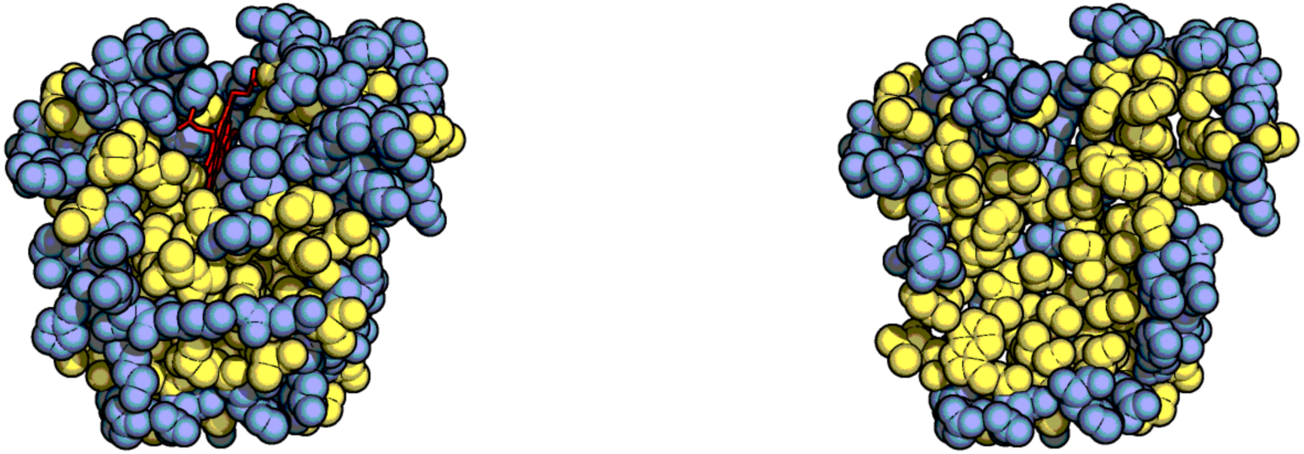


Hydrophobic core

Hydrophobic side chains go into the core of the molecule – but the main chain is highly polar.

The polar groups (C=O and NH) are neutralized through formation of H-bonds.

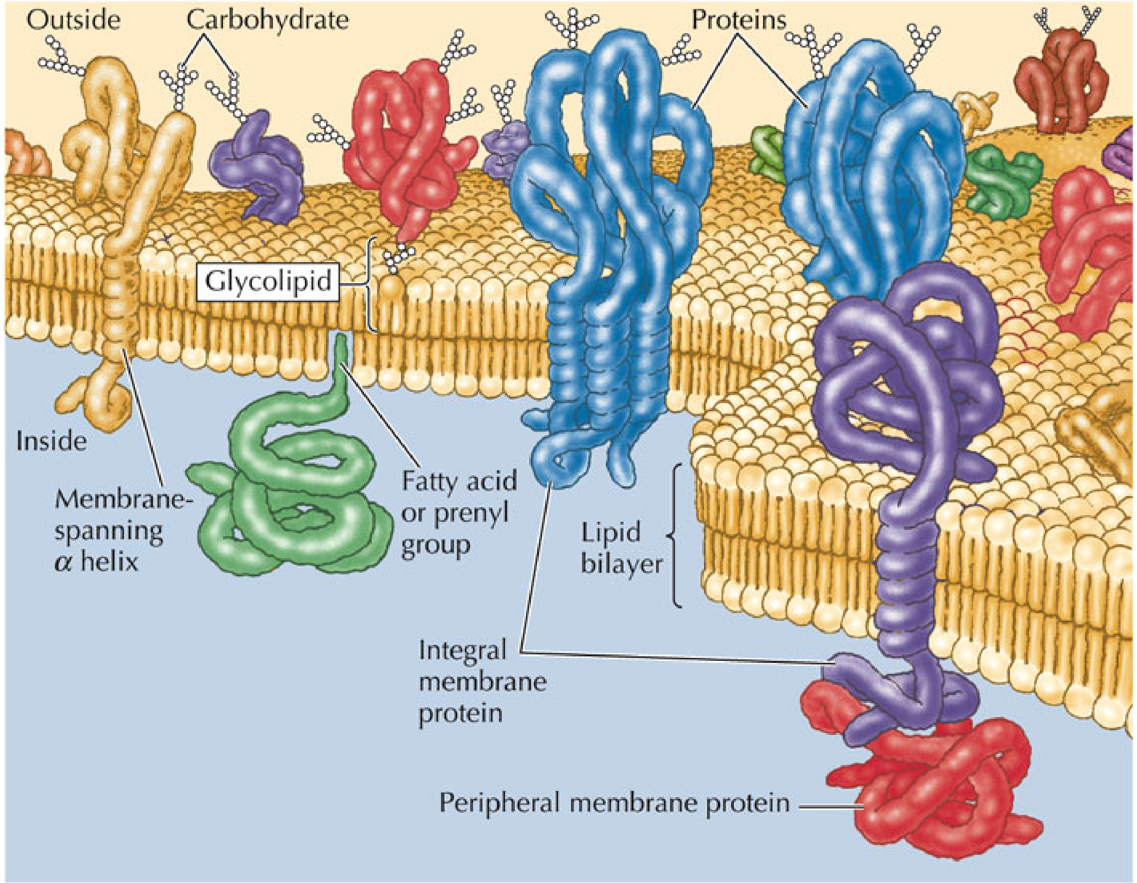
Myoglobin



surface

buried

Membrane proteins

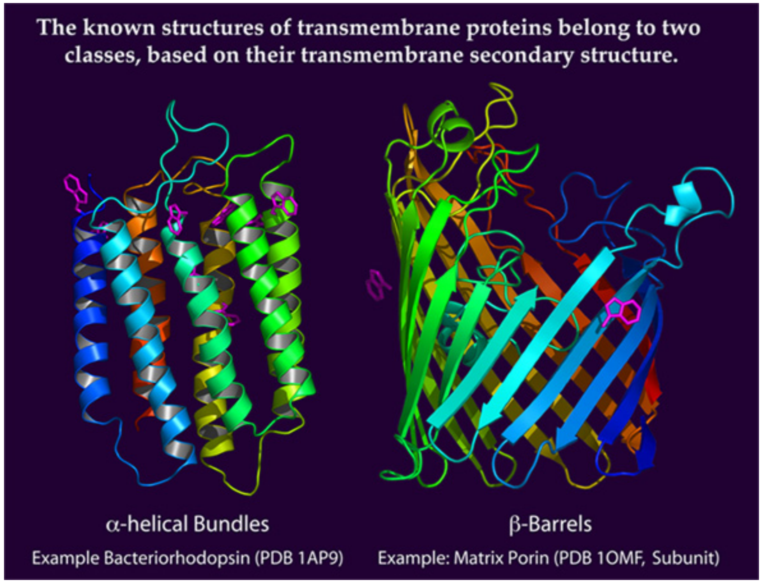

Outside
Carbohydrate
Proteins
Glycolipid
Inside
Membrane-spanning α helix
Fatty acid or prenyl group
Lipid bilayer
Integral membrane protein
Peripheral membrane protein

THE CELL, Fourth Edition, Figure 2.25 © 2006 ASM Press and Sinauer Associates, Inc.

Important for:

Energy production
Transport
Cell-cell junction
Signaling

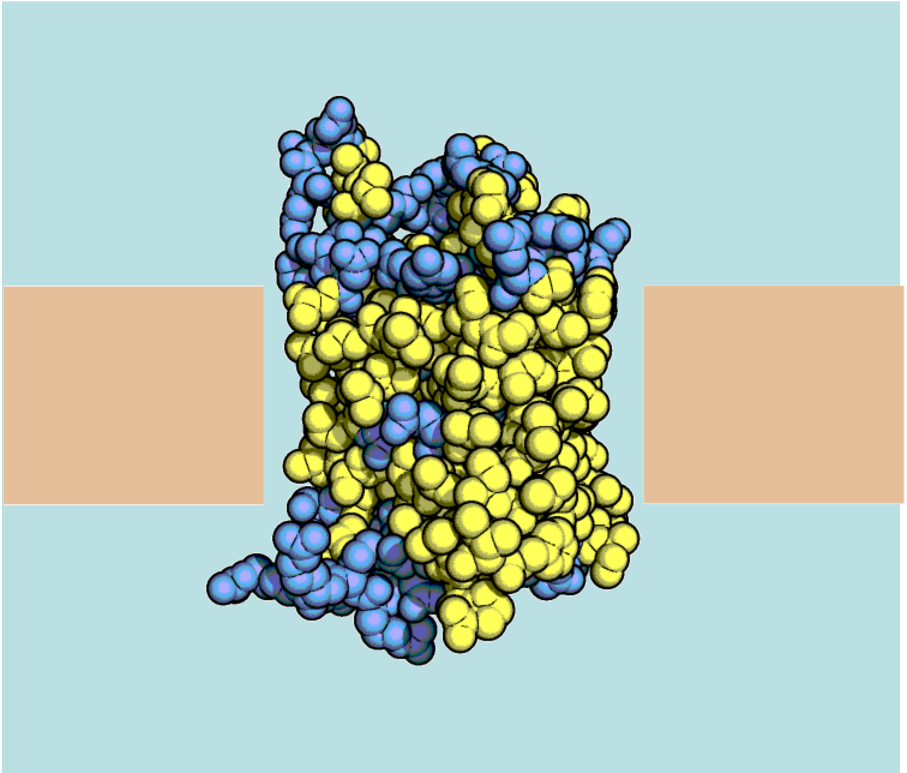
Drug targets



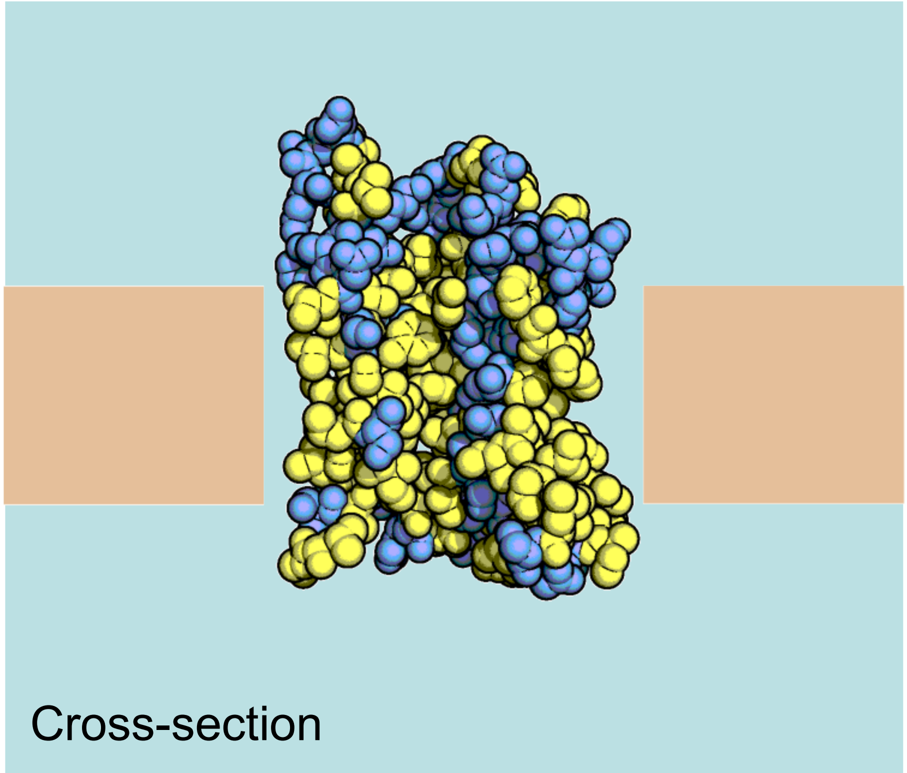
The known structures of transmembrane proteins belong to two classes, based on their transmembrane secondary structure.

α-helical Bundles
Example Bacteriorhodopsin (PDB 1AP9)
β-Barrels
Example: Matrix Porin (PDB 1OMF, Subunit)

Hydrophobicity of membrane proteins

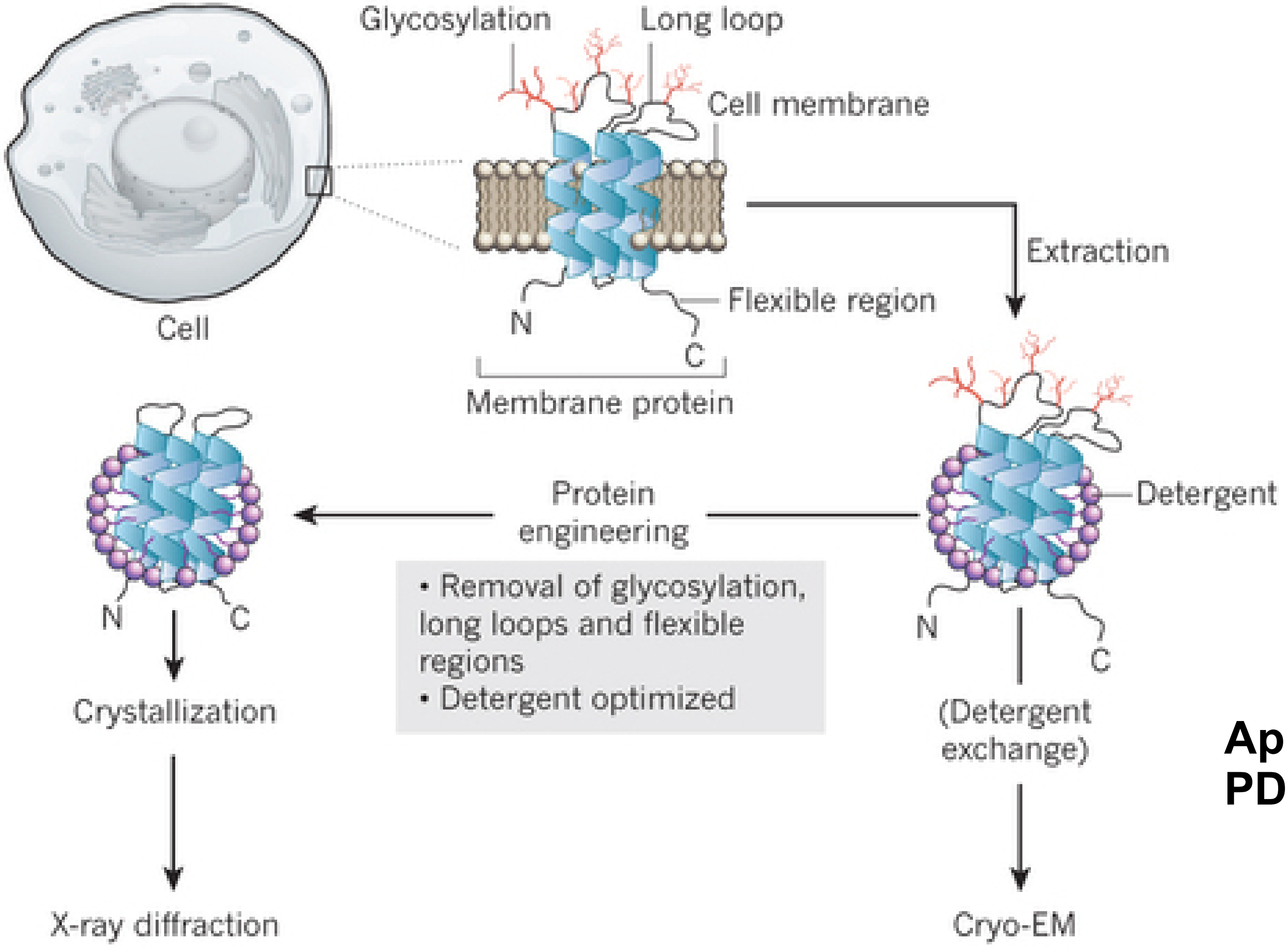


Aquaporin



Aquaporin

Structure determination of transmembrane proteins



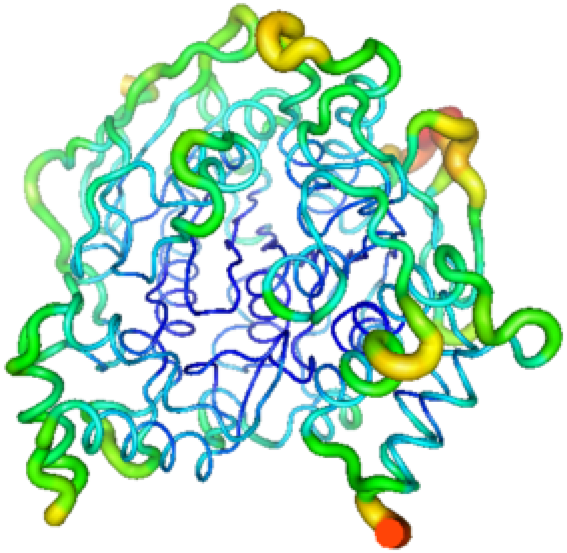
- Removal of glycosylation, long loops and flexible regions
- Detergent optimized

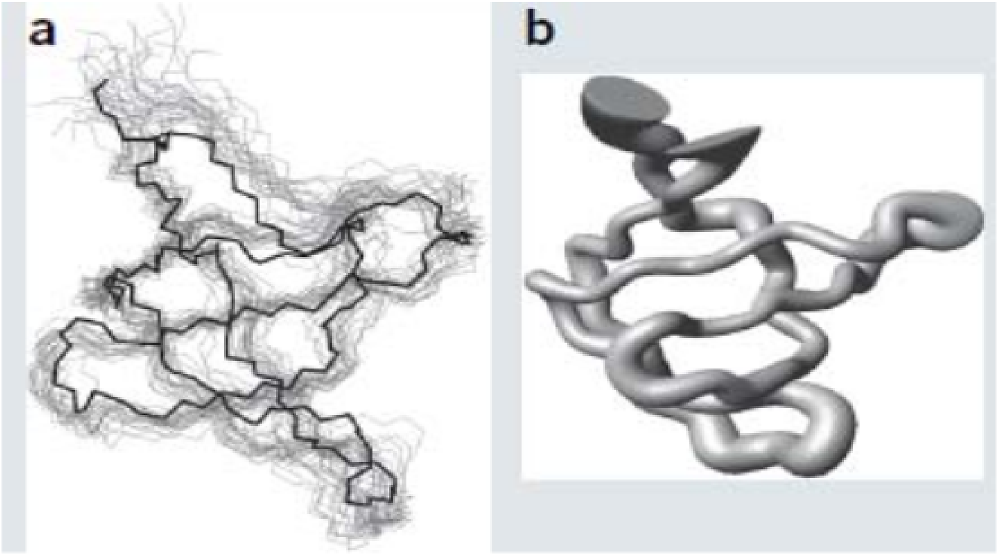
Approx. 2% of PDB structures

Proteins are dynamic molecules

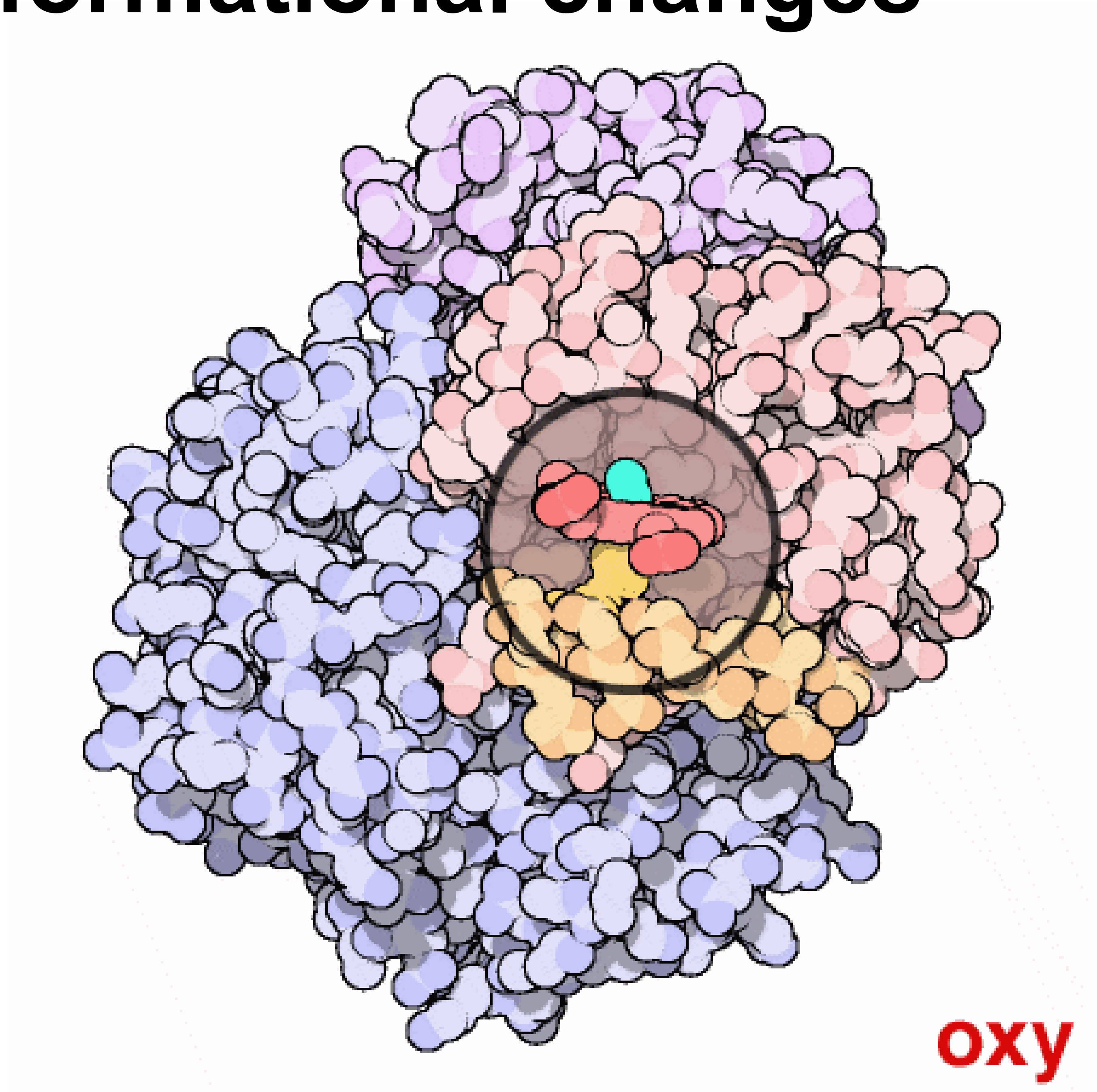
X-ray
B-factor

NMR
Structural variability

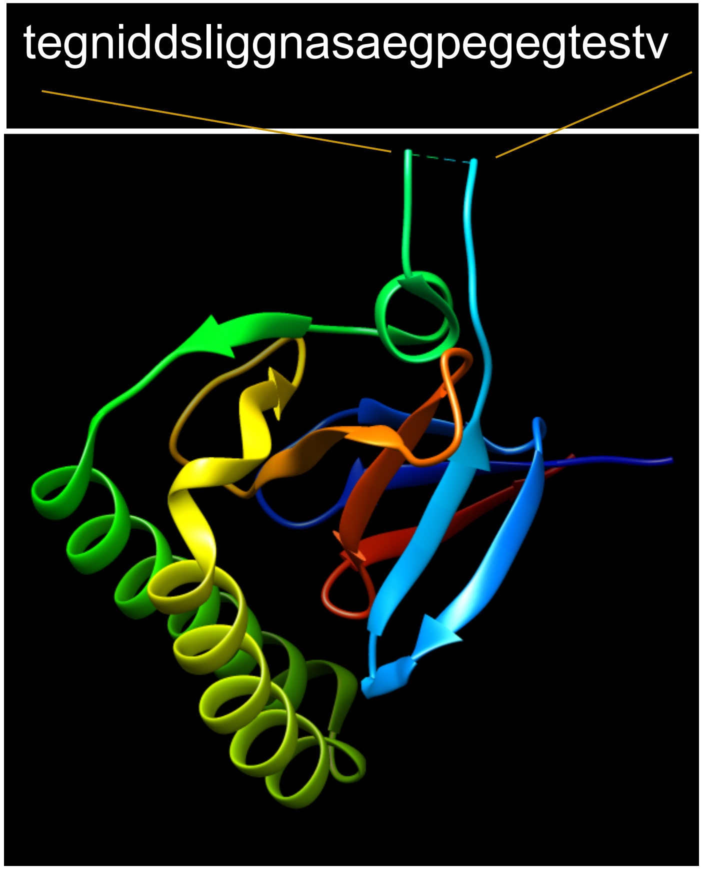




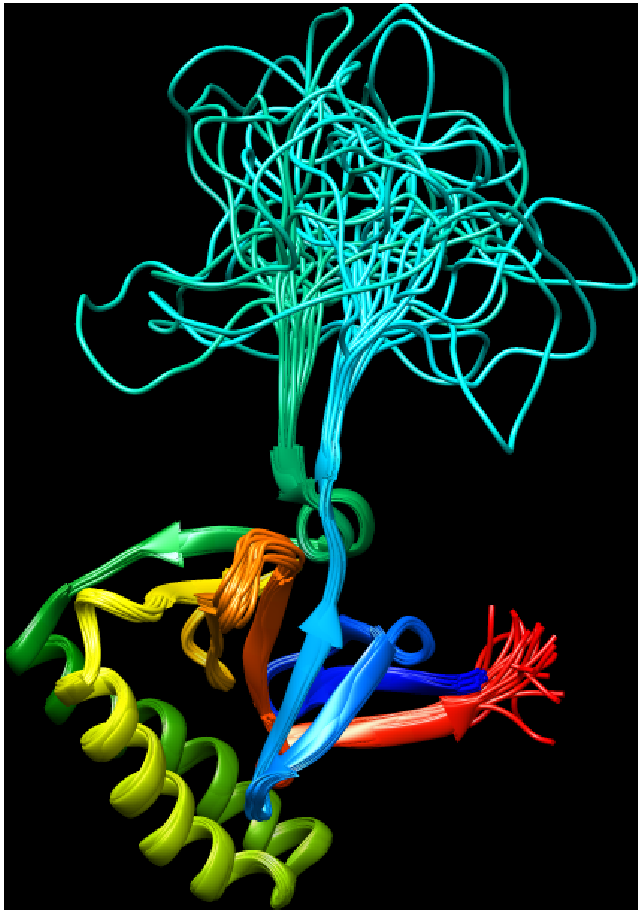
Conformational changes



Missing structure parts



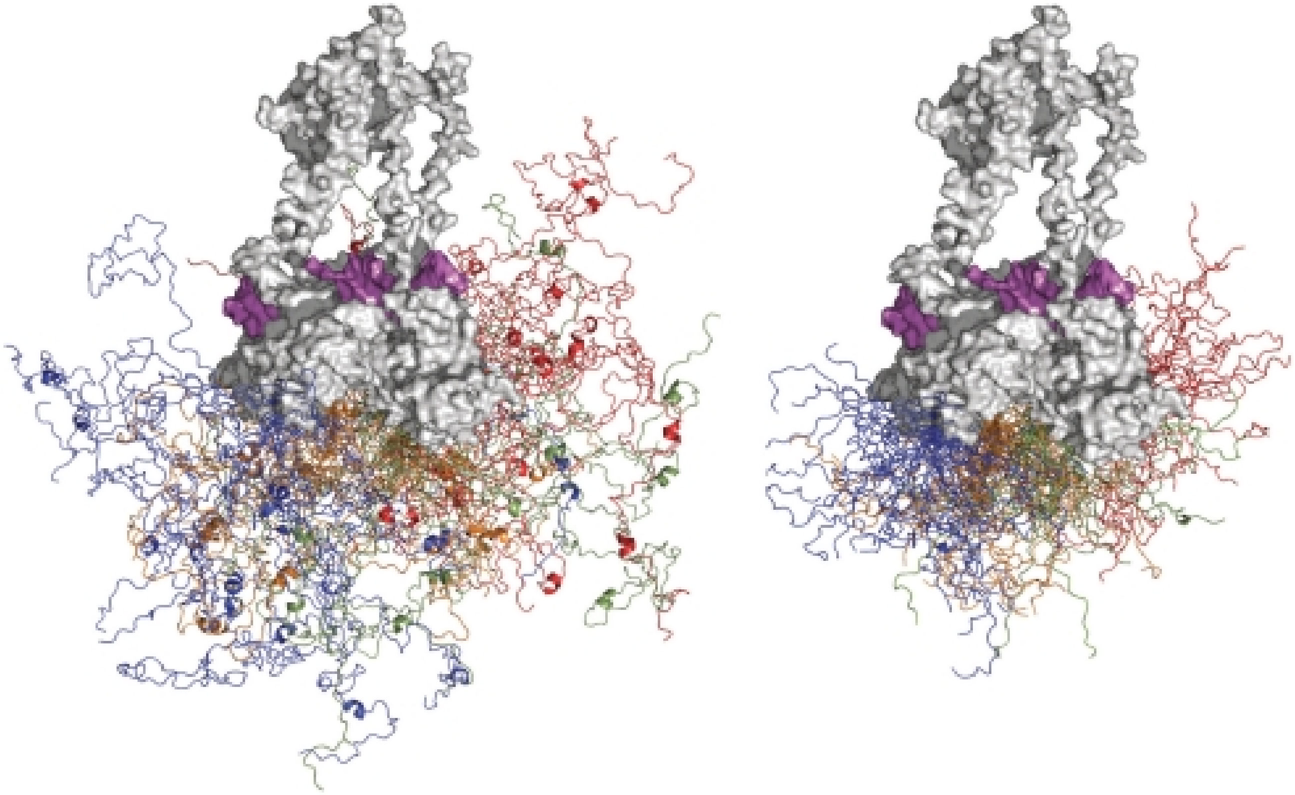
Missing regions in the protein structure



NMR structures with high structural variability

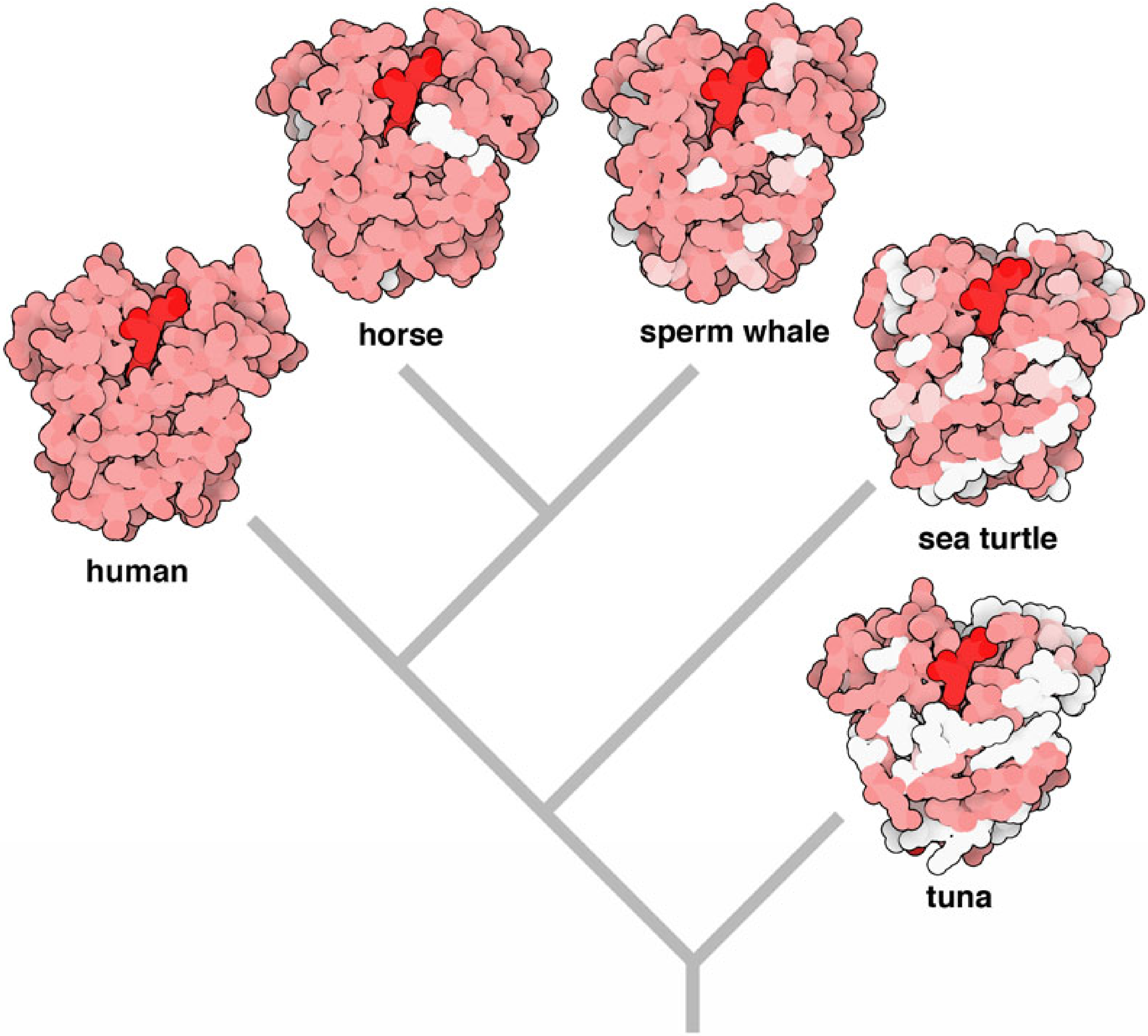
Intrinsically disordered proteins

Do not form a well-defined structure on their own under native(-like) conditions

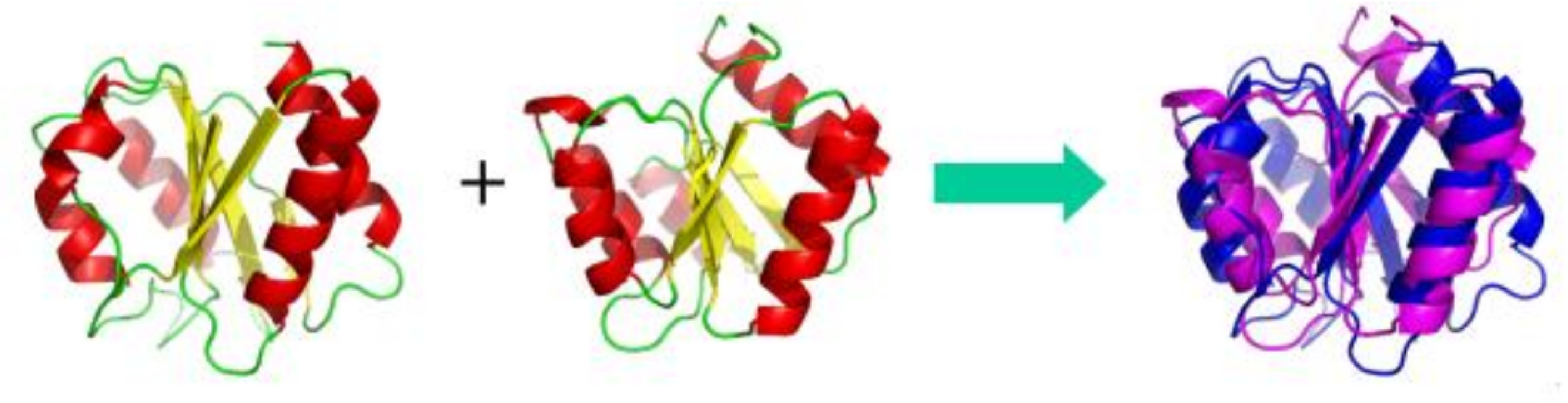


Human p53

Globin evolution

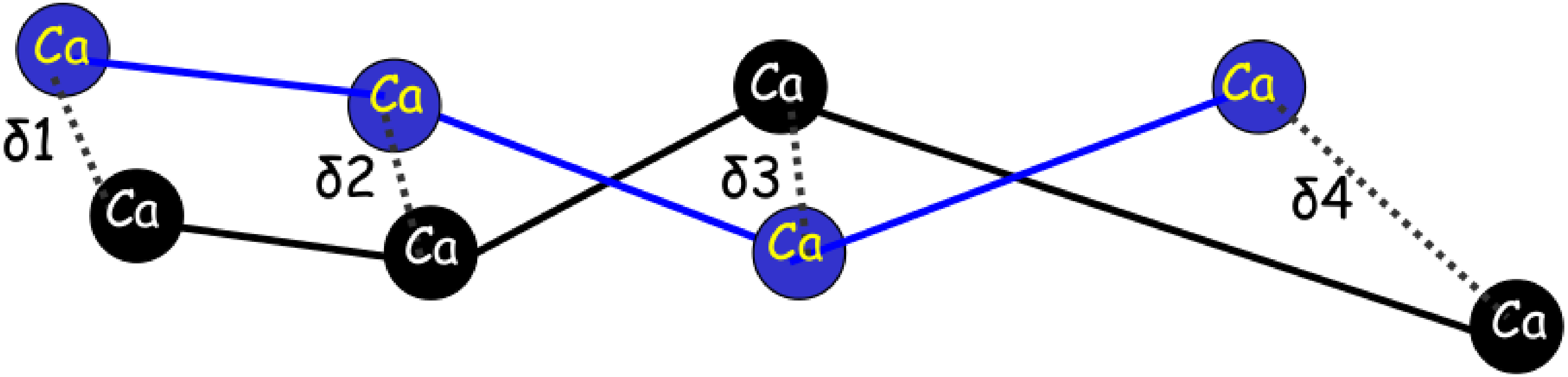


Similarity between two structures



Superposition: minimizing distances between positions

RMSD



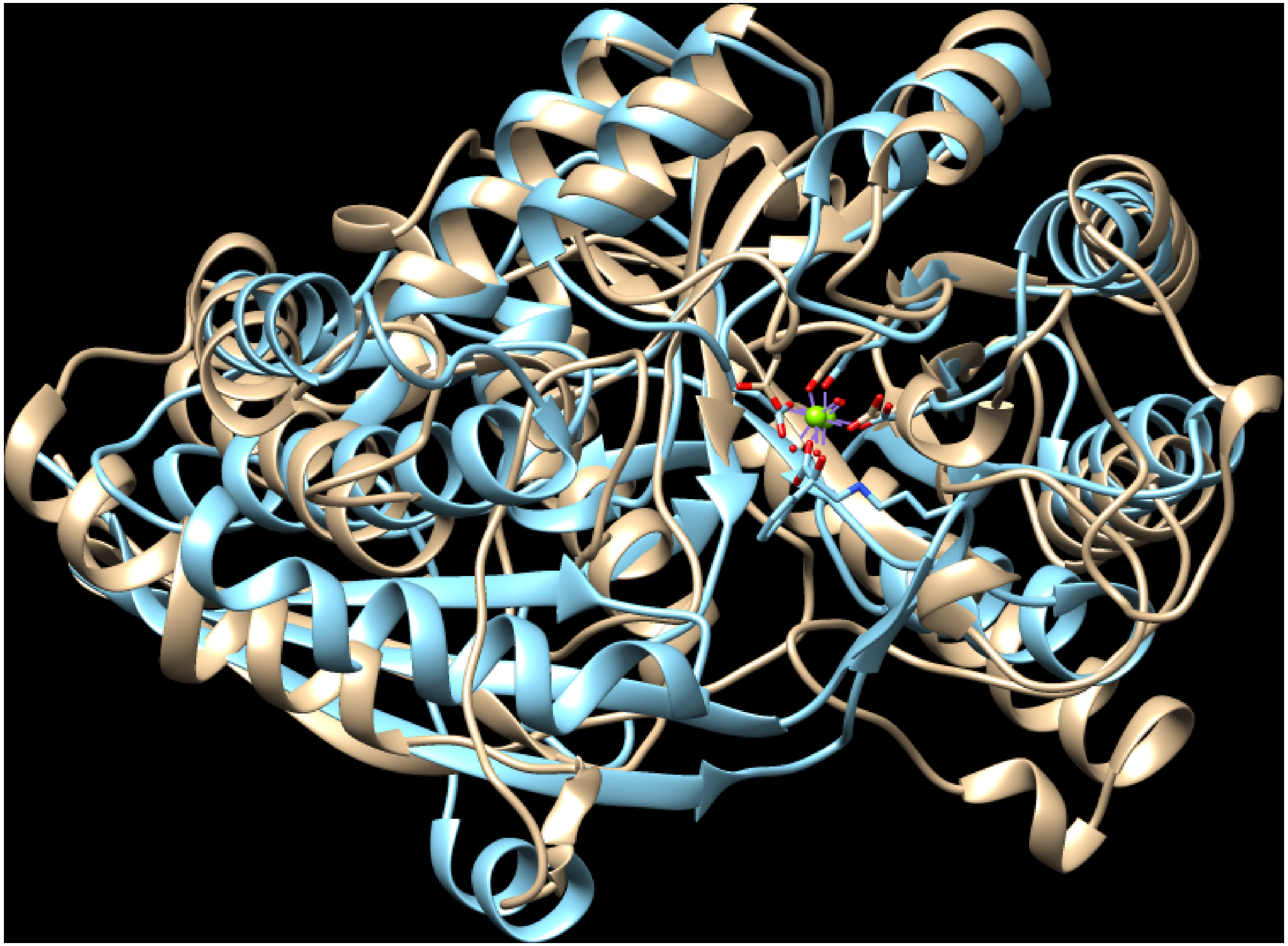
Root Mean Square Deviation (RMSD):

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

The most commonly used function for measuring structural similarity

RMSD is the average distance between equivalent atoms of superimposed structures

Structures of evolutionarily related proteins are usually similar



1ebhA: enolase
1mns : mandelate racemase

Sequence identity: 25%
Active center is very similar
Simlar chemical reactions
Different substrate

Sequence-structure relationship

- The structure is usually more conserved than the sequence
  - Structures typically tolerate more mutations
  - Due to physical effects some structures are more common
  - Analogue

- The number of folds is limited
  - Currently around 1,200 folds

Structural classification

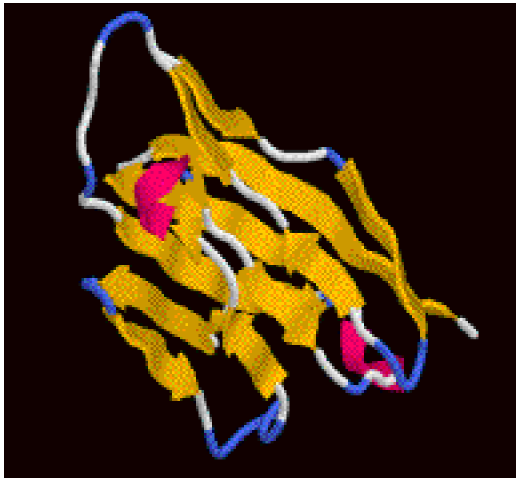
We can group similar and evolutionarily related protein structures using classification

Example

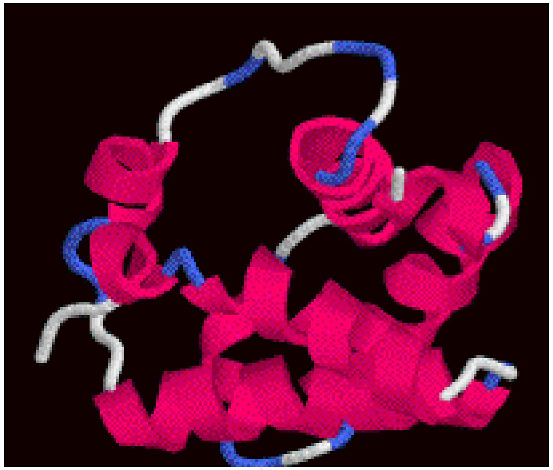
- CATH
    - http://www.cathdb.info/
- SCOP
    - http://scop2.mrc-lmb.cam.ac.uk/

Structural classes

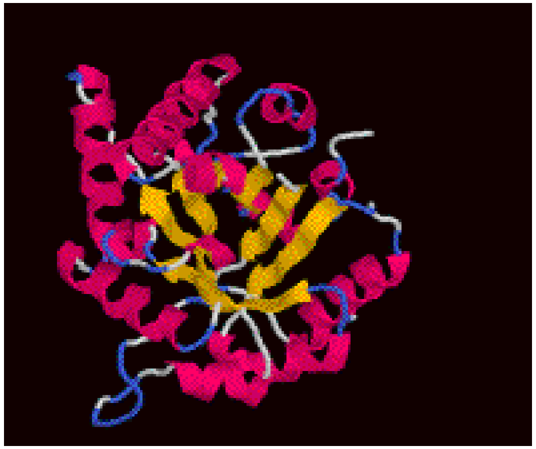
All β



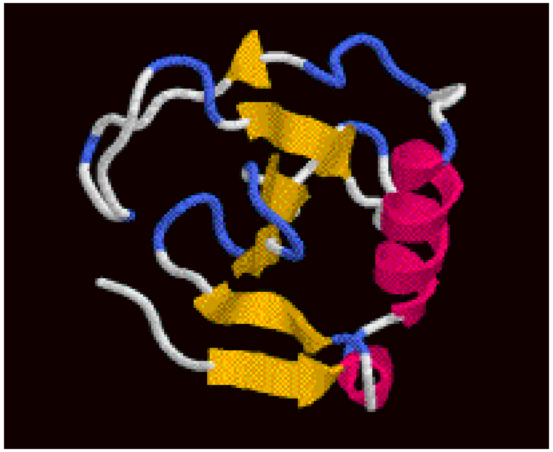
All α



α/ β



α+β

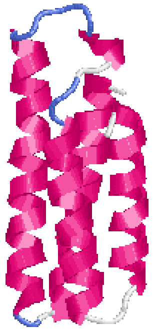


Fold - topology

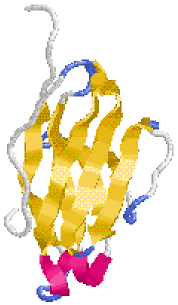
Proteins belonging to the same fold contain roughly the same secondary structure elements in the same order and similar spatial configuration.

globin

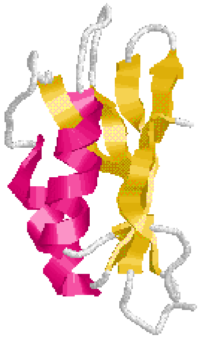
trefoil



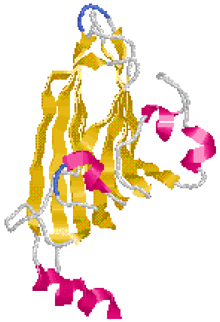
up-down



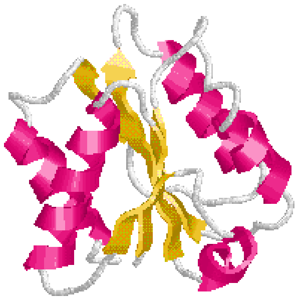
immunoglobulin



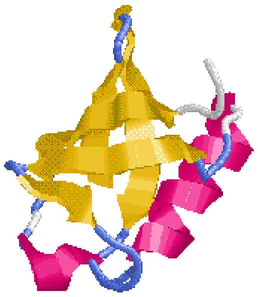
α/β sandwich



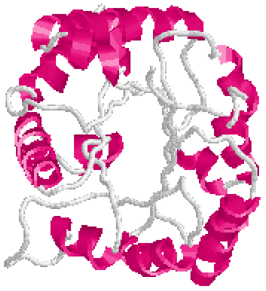
jelly roll



doubly wound



UB α/β roll



TIM barrell

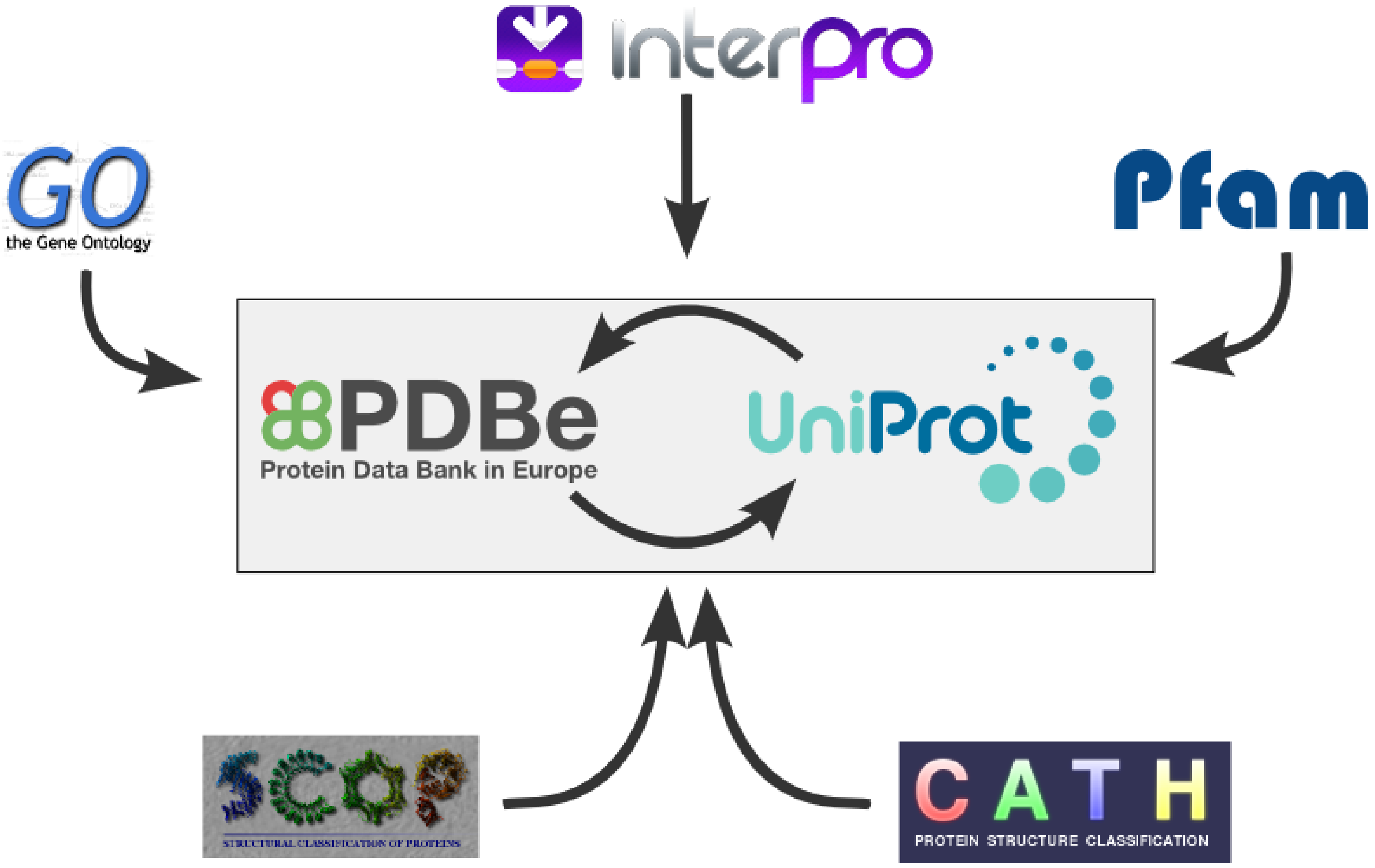
Homolous and analogous structures

Homolous proteins evolved from a common ancestor via divergence, and share the same fold

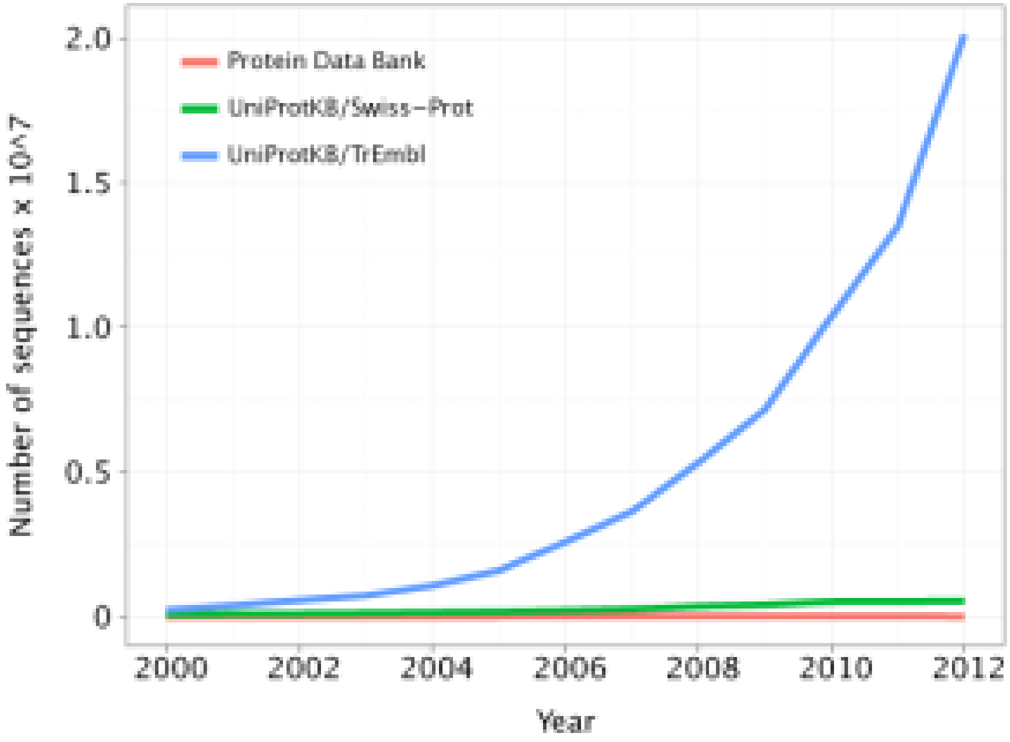
Analogous proteins share the same fold but do not have and evolutionary relationship (or it is undetectable)

- Some folds are more common (due to physical effects)
- Number of folds is limited (1-2,000 folds)

Cross-references with other databases


InterPro
GO
the Gene Ontology
Pfam
PDBe
Protein Data Bank in Europe
UniProt
SCOP
STRUCTURAL CLASSIFICATION OF PROTEINS
CATH
PROTEIN STRUCTURE CLASSIFICATION

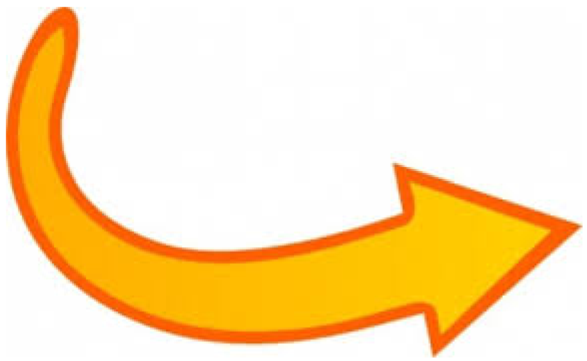
Sequence-structure gap

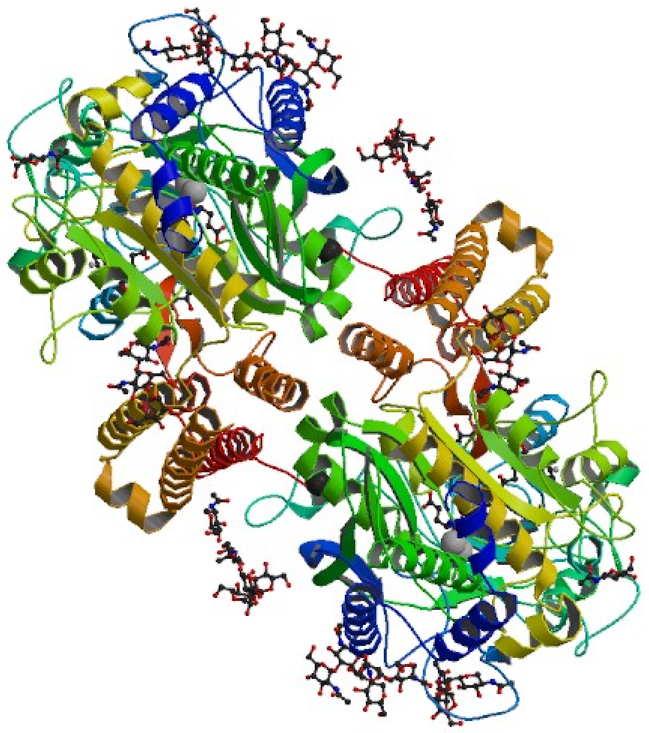


	2014
Sequences	50,000,000
Structures	100,000

Tertiary structure predictions

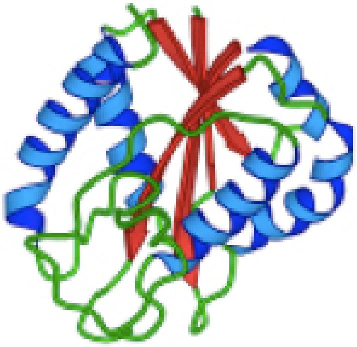
```
>Protein
RSKSSNEATNITPKHNMKAFLDELKAENIKKFLYNFTQIPHLAGTEQNFQLAKQIQSQWKEFGLDSVELAHYDVLLSYPN
KTHPNYISIINEDGNEIFNTSLFEPPPPGYENVSDIVPPFSAFSPQGMPEGDLVYVNYARTEDFFKLERDMKINCSGKIV
IARYGKVFRGNKVKNAQLAGAKGVILYSDPADYFAPGVKSYPDGWNLPGGGVQRGNILNLNGAGDPLTPGYPANEYAYRR
GIAEAVGLPSIPVHPIGYYDAQKLLEKMGGSAPPDSSWRGSLKVPYNVGPGFTGNFSTQKVKMHIHSTNEVTRIYNVIGT
LRGAVEPDRYVILGGHRDSWVFGGIDPQSGAAVVHEIVRSFGTLKKEGWRPRRTILFASWDAEEFGLLGSTEWAEENSRL
LQERGVAYINADSSIEGNYTLRVDCTPLMYSLVHNLTKELKSPDEGFEGKSLYESWTKKSPSPEFSGMPRISKLGSGNDF
EVFFQRLGIASGRARYTKNWETNKFSGYPLYHSVYETYELVEKFYDPMFKYHLTVAQVRGGMVFELANSIVLPFDCRDYA
VVLRKYADKIYSISMKHPQEMKTYSVSFDSLFSAVKNFTEIASKFSERLQDFDKSNPIVLRMMNDQLMFLERAFIDPLGL
PDRPFYRHVIYAPSSHNKYAGESFPGIYDALFDIESKVDPSKAWGEVKRQIYVAAFTVQAAAETLSEVA
```





Protein folding

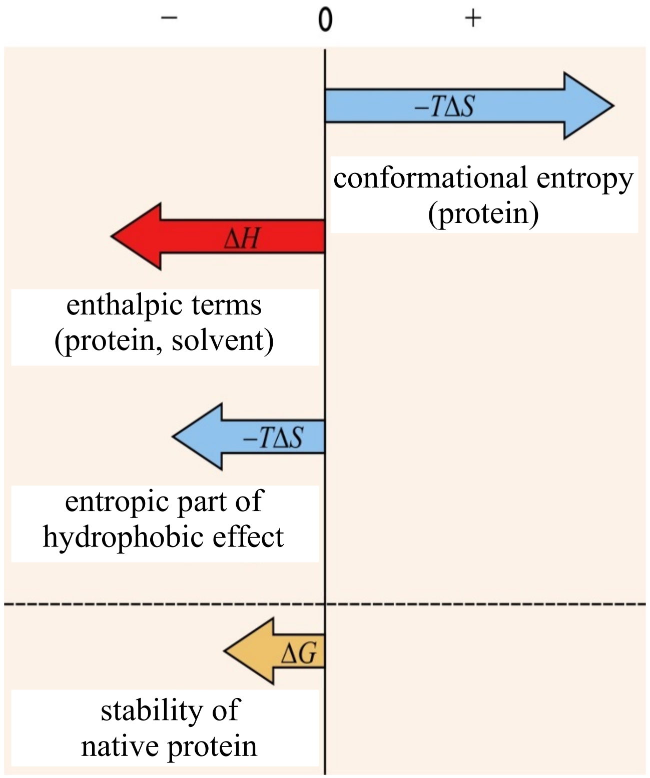
GFCHIKAYTRLIMVG…



Folding

(physics)

$\Delta G = \Delta H - T\Delta S$



conformational entropy
(protein)

enthalpic terms
(protein, solvent)

entropic part of
hydrophobic effect

stability of
native protein

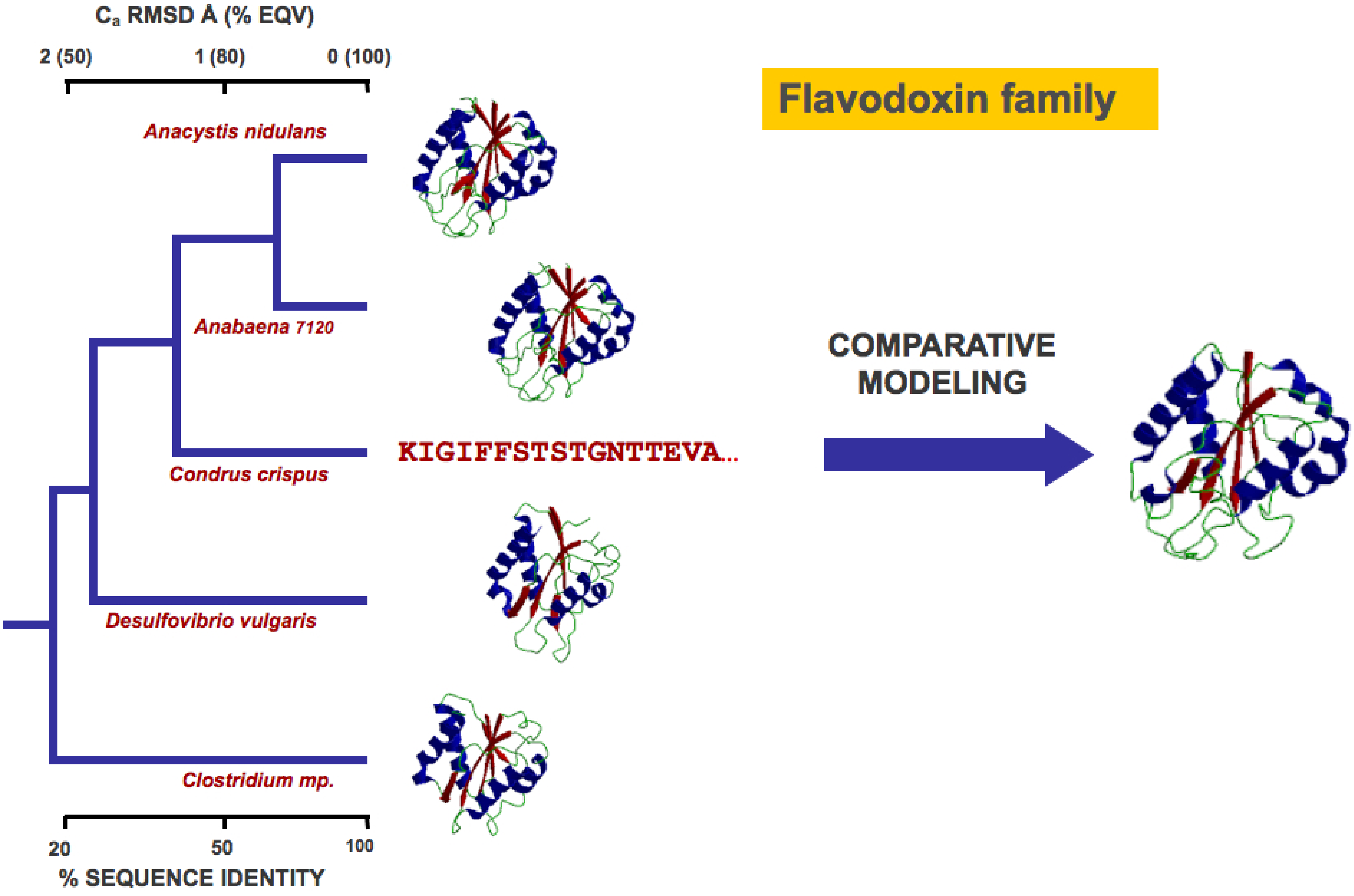
Determining tertiary structure based on physical principles

- large number of comformations, huge conformational space

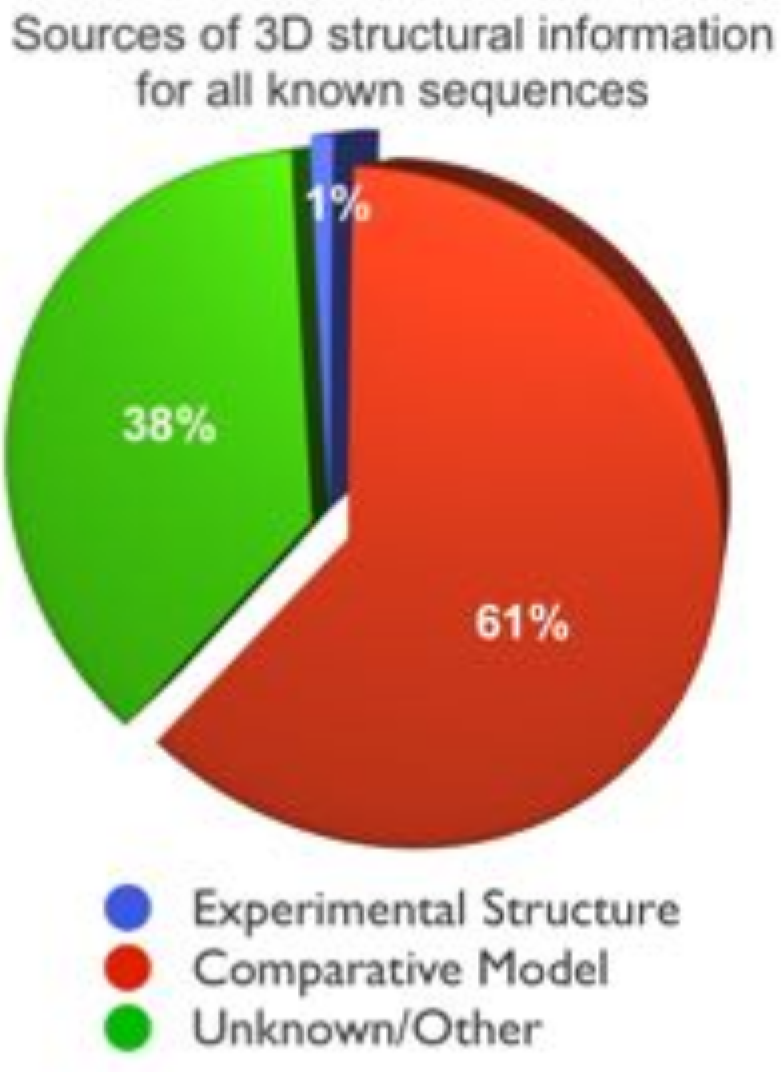
- the physical energy function is not known exactly

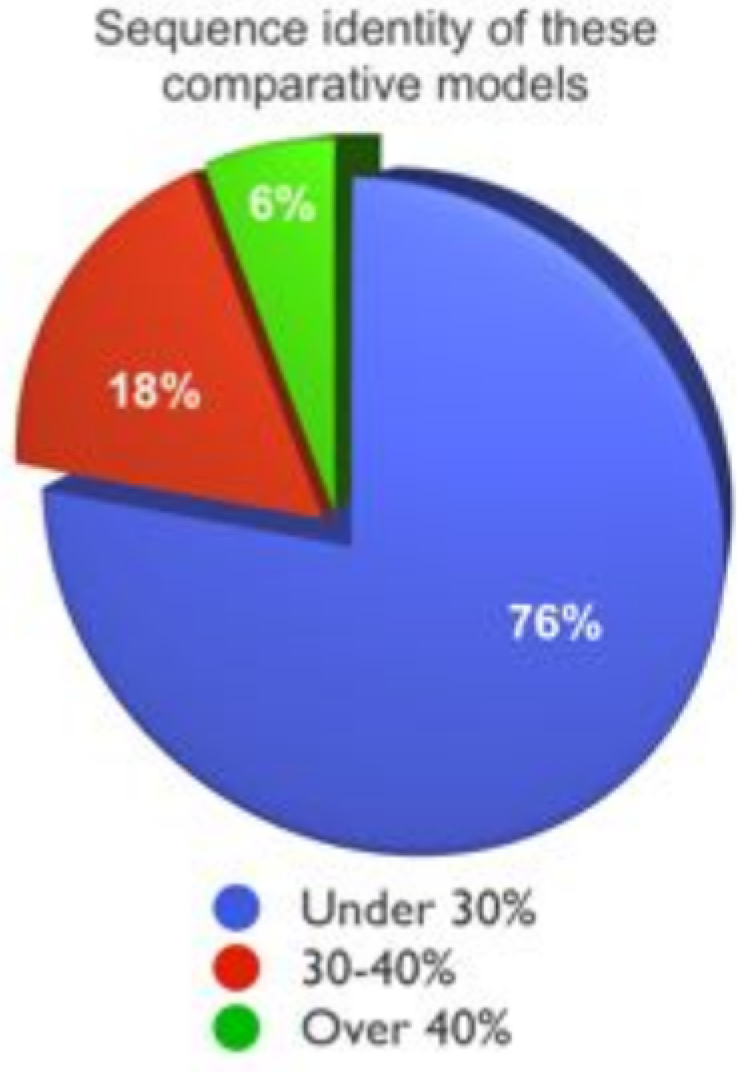
Comparative structure modeling

Flavodoxin family

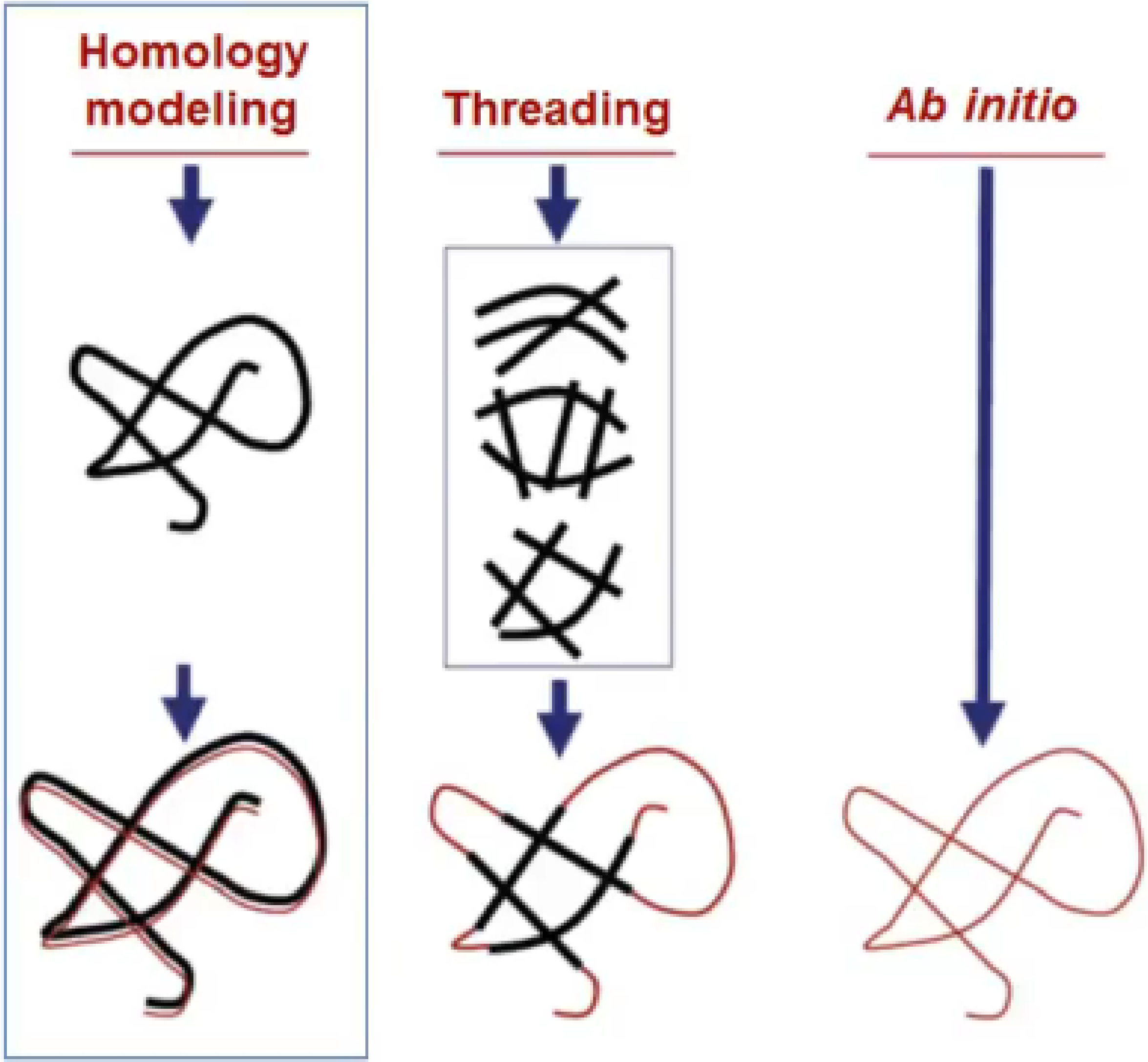


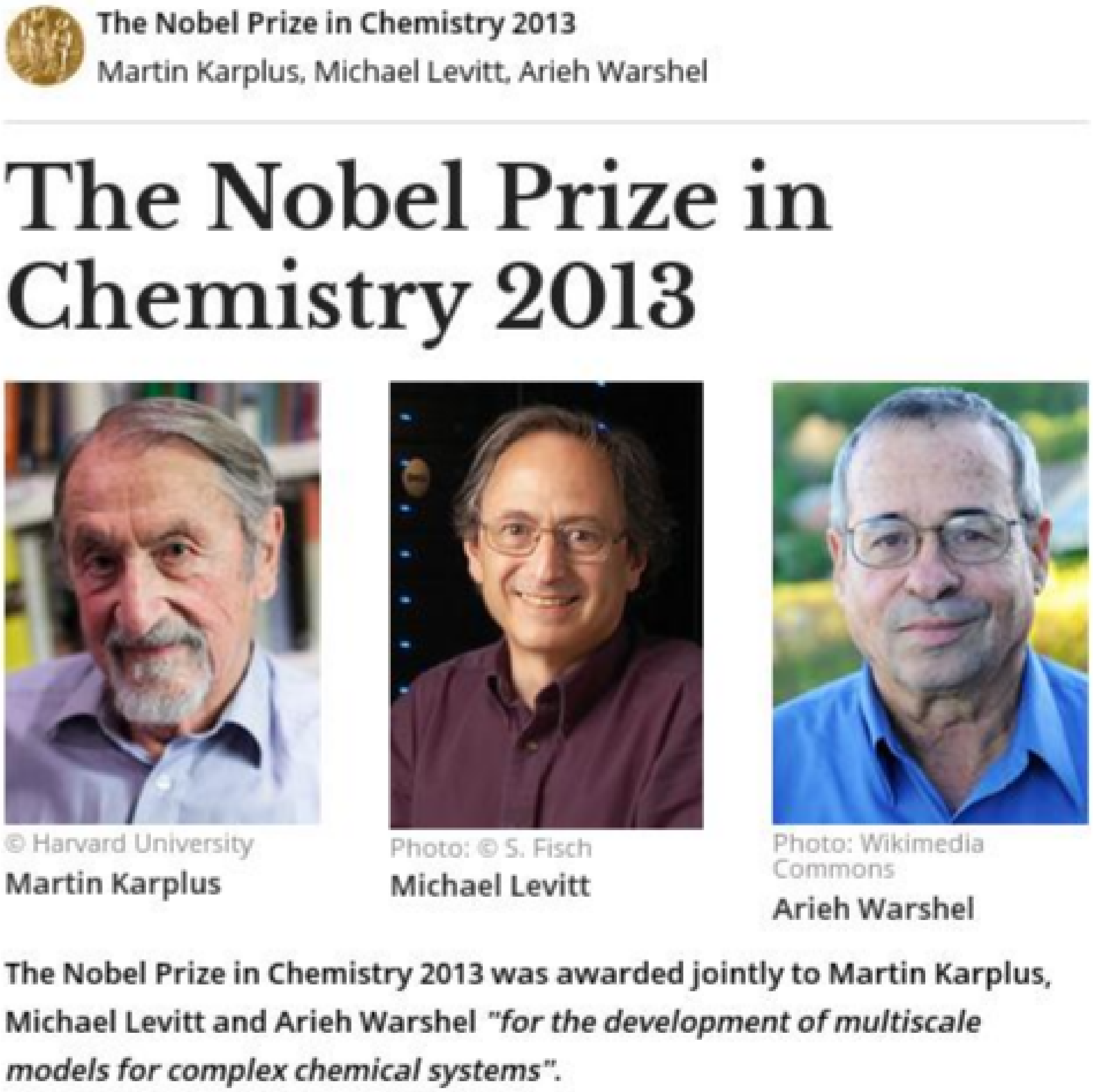
Structural coverage


Sources of 3D structural information for all known sequences


Sequence identity of these comparative models

Tertiary structure prediction approaches





The Nobel Prize in Chemistry 2013
Martin Karplus, Michael Levitt, Arieh Warshel

The Nobel Prize in Chemistry 2013

© Harvard University
Martin Karplus

Photo: © S. Fisch
Michael Levitt

Photo: Wikimedia Commons
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

M.L.:

It's sort of nice in more general terms to see that computational science, computational biology is being recognized." He added, "It's become a very large field and it's always in some ways been the poor sister, or the ugly sister, to experimental biology."