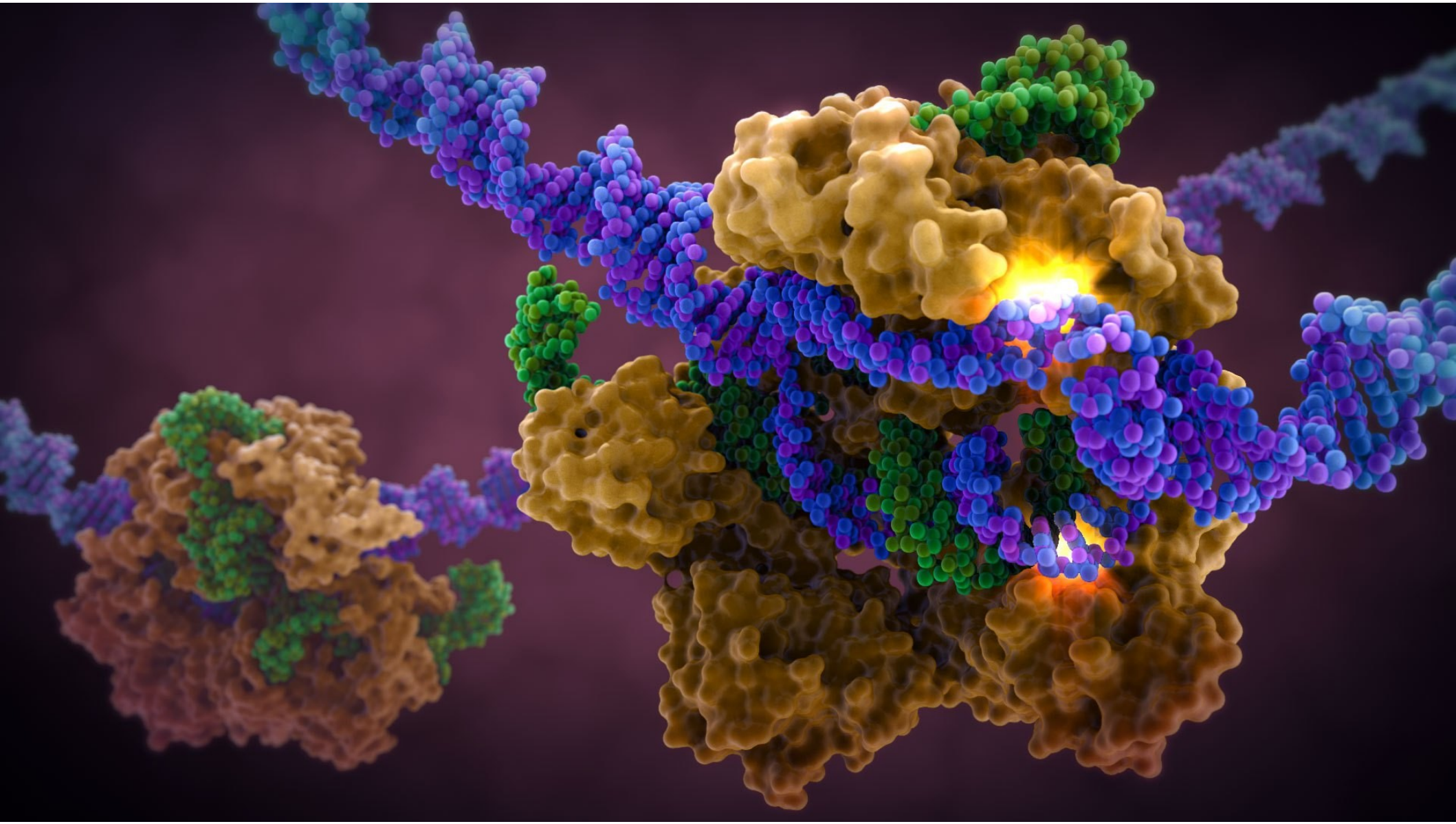
Protein structures



Erdős Gábor, 2017.01.11

Why are we interested?

Function is related to the **structure**

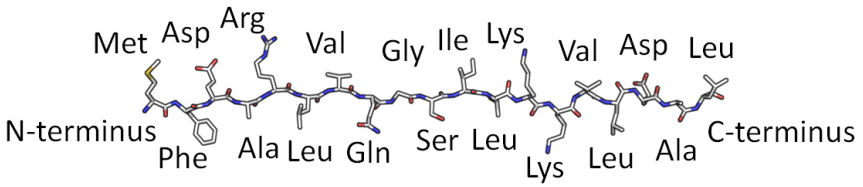
- Understand biological processes (DNA, RNS, enzymes, hormones, receptors)
- Diseases
- Drug design, protein – drug interactions

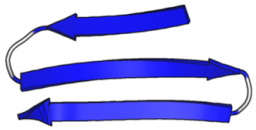
Primary

Secondary

Tertiary

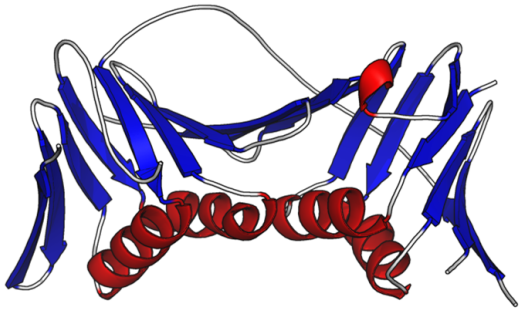
Quaternary

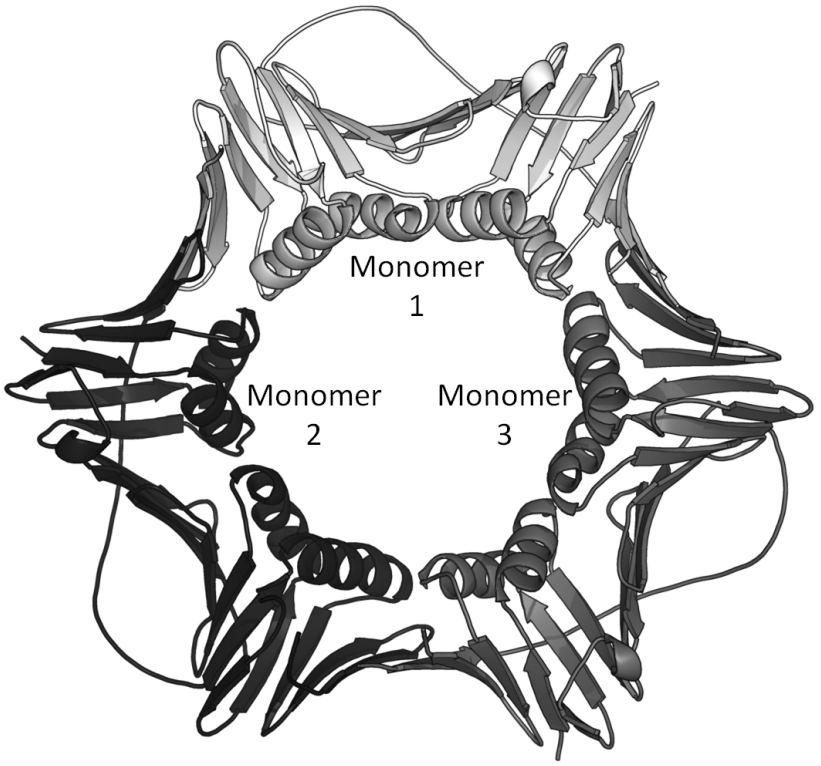




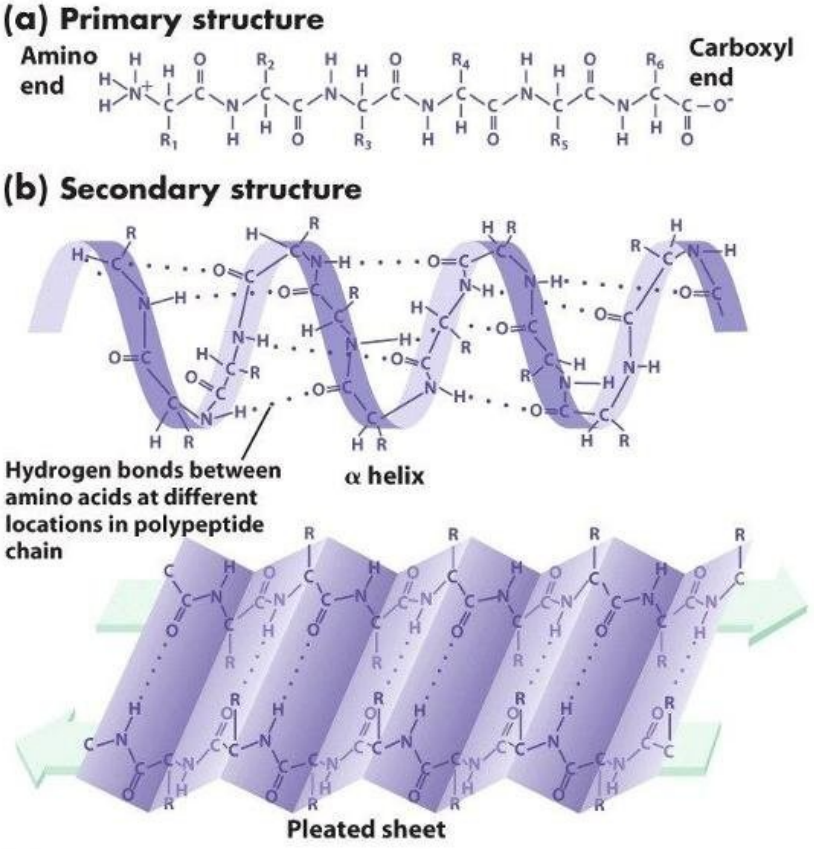
β-Sheet (3 strands)

α-helix



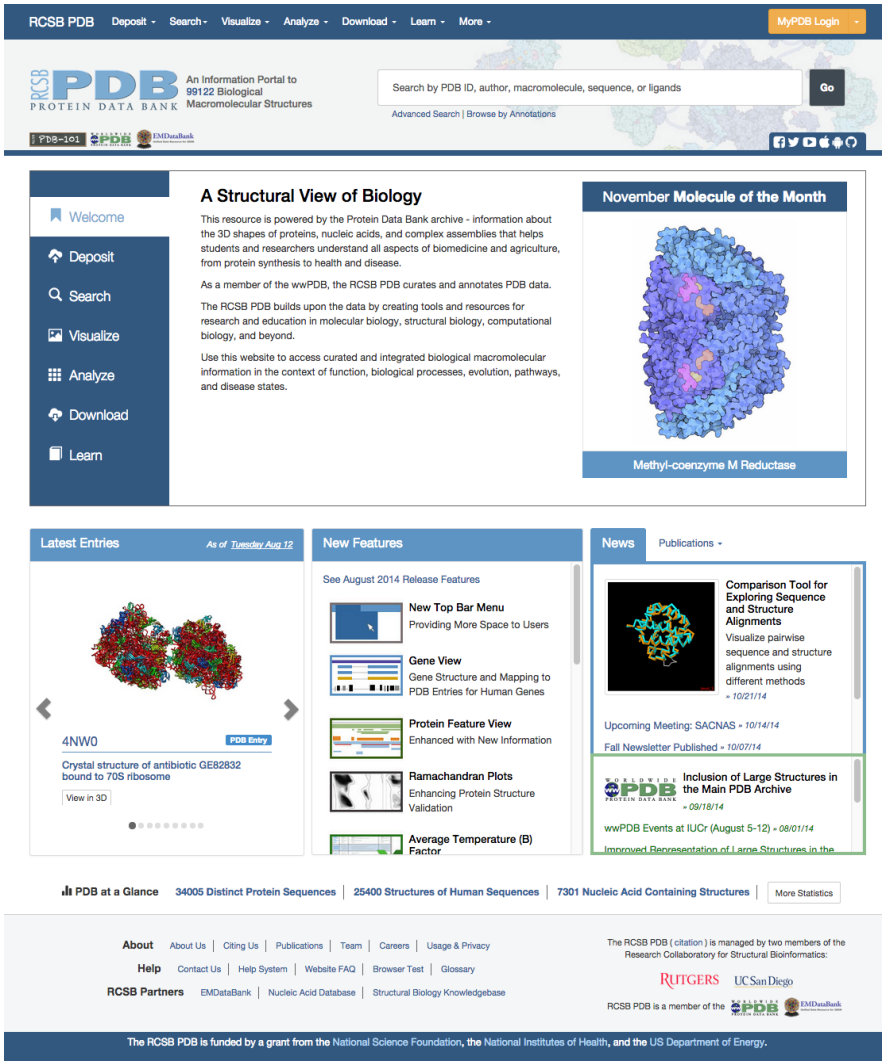


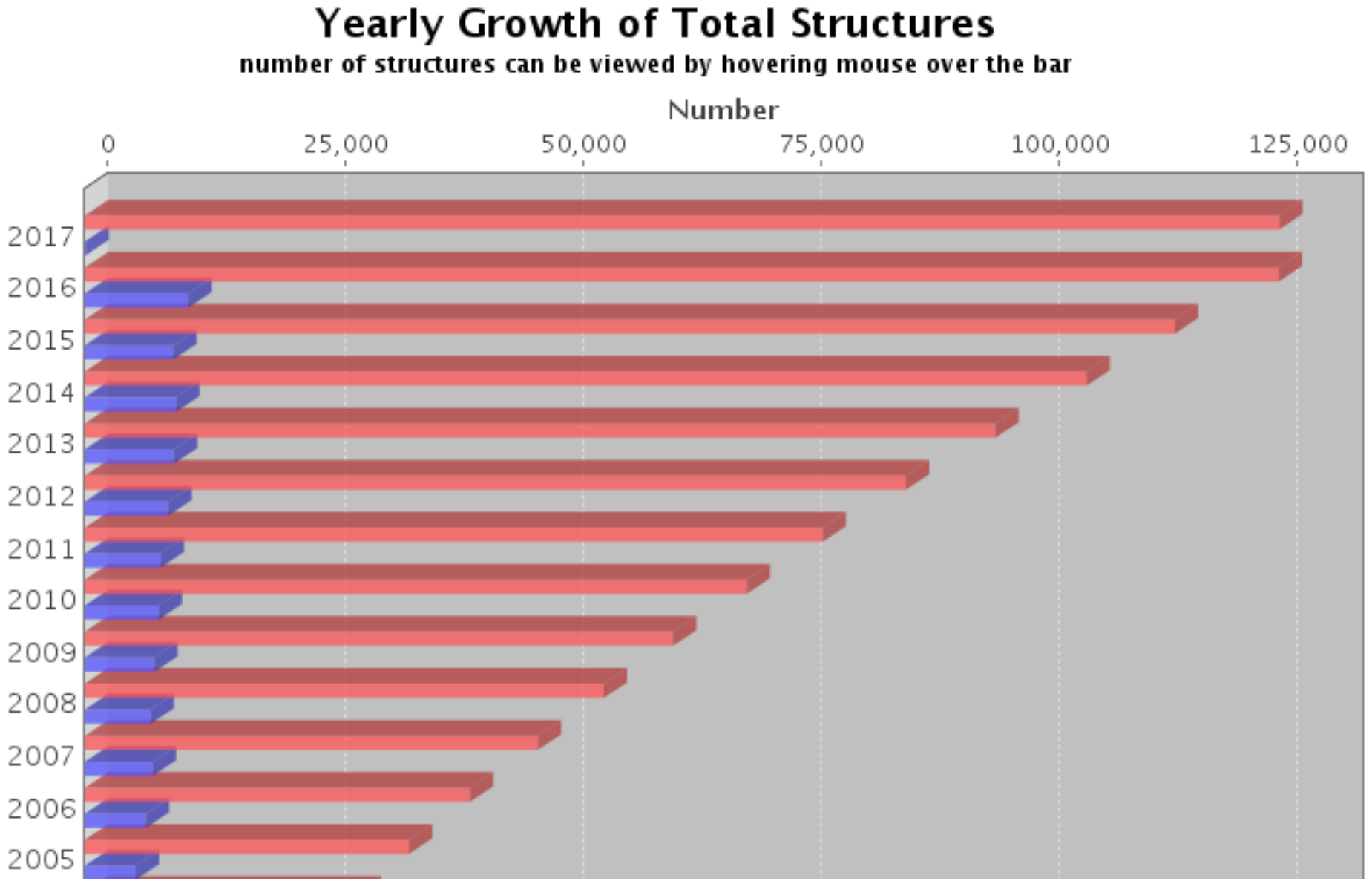
"**Protein structure** is the three-dimensional arrangement of atoms in a protein molecule"



How to find a specific structure?

- Database
- Wordwide Protein Data Bank (wwPDB)
- 3 entries:
  - PDB Europe
  - PDB Japan
  - **RSCB PDB**
    - Easy to use
    - Search by name, ID
    - Direct links
    - Well maintained




Yearly Growth of Total Structures
number of structures can be viewed by hovering mouse over the bar
Number
0
25,000
50,000
75,000
100,000
125,000
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005

Where do these structures come from?

Most structures are solved by x-ray crystallography (86%)

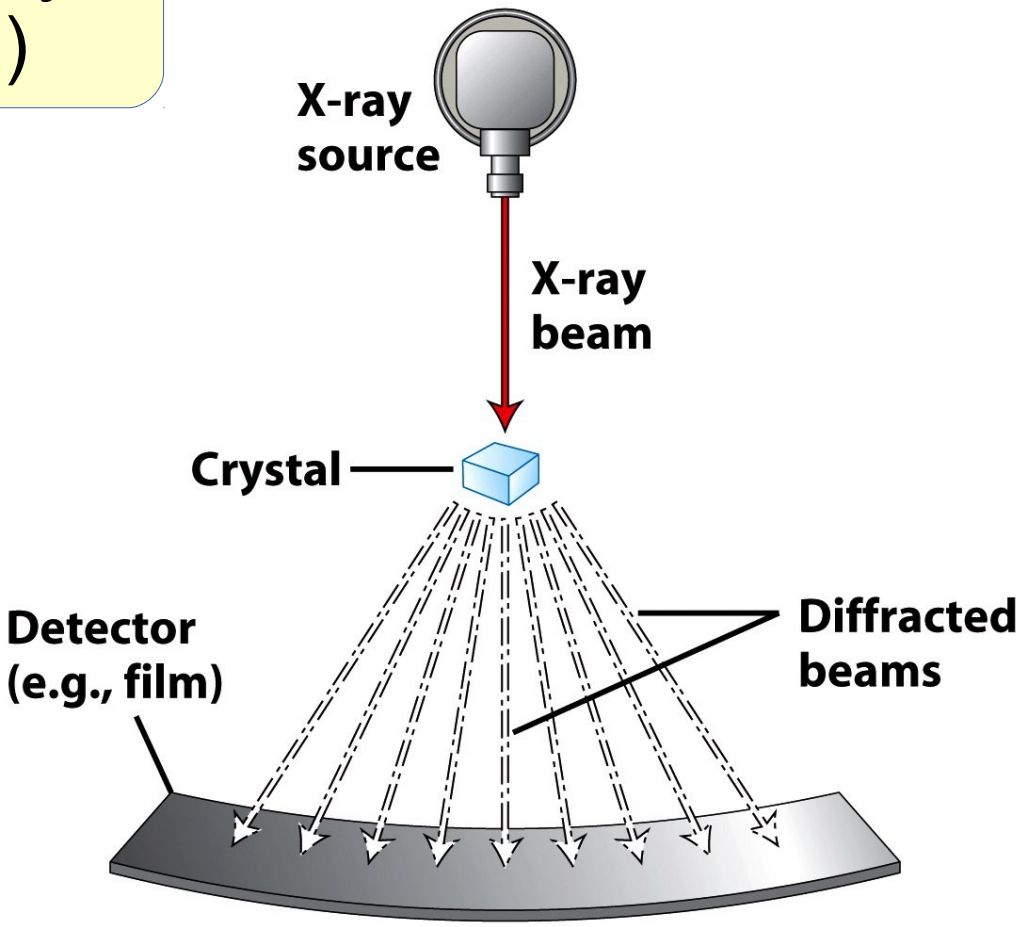
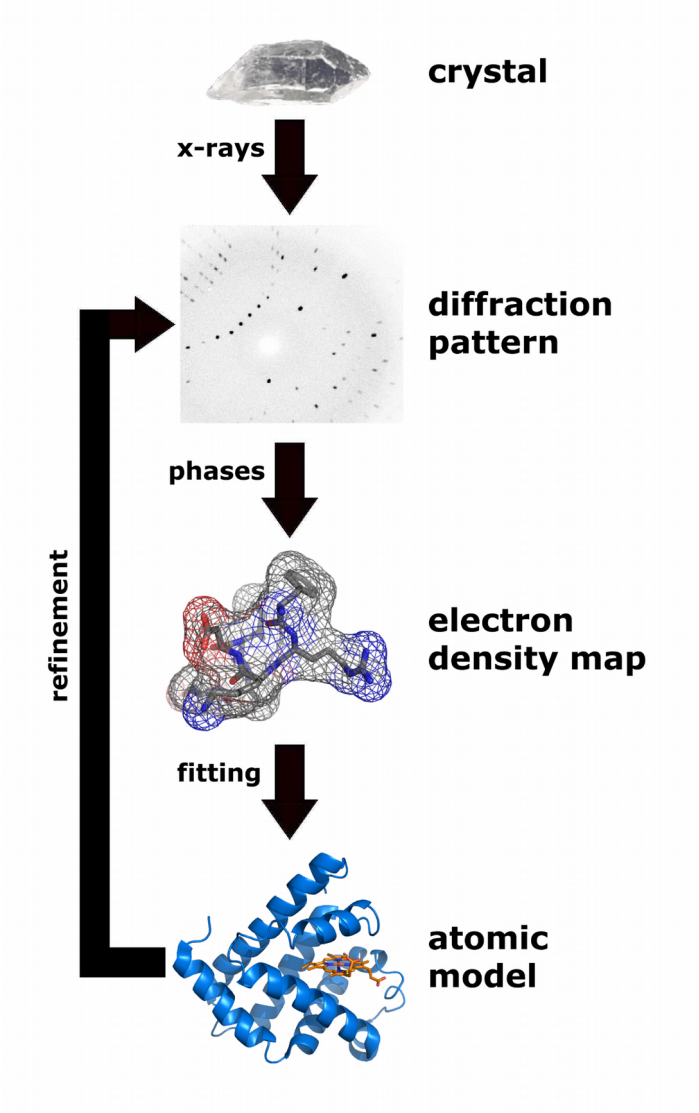


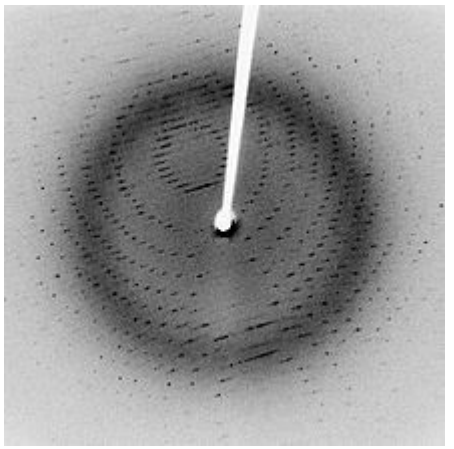
Figure 3-42a
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Where do these structures come from?

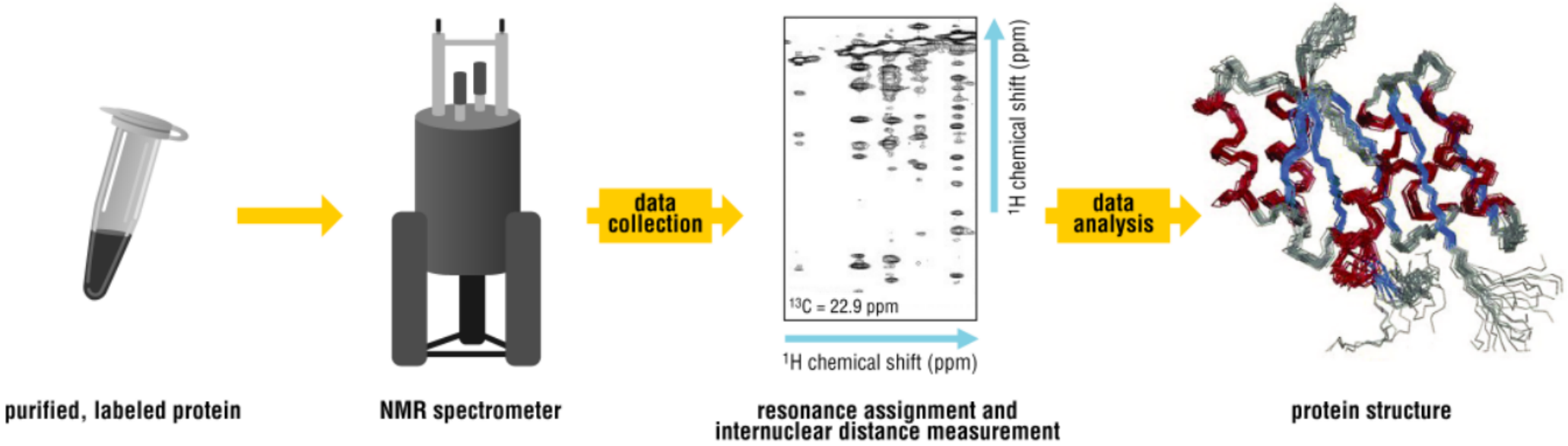
Most structures are solved by x-ray crystallography (86%)



- Short wavelength (ca. 1.5 Å), so we can measure the distance between atoms
- Brag equation
- The output is an electron density
- Crystallization artefacts
- Not physiological
- Hydrogens are not visible



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



Some with NMR (13%)

- Solution
- Usually a crowd of structures that fulfill the constrains
- Only small molecules (> 35kDa)
- Less accurate
- Can be used on flexible proteins

Very little by electron microscopy (1%)

TEM (Transmission Electron Microscopy)

- Capable of creating sub angstrom resolution images
- Until recently it was unable to work with biological samples
- Two techniques emerged which can deal with biological samples

Problems with biological samples

- EM is done in high vacuum
- Samples needs to be radiation resistant
- N, O, C (typical biological atoms) have the same EM scatter as water (usual solution)

Negative Staining

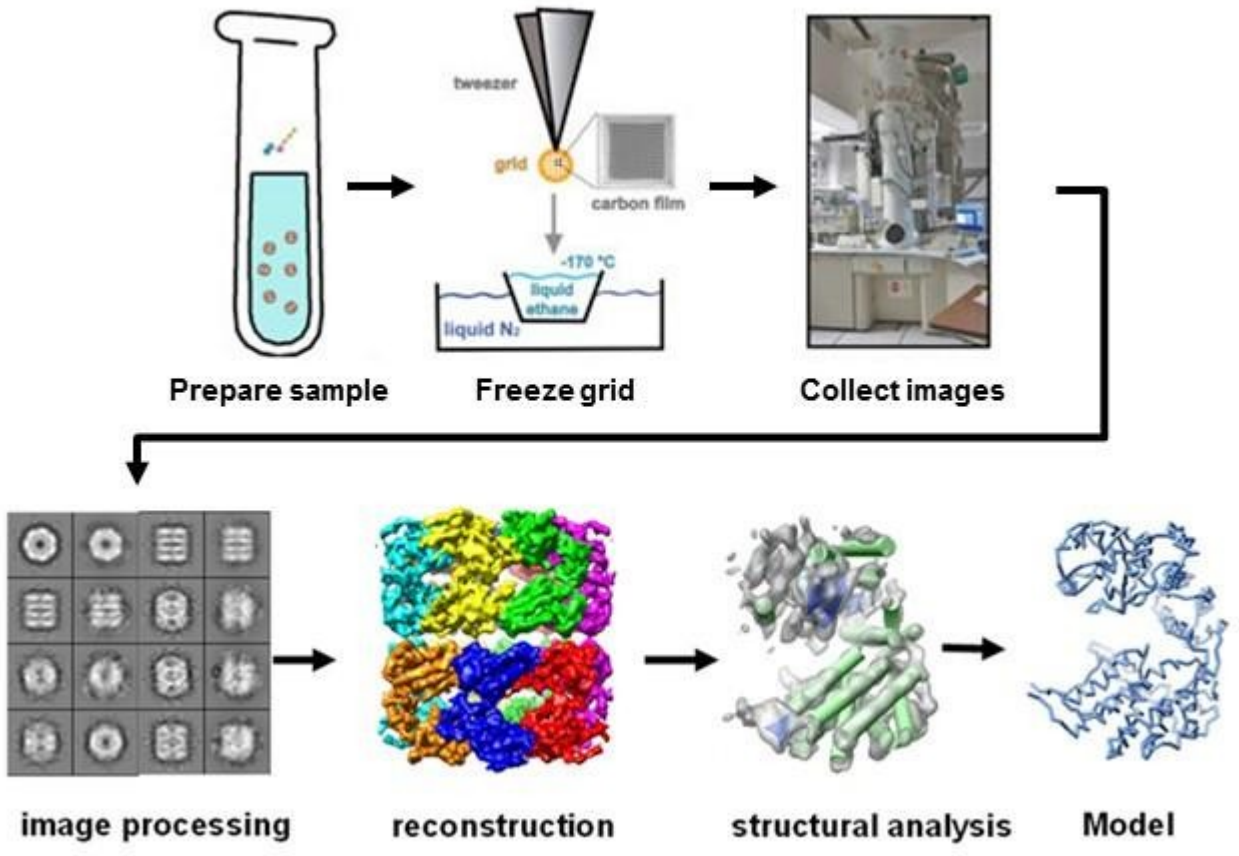
- Sample is embedded with a heavy salt (usually uranium)
- Dried to a thick layer in low concentration

This technique can overcome all three limitations, but the protein can lose its structure due to the drying

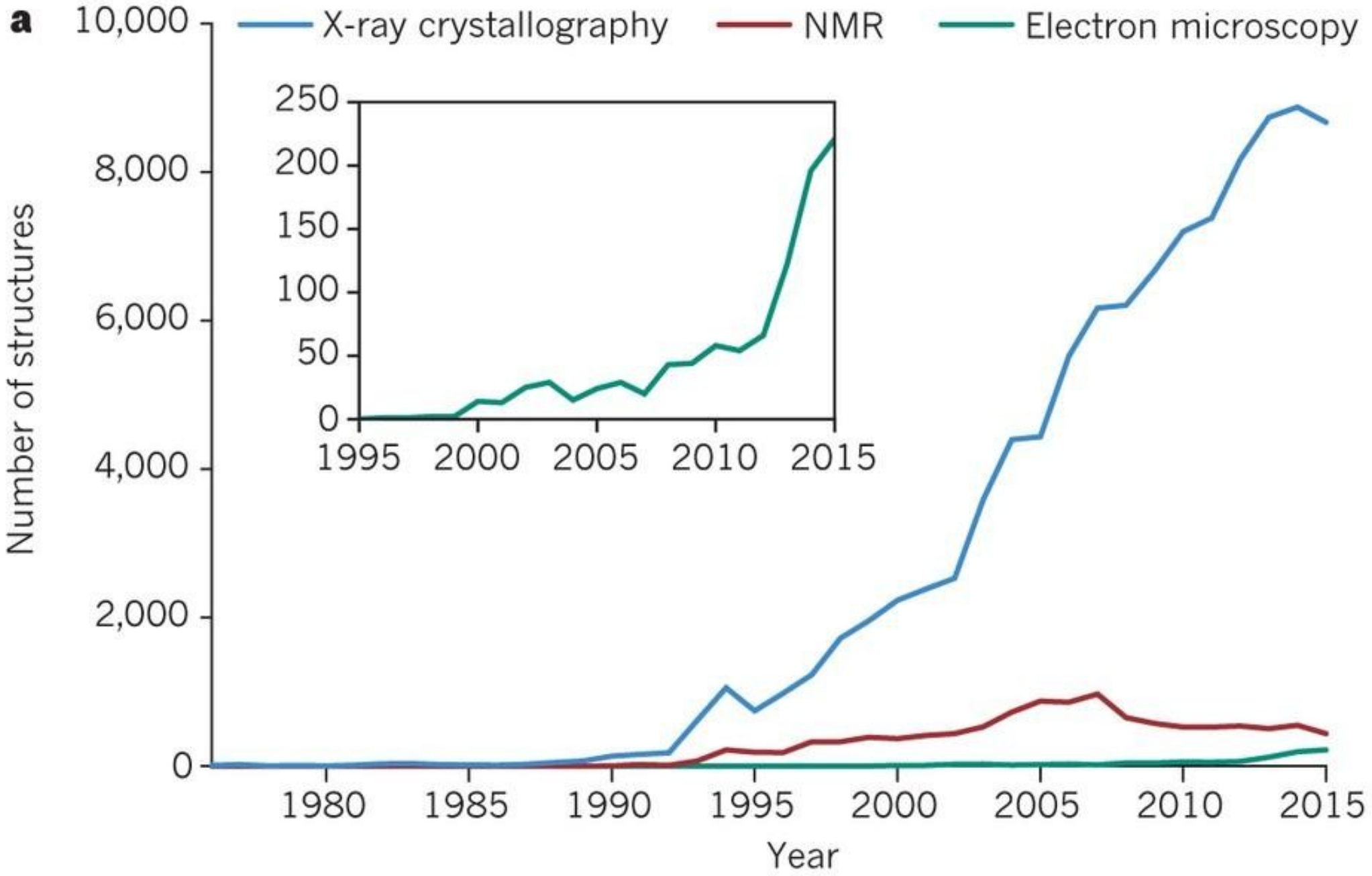
Cryo – EM

- Normal solution of the protein is used
- Solution is **quickly** frozen
- Water cannot form a crystal
  → vitrified water

Vacuum resistance	✓
Radiation resistance	✓
Contrast	✗



https://www.ibiology.org/ibioseminars/techniques/eva-nogales-part-1.html


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

How to find a structure?

In the PDB each atomic coordinate file has a unique ID which contains 4 specific characters First is always a number, a followed by 3 characters that can be either letters or numbers.

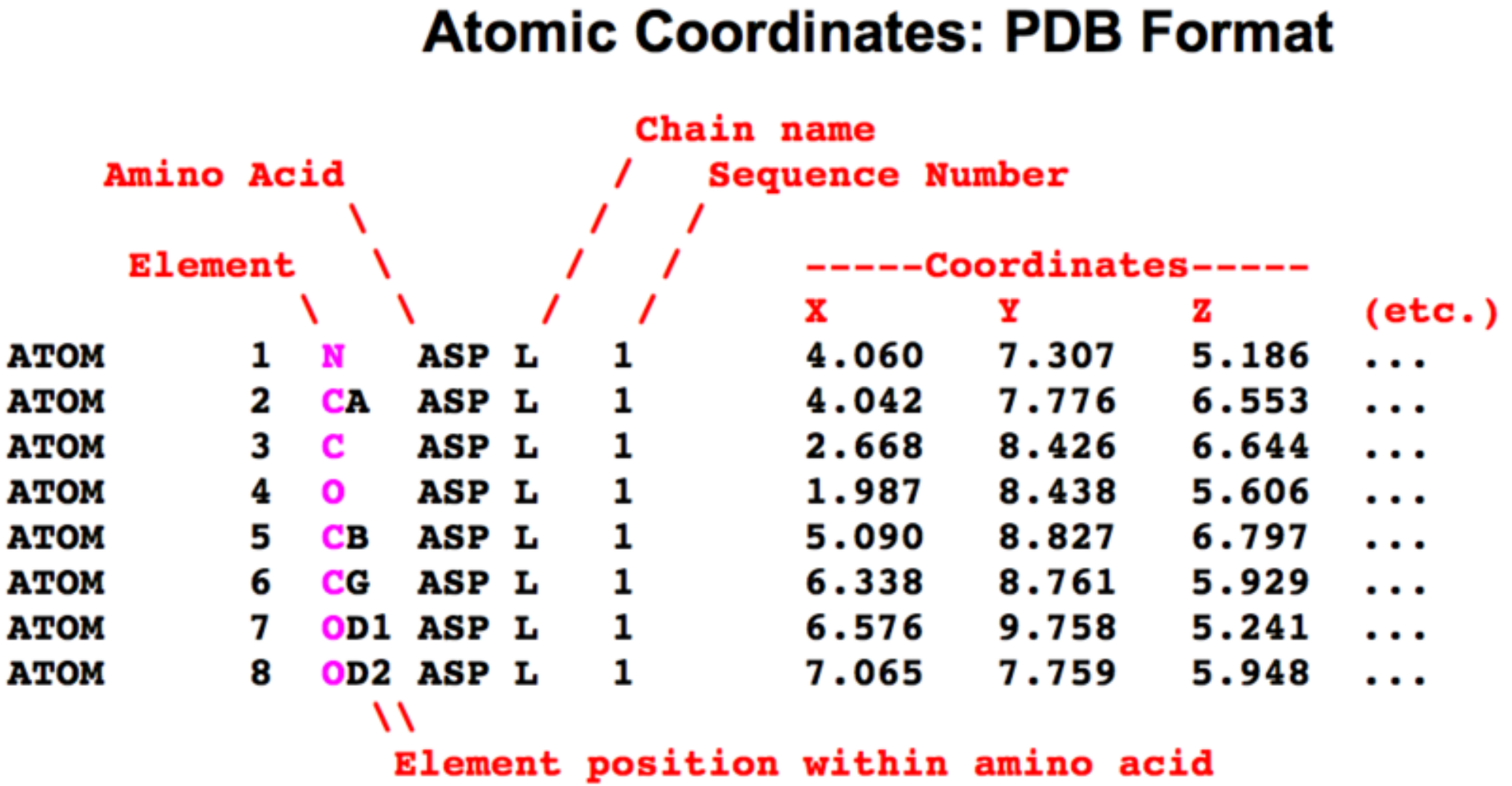
F.e.:

- 1mbn - The very first structure from 1973, the **myoglobin**
- 1tna - Form 1975 yeast phenylalanine transfer RNS, the first RNS structure
- 1bna - The first DNA double helix solved in 1980 with X-ray, (confirmation of the Watson & Crick modell from 1953).
- 2hhd - human **hemoglobin,** (deoxy form)
- 9ins - **insulin**

The .pdb file format

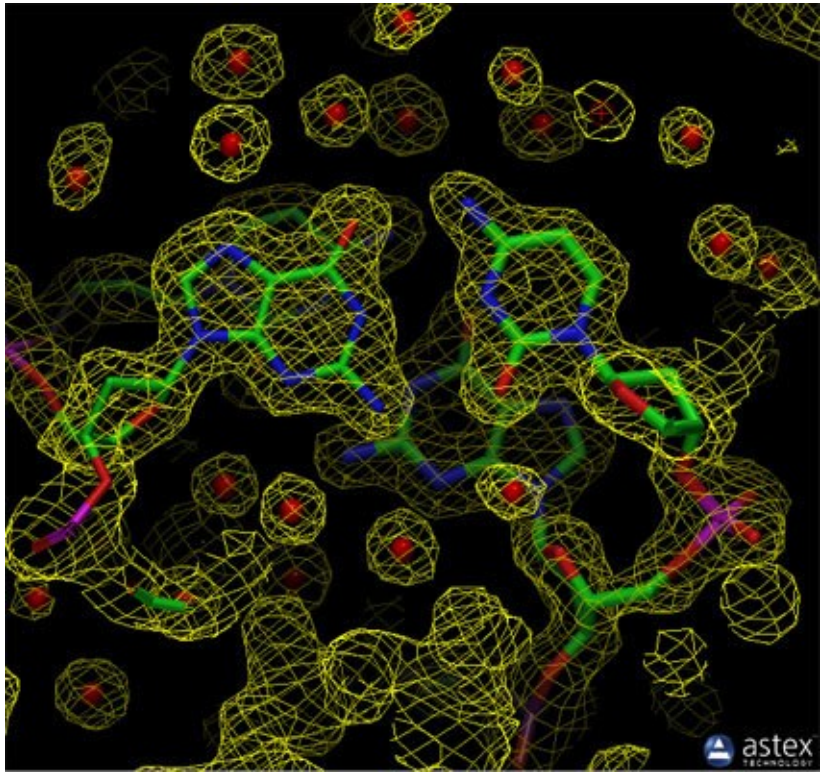
```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   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
...
```

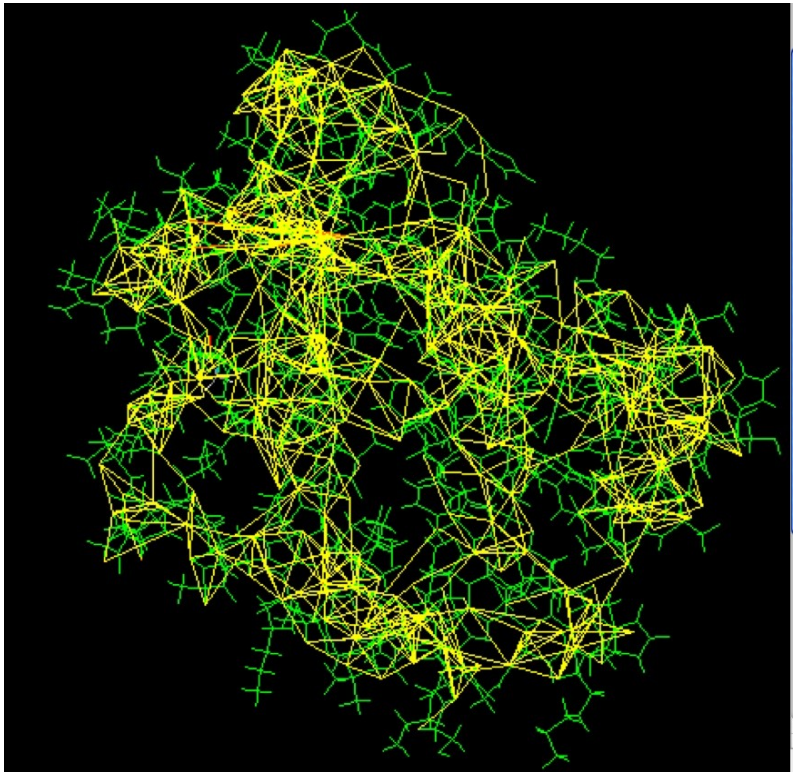
The .pdb file format



Model

The model obtained with NMR or X-ray is not a direct one. Through the measurement we derive the model from a collection of data (NMR spectrum or diffraction pattern).



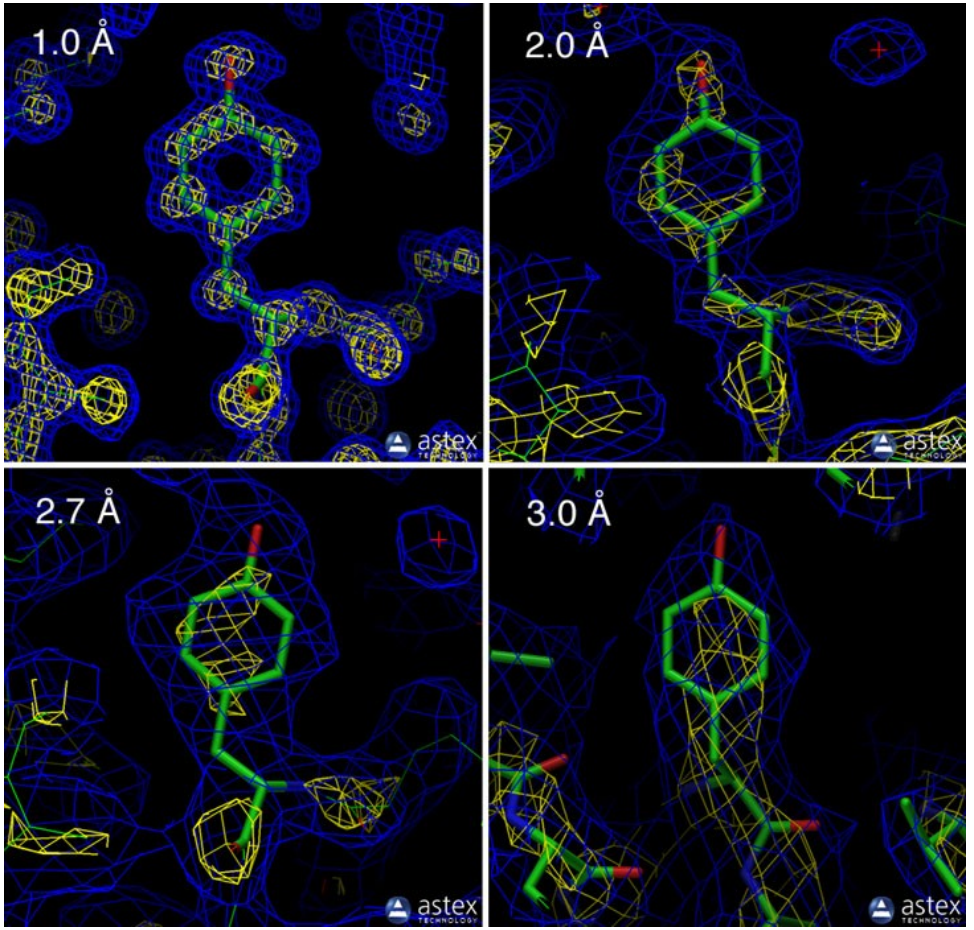


Resolturion

- Indicates how much can we trust in a position of specific atom

Low: <3.0A
Average: 1.8-3.0A
Good: 1.0 – 1.8A
Atomic: >1.0A

The resolution varies through the structure!



Resolution

Resolution	Meaning
0.5 - 1.5	In general, structures have almost no errors at this resolution. Structures used for Rotamer libraries and geometry studies.
1.5 - 2.0	Few residues have wrong rotamer. Many small errors can normally be detected. Folds are extremely rarely incorrect.
2.0 - 2.5	Fold likely correct, low proportion of sidechains in wrong rotamer. Many small errors can normally be detected. Fold normally correct and number of errors in surface loops is small.
2.5 - 3.0	Fold likely correct except that some surface loops might be mismodelled. Several long, thin sidechains (lys, glu, gln, etc) and small sidechains (ser, val, thr, etc) likely to have wrong rotamers.
3.0 - 4.0	Fold possibly correct, but errors are very likely. Many sidechains placed with wrong rotamer.
>4.0	Individual coordinates meaningless

How good is a structure?

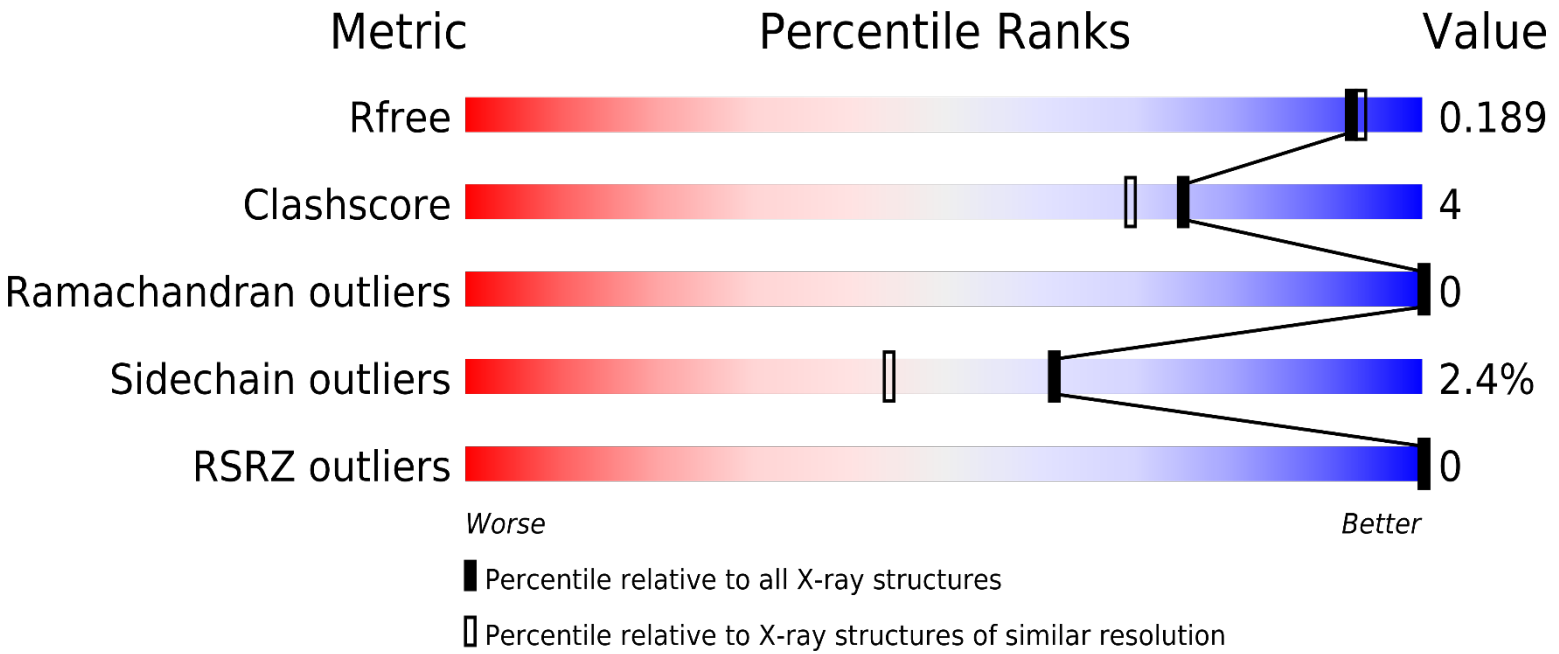
How much is the resolution of the collected data? (In case of X-ray)

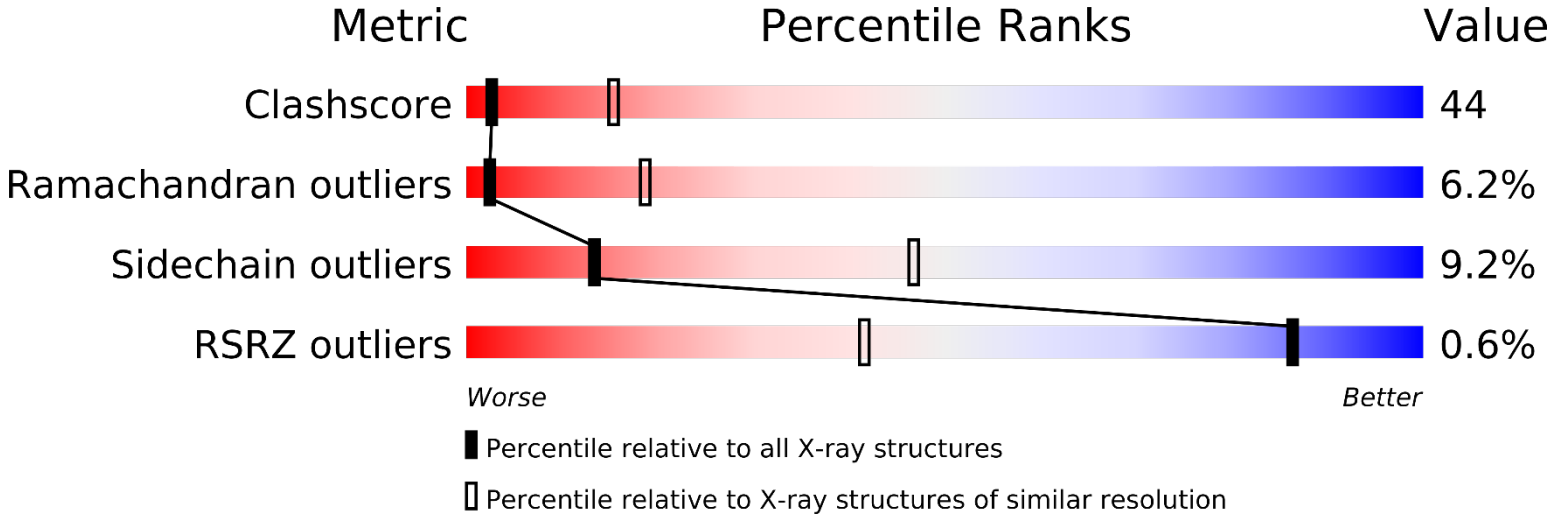
R-factor/FreeR-factor (X-ray)
How well does the data fit to the experimental mesurements?

There is no standard for NMR structures

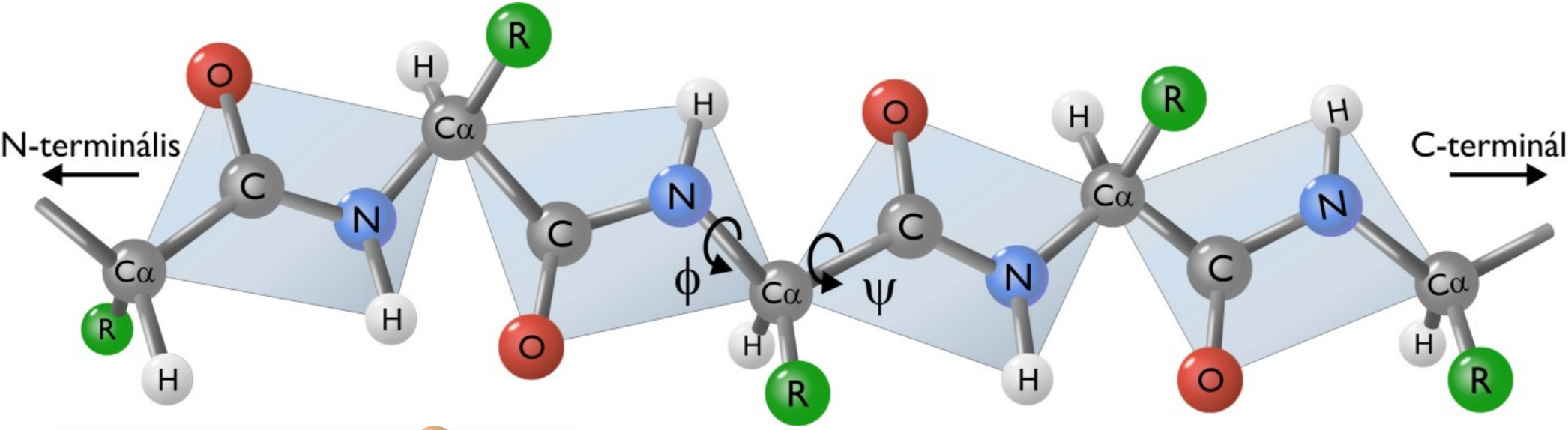
Ramachandran Plot
Geometry and stereo-chemistry

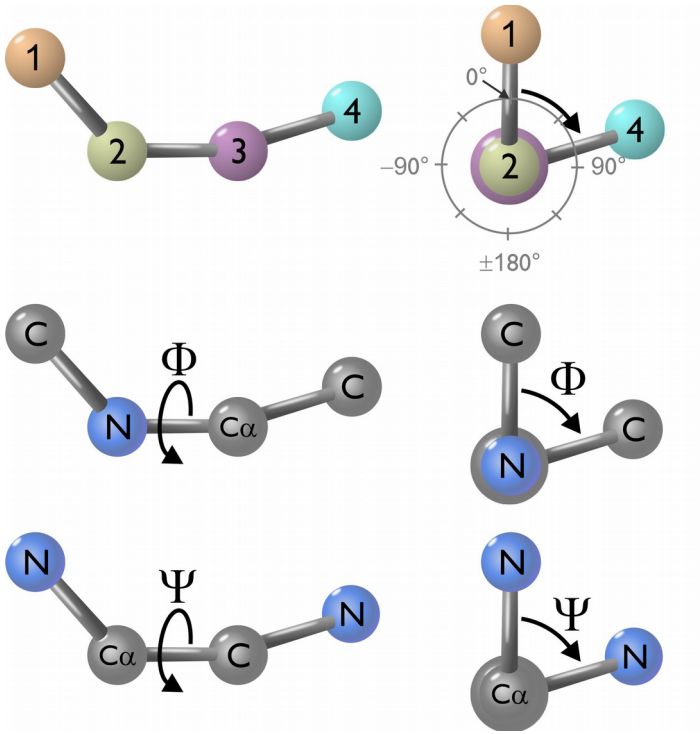
How similar are these to good known structures


Metric
Percentile Ranks
Value
Rfree
0.189
Clashscore
4
Ramachandran outliers
0
Sidechain outliers
2.4%
RSRZ outliers
0
Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution


Metric
Percentile Ranks
Value
Clashscore
44
Ramachandran outliers
6.2%
Sidechain outliers
9.2%
RSRZ outliers
0.6%
Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution

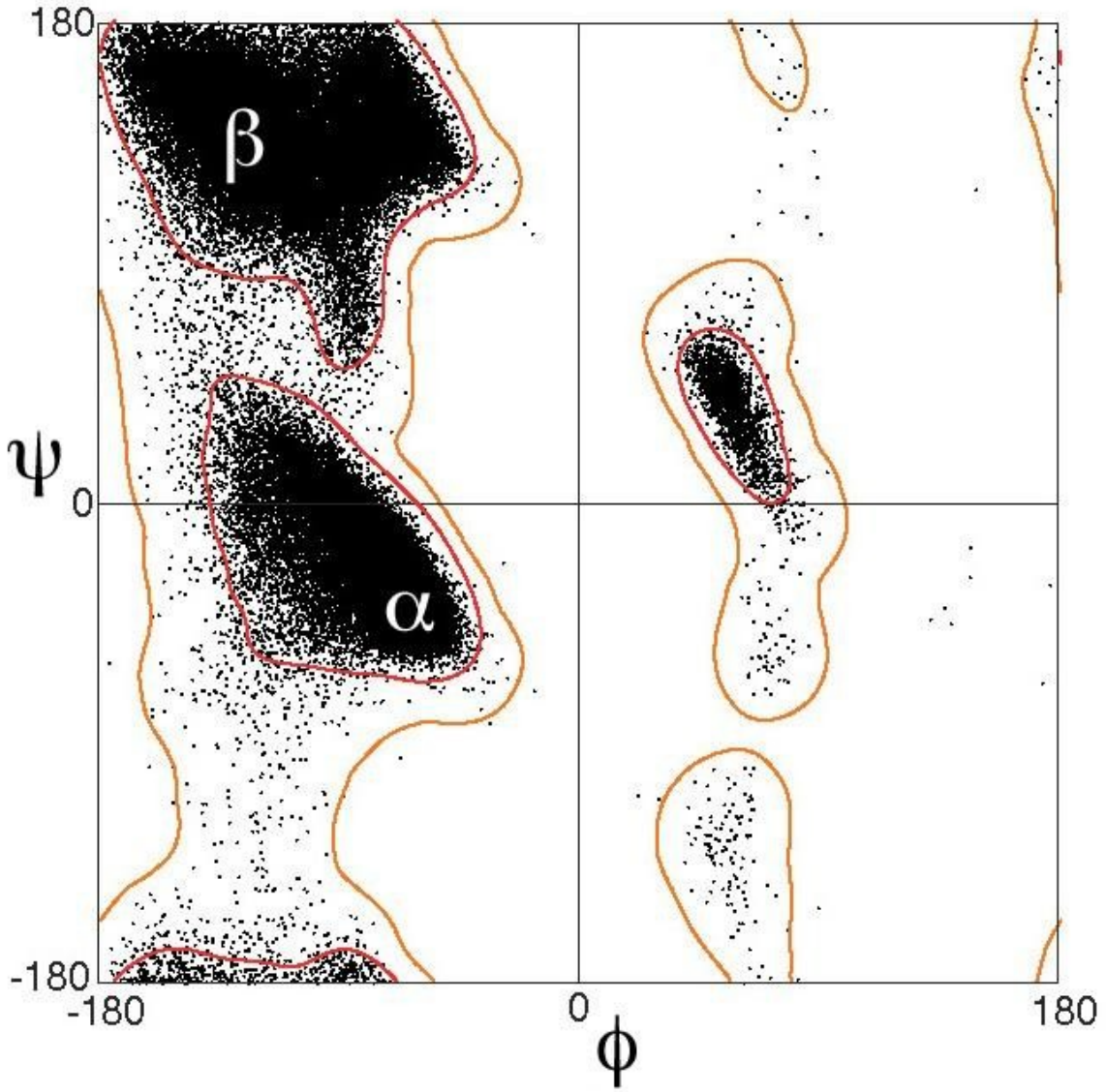
Conformation of the backbone





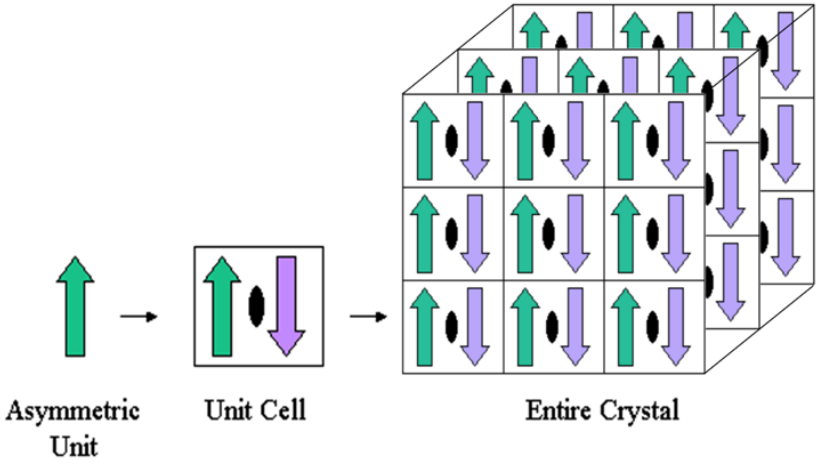
The bonds in the backbone of the peptide chain may only have some specific (φ, ψ) torsion angles. These constrains creates specific preferred conformations.

Ramachandran plot

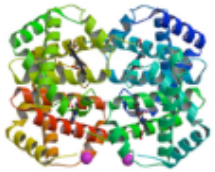
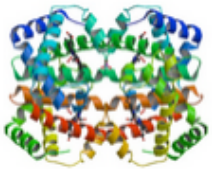
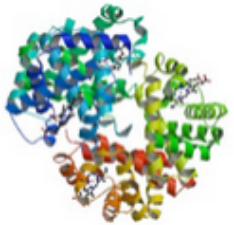
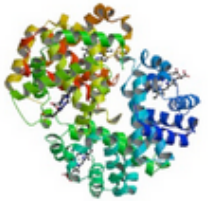


- Using the φ, ψ angles we can evaluate a structure
- Each secondary structure has a distinct region
- Glycines and prolines are not represented, they have special conformational preferences

Asymmetric Unit

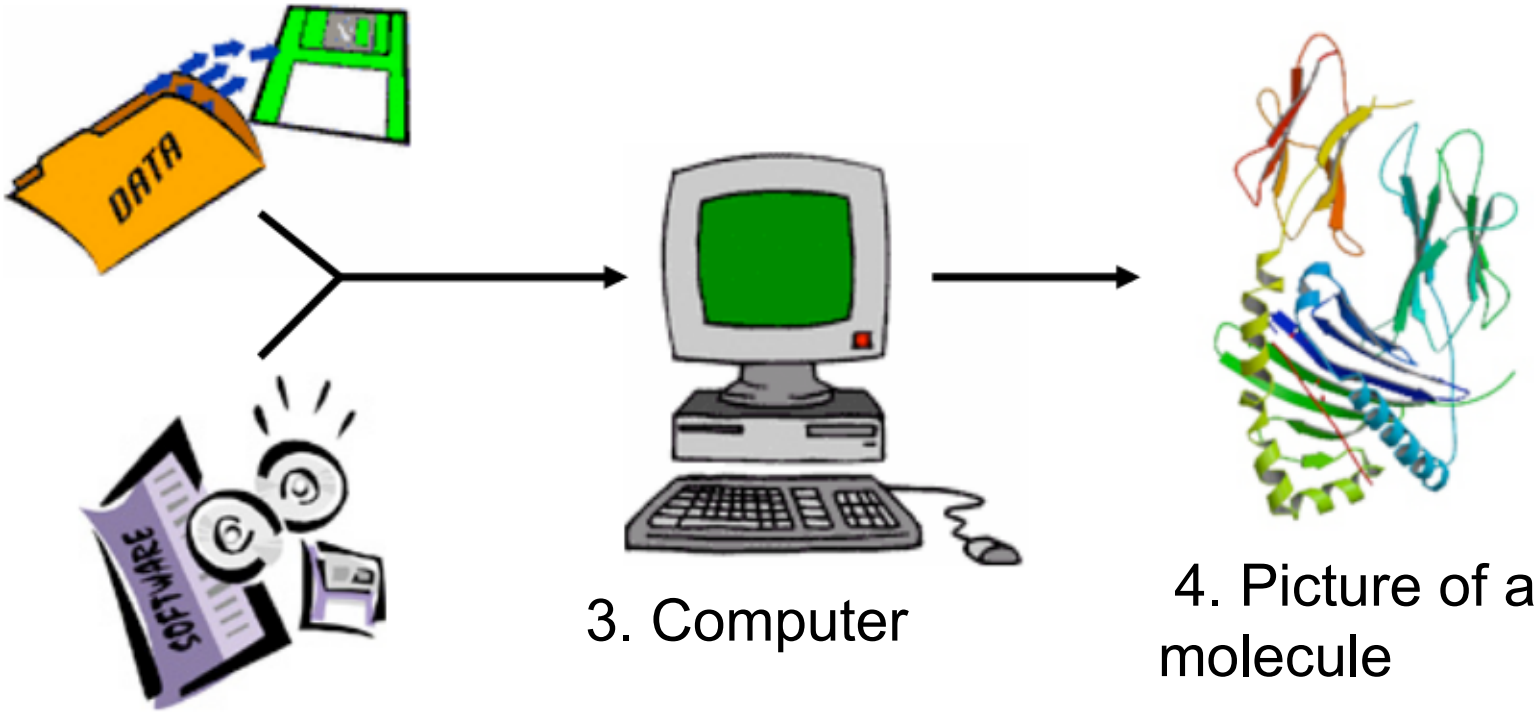


PISA

Biological assembly composed of one copy of the asymmetric unit	Biological assembly composed of multiple copies of the asymmetric unit	Multiple biological assemblies in the asymmetric unit
		
		
In entry **2hhb**, the biological assembly is **equivalent** to the asymmetric unit. No operations are necessary.	In entry **1hho** the biological assembly includes **two** asymmetric units. Application of a crystallographic symmetry operation (a 180 rotation around a crystallographic two-fold axis) produces the complete biological assembly.	In entry **1hv4** the biological assembly is **one-half** of the asymmetric unit. The entry contains **two** structurally similar, but not entirely identical copies of the biological assembly within the crystal asymmetric unit.

Visualization of structures

1. PDB file



3. Computer

4. Picture of a molecule

2. Program for visualization

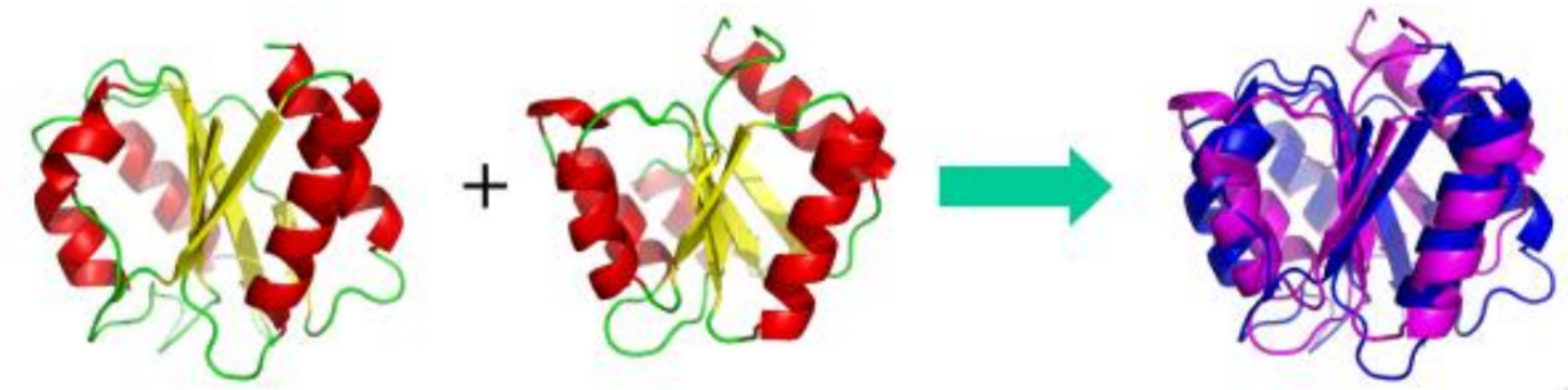
F.E.: Rasmol, Pymol, Chimera,
VMD, Jmol, Swiss PDB viewer

Why Chimera?

- Free licenses
- Well maintained
- Intuitive
- Not just for visualization
- Subjective

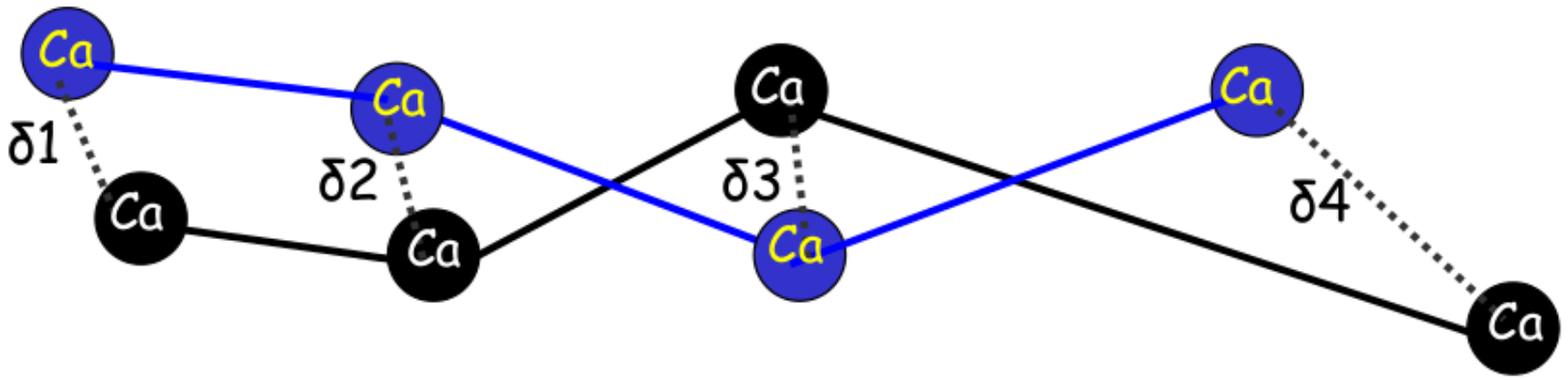
http://www.cgl.ucsf.edu/chimera/

How similar are two structures?



Superposition: Minimizing the distance of positions

RMSD



Root Mean Square Deviation (RMSD):

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

The most common function used to measure structural similarity

RMSD is the average distance between equivalent atoms of superimposed proteins (generally Ca).

Structural alignment

What are the equivalent positions?

- If the sequences are alike we can use that

- In case of low sequence identity we can use the structure

RMSD depends on the equivalent positions

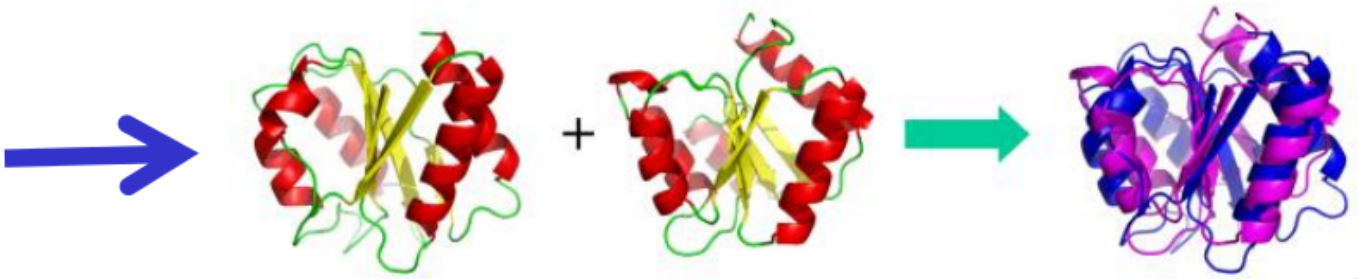
Linker regions may cause problems

MatchMaker: Structure Alignment (Chimera)

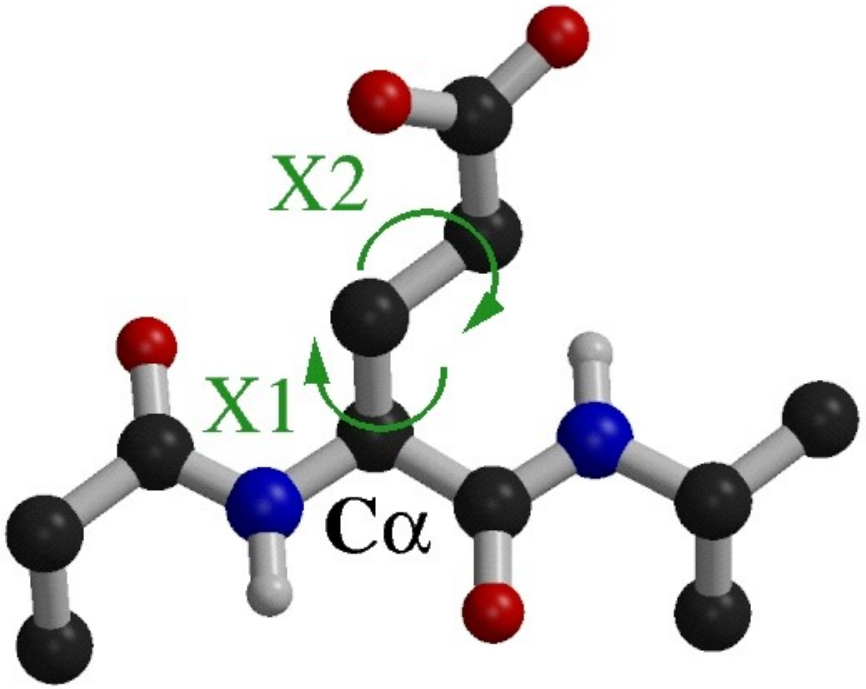
1. Constructs a pairwise sequence alignments as a guide to superimpose the structures.
2. Improves the sequence alignment using a scoring matrix that can include residue similarity, secondary structure and gap penalties.
3. Structural Fitting using one point per residue: *CA* atoms in amino acids and *C4'* atoms in nucleic acids.

```
3chy_: KBCCDDDDFBF--KLMMMMMMMMMM
2fox_: ---DDDDFKOMMMMMMMMMMMMMMM

3chy_: LM--LMMMMMMMPMKLMMPCCDDDI
2fox_: LPACFKLMMMMMMMMMMMGHIACDI
```



Rotamers



Rotamers are usually defined as low energy side-chain conformations

Rotamer libraries