What are the main methods used to generate the PDB data?

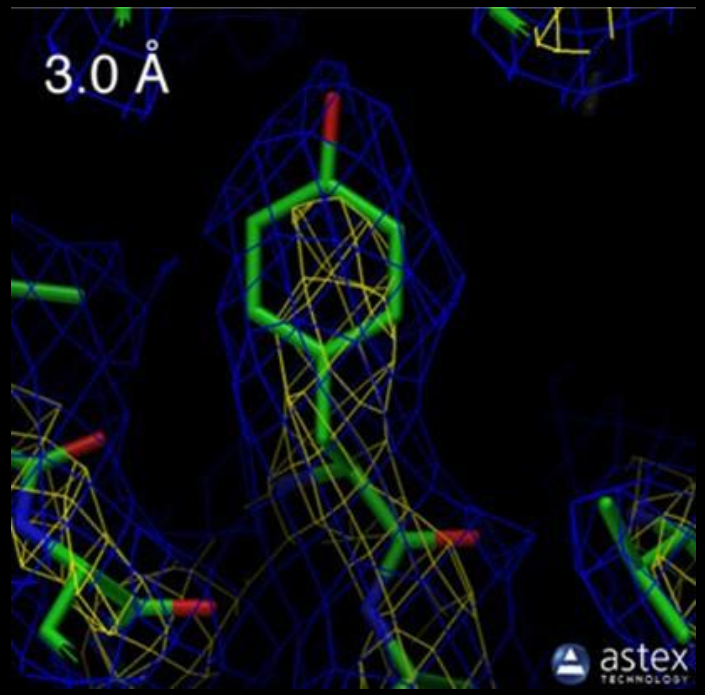
X-Ray
Cryo-EM
NMR

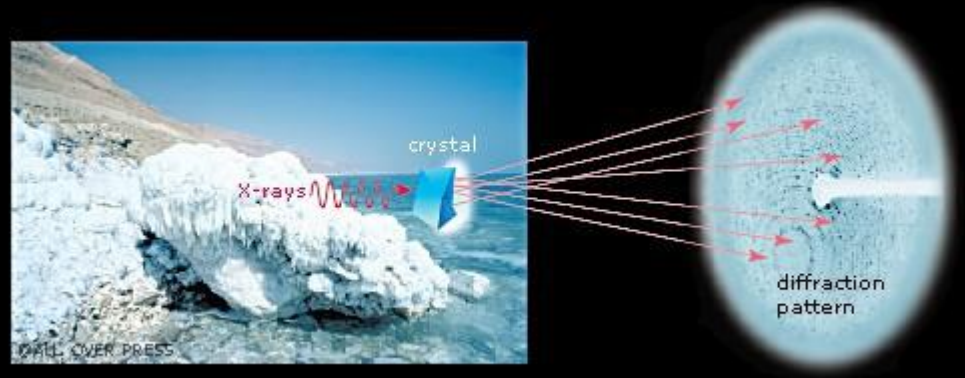
X-ray Protein Structure Determination

~ 90 % are X-ray crystallography determined structures

Crystals → X-ray diffraction →

Electron density maps (EDMs)





The structure is solved by fitting the amino acids into the electron density map.

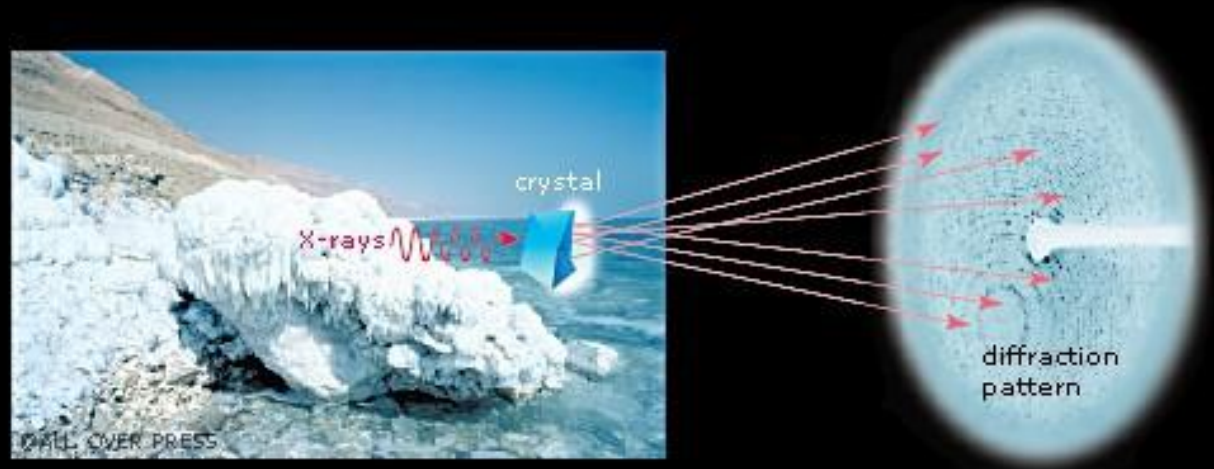
X-ray Protein Structure Determination

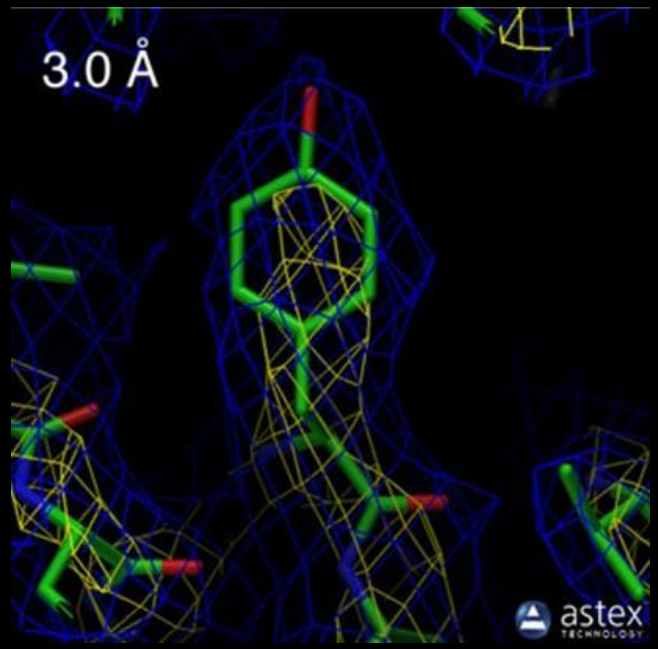
~ 90 % are X-ray crystallography determined structures

Crystals → X-ray diffraction →

Electron density maps (EDMs)

The structure is solved by interpretation of electron density maps and fitting the primary amino acid sequence into the electron density map

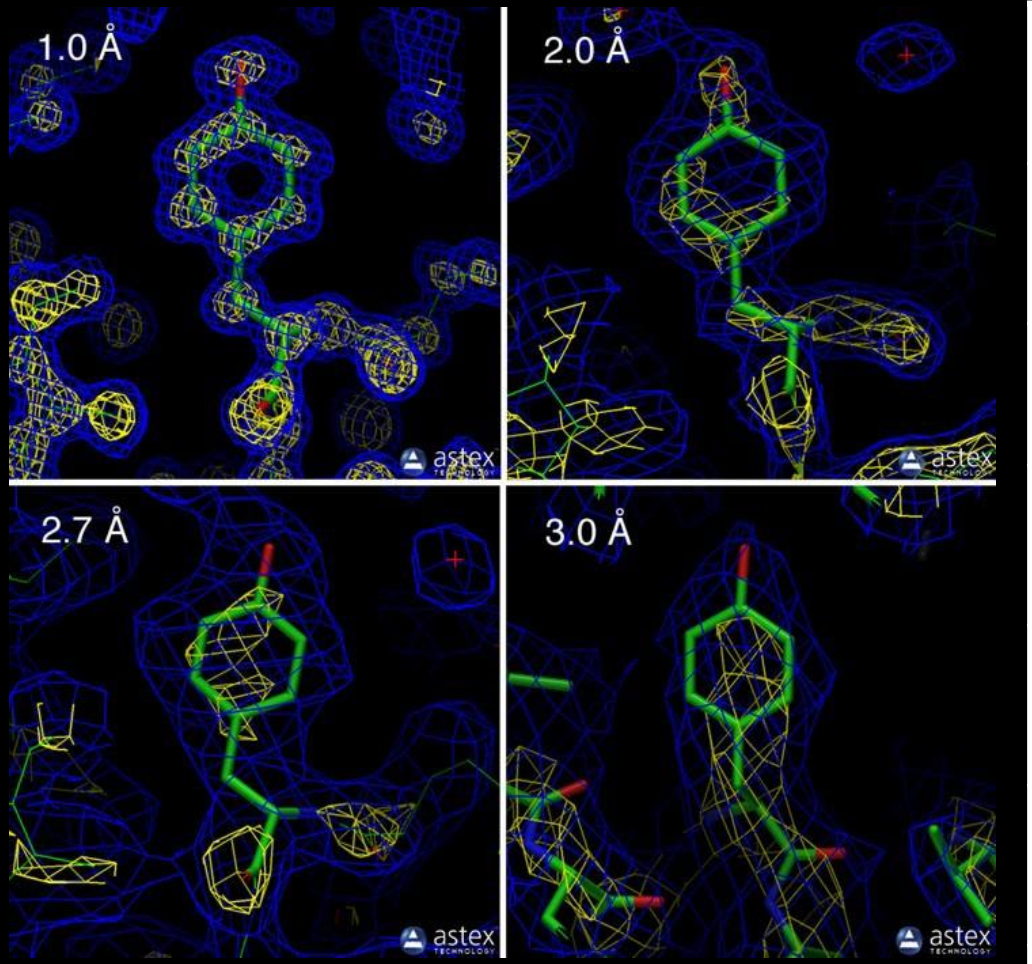




Reliability factor (R)
R-value (work) vs. R-free – parameter fitting successful?
About 0.2-0.3 (R-work and R-free should be similar: 0 is perfect and 0.6 is random)

X-ray Resolution

1 Ångström (Å)= 0.1 nm



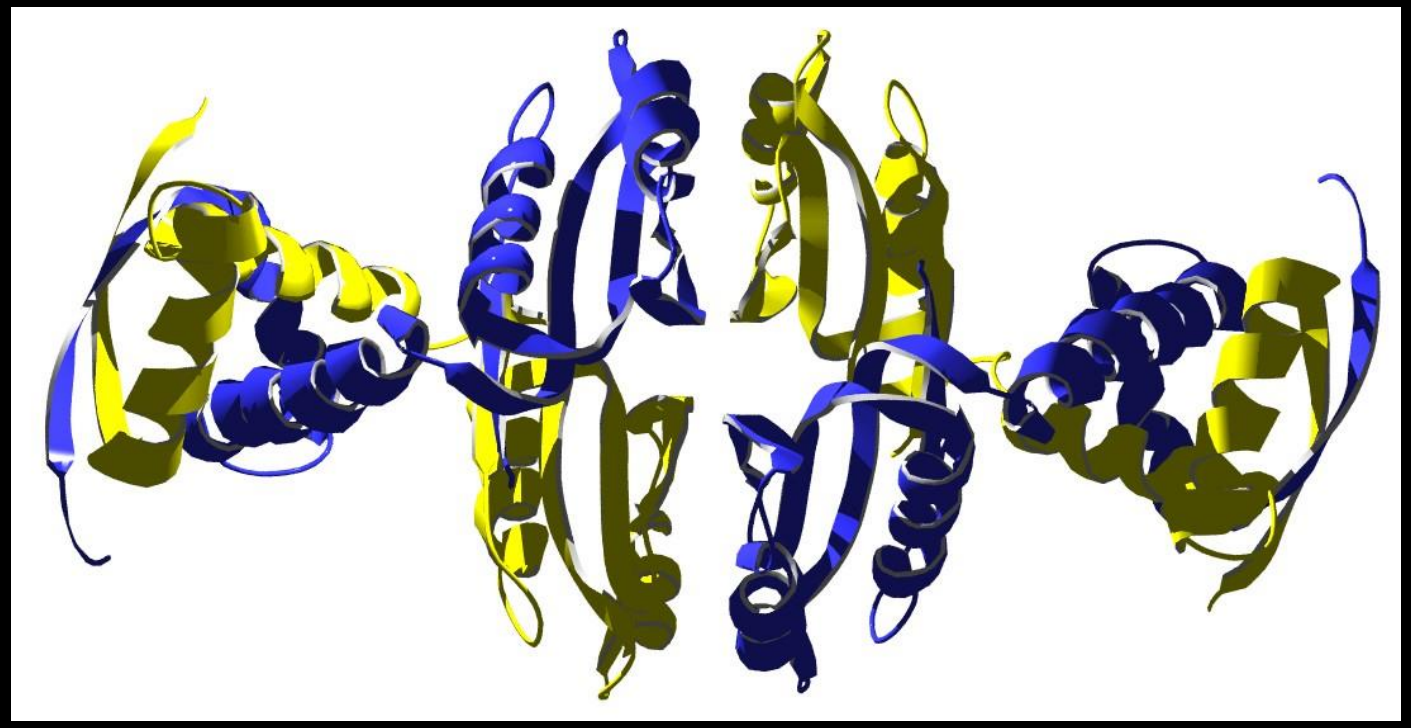
www.pdb.org

Resolution	EDM shows
5.0 Å	Broad helices
3.0 Å	Side chains
2.5 Å	Water molecules
2.0 Å	Reliable distances
1.5 Å	Individual atoms
1.0 Å	Hydrogens

X-ray *crystallography*

Asymmetric symmetry - may not correspond to the functional (biological) unit.

Static snapshot of a dynamic protein structure…



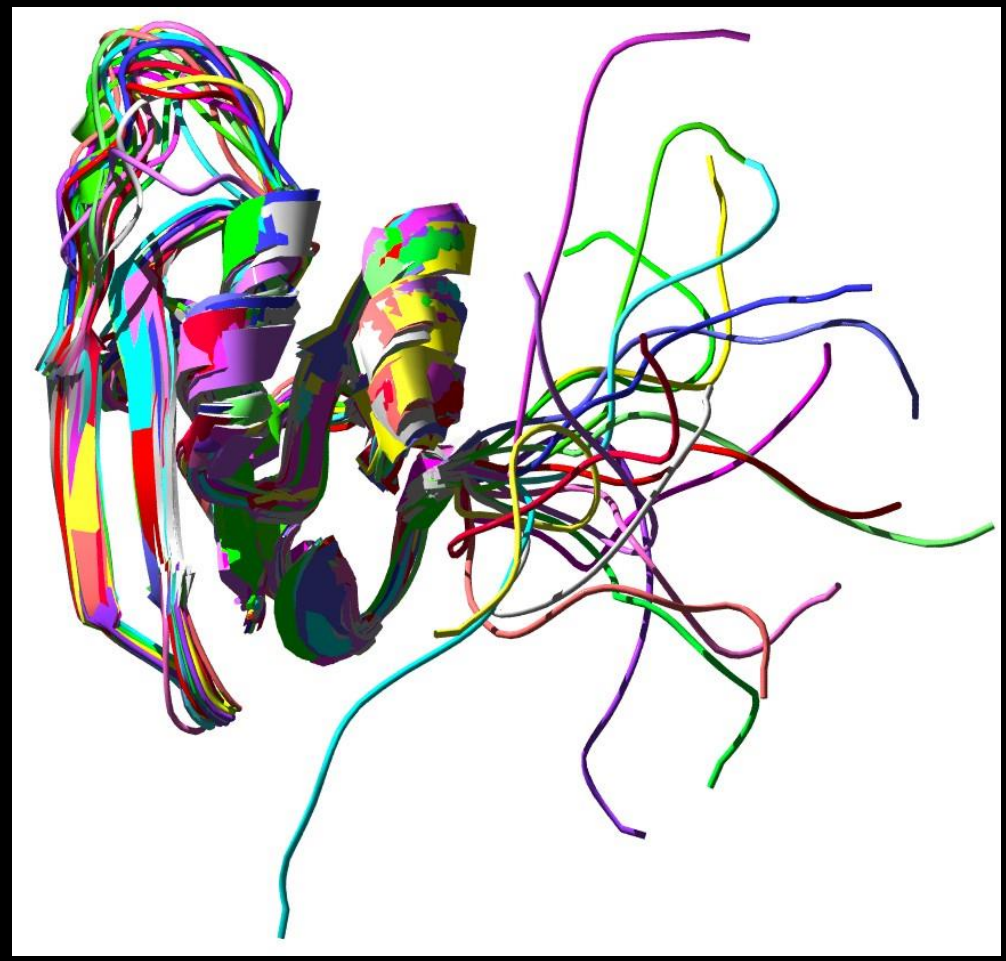
NMR - Nuclear magnetic resonance

Works best for
smaller proteins

Shows dynamics.

Yields an ensemble of structures as the protein is dynamic in solvent.

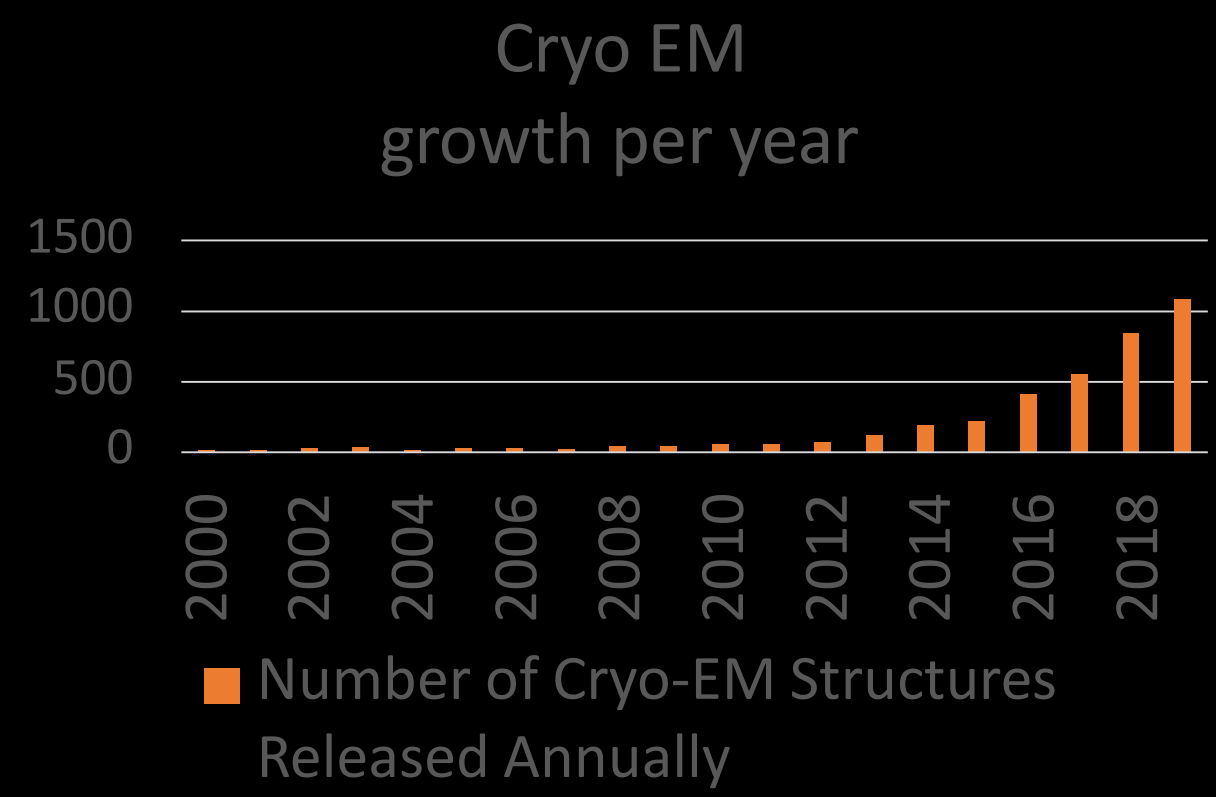
Large PDB files with the best
(lowest energy) conformations.



Cryo-Electron microscopy

- Often used to determine viral structures or bigger protein complexes

- FREEZE! Don’t move. Proteins are present as conformational ensembles and combining the images from lots of freeze samples show the conformational ensemble.

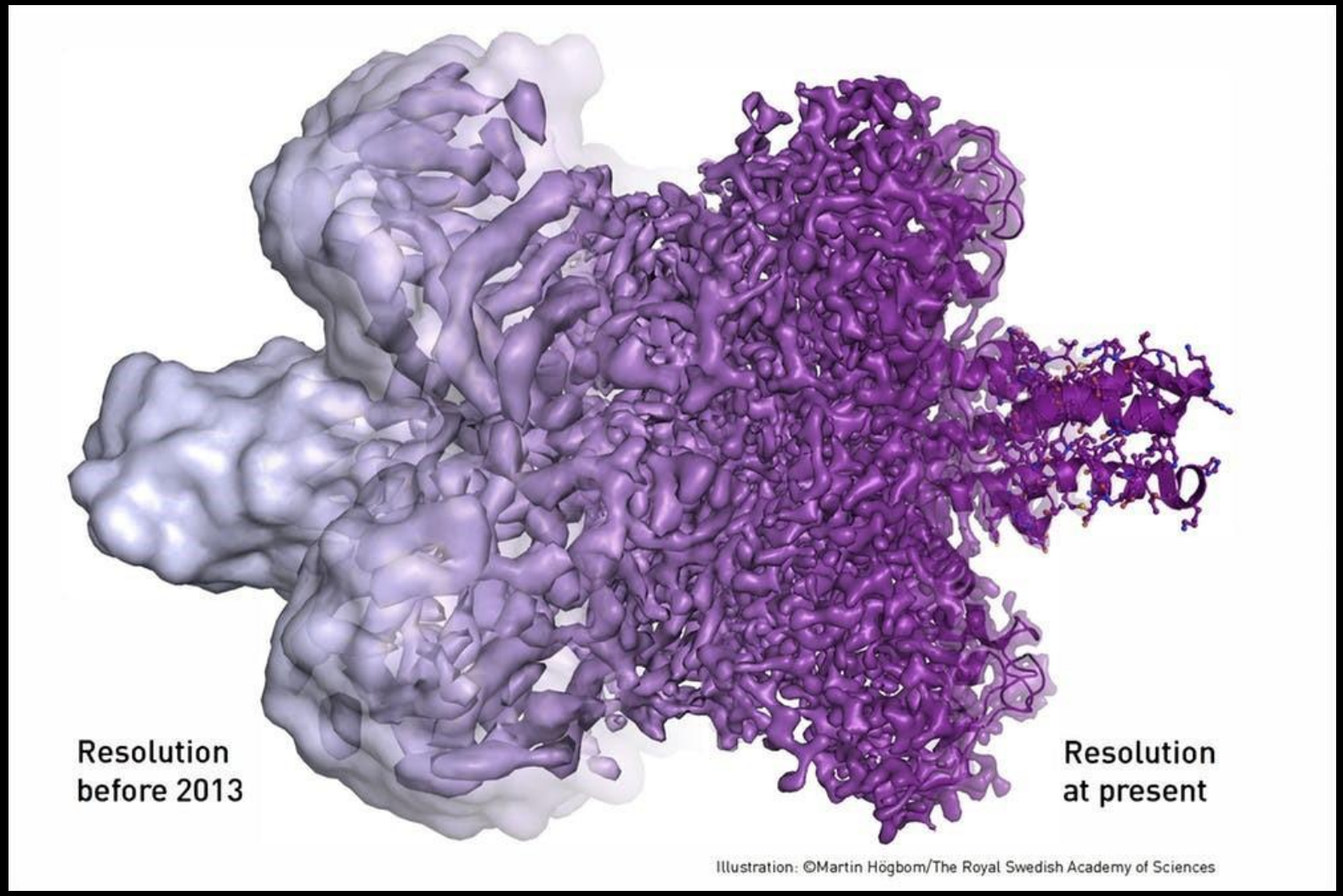


www.pdb.org

Cryo-EM

Low resolution? It is improving. Rapidly.

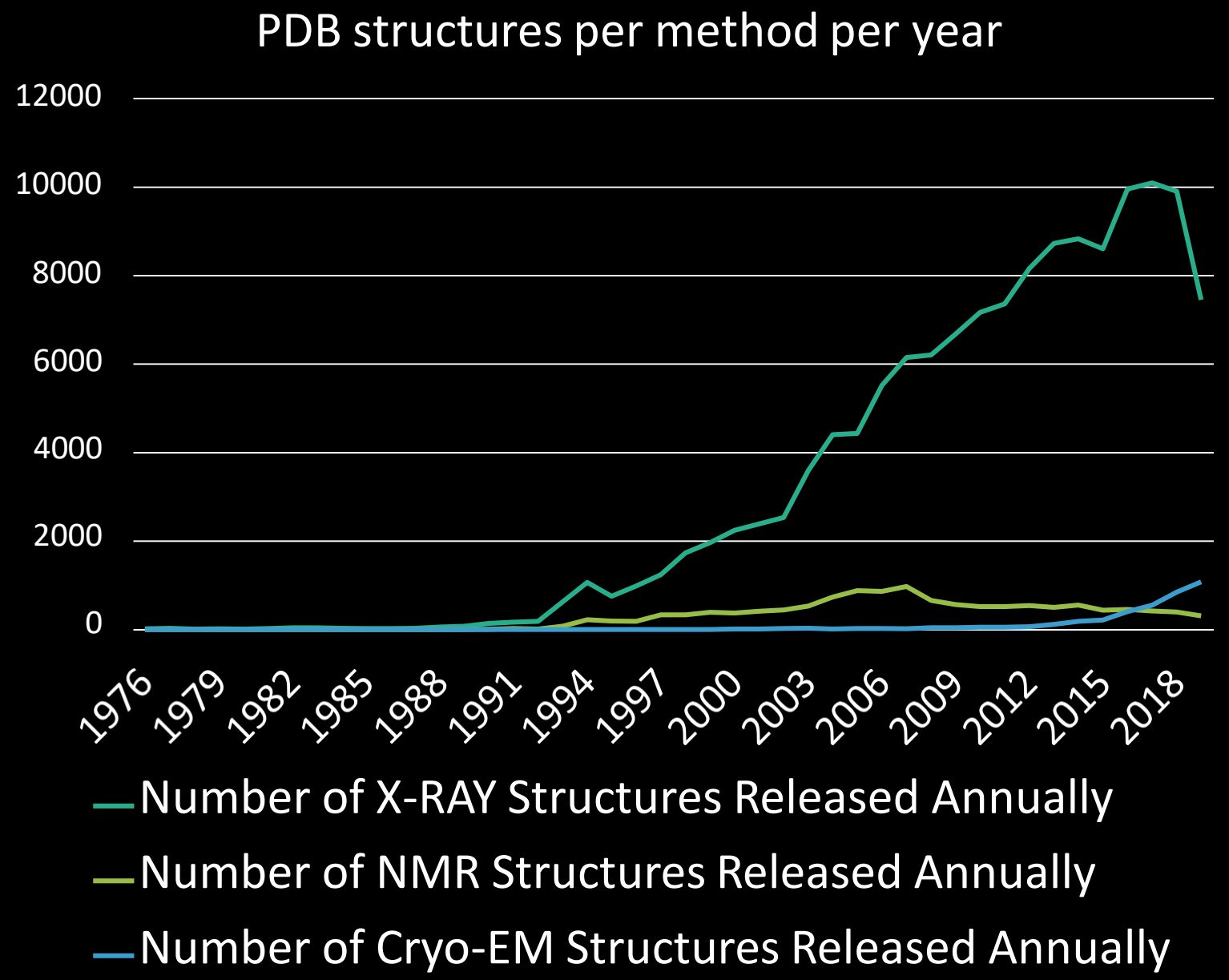
Nobelprize in Chemistry 2017



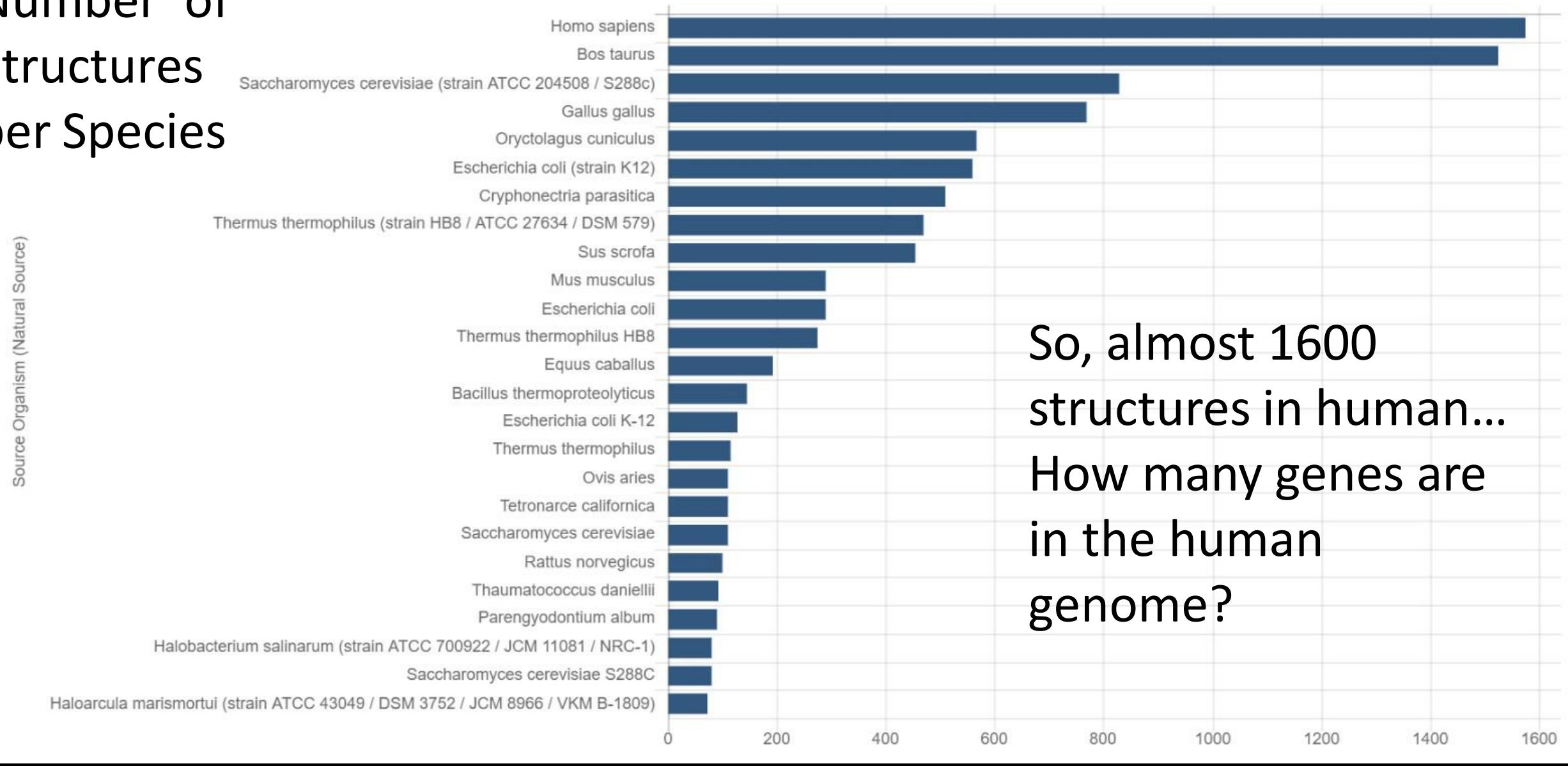
RCSB PDB PROTEIN DATA BANK

156365 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Method	Total	Trend
X-ray	89.3%	↓
NMR	8.2%	↓
EM	2.4%	↑



Number of Structures per Species



So, almost 1600 structures in human… How many genes are in the human genome?

Structure vs Sequence

We have a lot more sequences than structures, but since structures seem to change more slowly, we can use the structures that we do have to predict structure for more of the sequence data.

Do you think we can predict structures for all known human sequences?

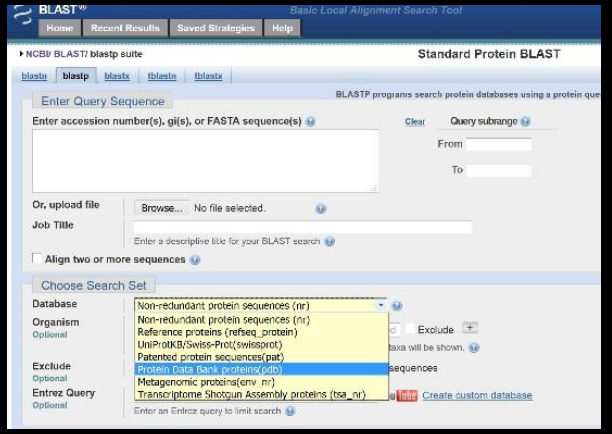
We cannot (yet) predict 3D structures for all proteins, but we can predict secondary structures!

Homology modeling – prediction of 3D structure

Homology modeling

- Protein structure is generally more conserved than sequence.
- We can use the known protein structures to model related protein sequences.
- How do we find protein structures of proteins that are related to our protein of interest?

- BLAST against PDB: www.pdb.org or



Homology modelling

- The starting sequence (the query used for BLAST) is the **target** to be modelled.
- The BLAST hits are potential **templates** for the model.
- The model will only be for the part of the target sequence that overlaps with the template.

- **No template? No model.**

  Homology modeling only works if there is a template to base the model on.

Templates and Alignment options

Finding a template:

- Refined template finding algorithm in Swiss model that tries to find the better template.
- If you happen to know which part or which conformation you need to model, identify the template or the template characteristics on your own.
- GMQE (Global quality estimation score) – higher number better model (better, or longer cover)

Alignments:

- Pairwise sequence alignment if very similar sequences
- A multiple sequence alignment (MSA) can improve the MSA for more divergent sequences. Which sequences should be included in the MSA?

Workflow of homology modeling using SWISS-MODEL workspace.

Automated mode

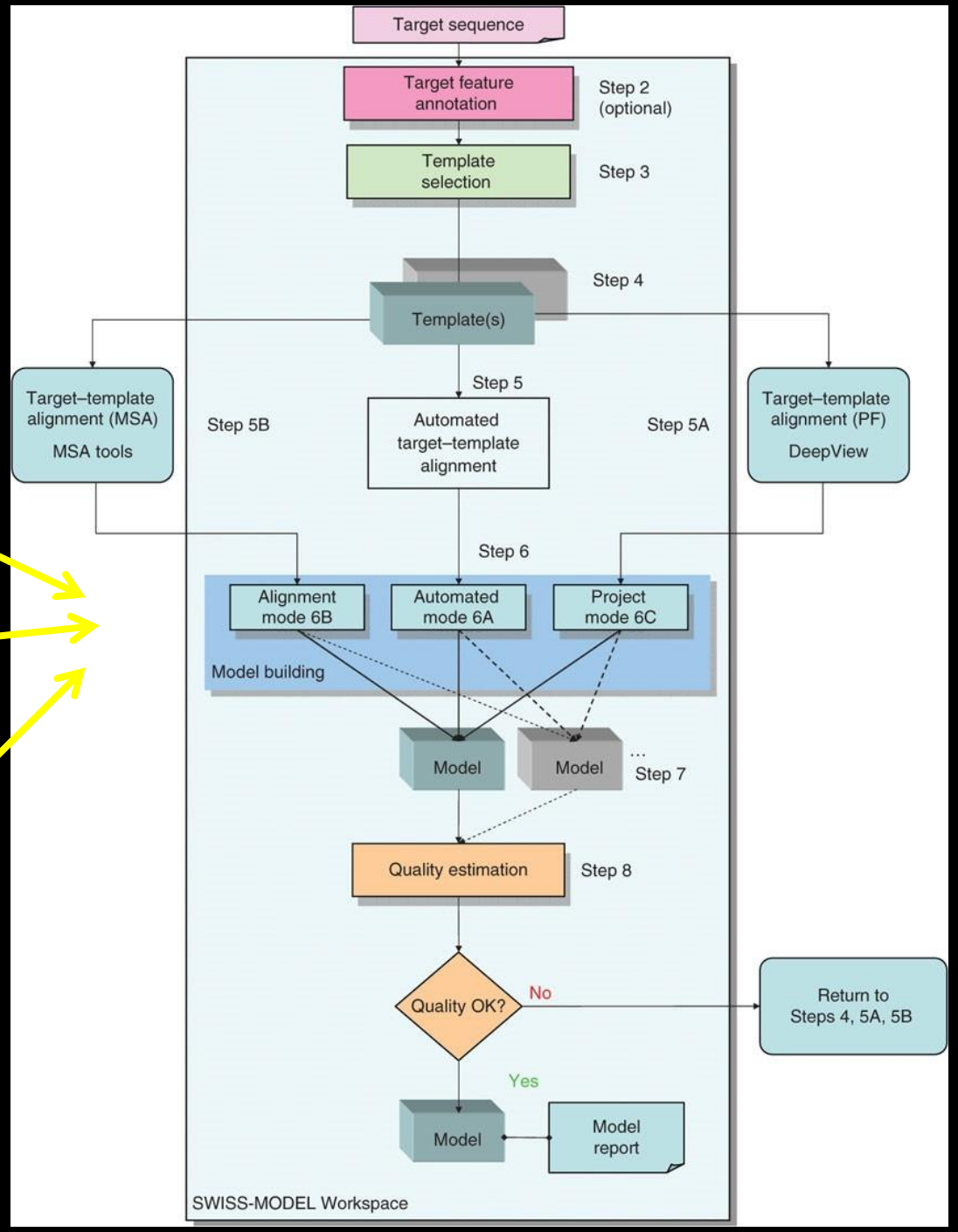
Provide a target sequence, Template optional

Alignment mode

Provide an alignment including target and template

Project mode

Build a project file in SwissPDBviewer. Oligomers, multiple sequence alignments



[Protein structure homology modeling using SWISS-MODEL workspace]

Bordoli et al. Nature Protocols 4, 1 - 13 (2008)

Modeling

- Framework
  - Based on the topological arrangement of corresponding atoms
  - Side chains with fully incorrect geometries are removed
- Loops are created from a loop fragment database
- Sidechains are added from a rotamer library. The most common conformations for the sidechains are called rotamers.
- Verify structure and packing
- Model refinement by energy minimization (according to a force field*)
- Summarized in the modeling log

A good model depends upon

1. A good template
2. A good alignment between the template and the target

What is a Statistical Potential?

This is also called a knowledge-based potential.

It compares how similar a model is to what is known for experimental protein structures.

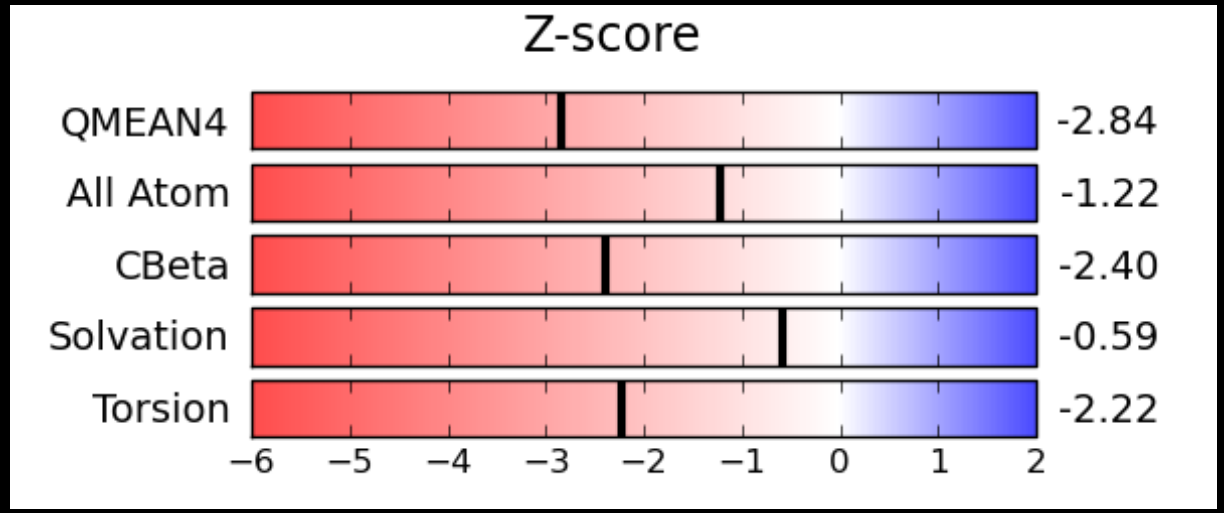
Homology model evaluation

QMEAN – statistical potential based on

1. Torsion: torsion angle potential over three consecutive amino acids

2 & 3. All atom and C-beta: two pairwise distance-dependent potentials are used to assess all-atom and C-beta interactions

4. Solvation: A solvation potential describes the burial status of the residues



The more similar the distribution for the model is to the observed distribution in real experimental structures, the higher score. The higher score, the better model.

Target model = Model_01 Template = 2phm.1.A

Are there better (blue) or worse (red) regions in the model?

Identify alignment errors.

Maybe the alignment can be improved if you make a multiple sequence alignment instead?