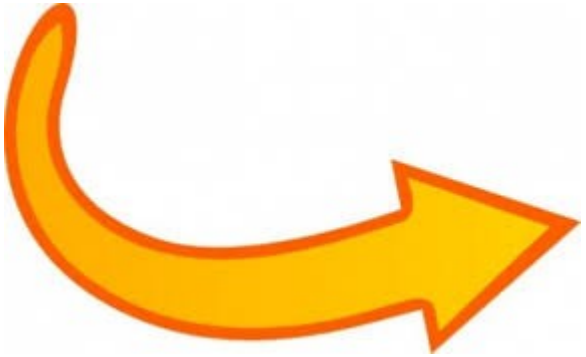
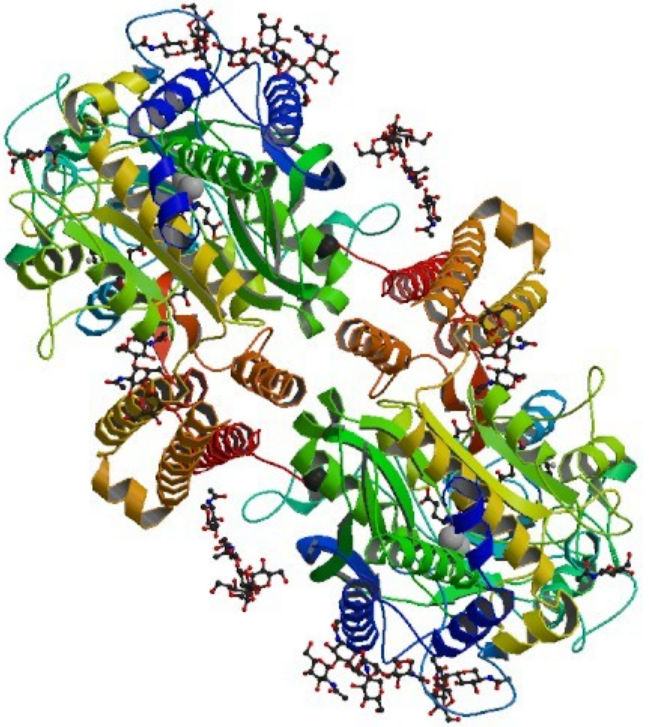
8.
3D Predictions

Structure prediction

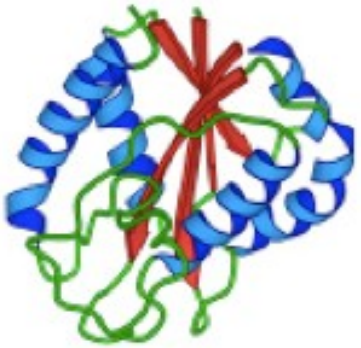
```
>Protein
RSKSSNEATNITPKHNMKAFLDELKAENIKKFLYNFTQIPHLAGTEQNFQLAKQIQSQWKEFGLDSVELAHYDVLLSYPN
KTHPNYISIINEDGNEIFNTSLFEPPPPGYENVSDIVPPFSAFSPQGMPEGDLVYVNYARTEDFFKLERDMKINCSGKIV
IARYGKVFRGNKVKNAQLAGAKGVILYSDPADYFAPGVKSYPDGWNLPGGGVQRGNILNLNGAGDPLTPGYPANEYAYRR
GIAEAVGLPSIPVHPIGYYDAQKLLEKMGGSAPPDSSWRGSLKVPYNVGPGFTGNFSTQKVKMHIHSTNEVTRIYNVIGT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





Protein folding

GFCHIKAYTRLIMVG…



Folding

(physics)

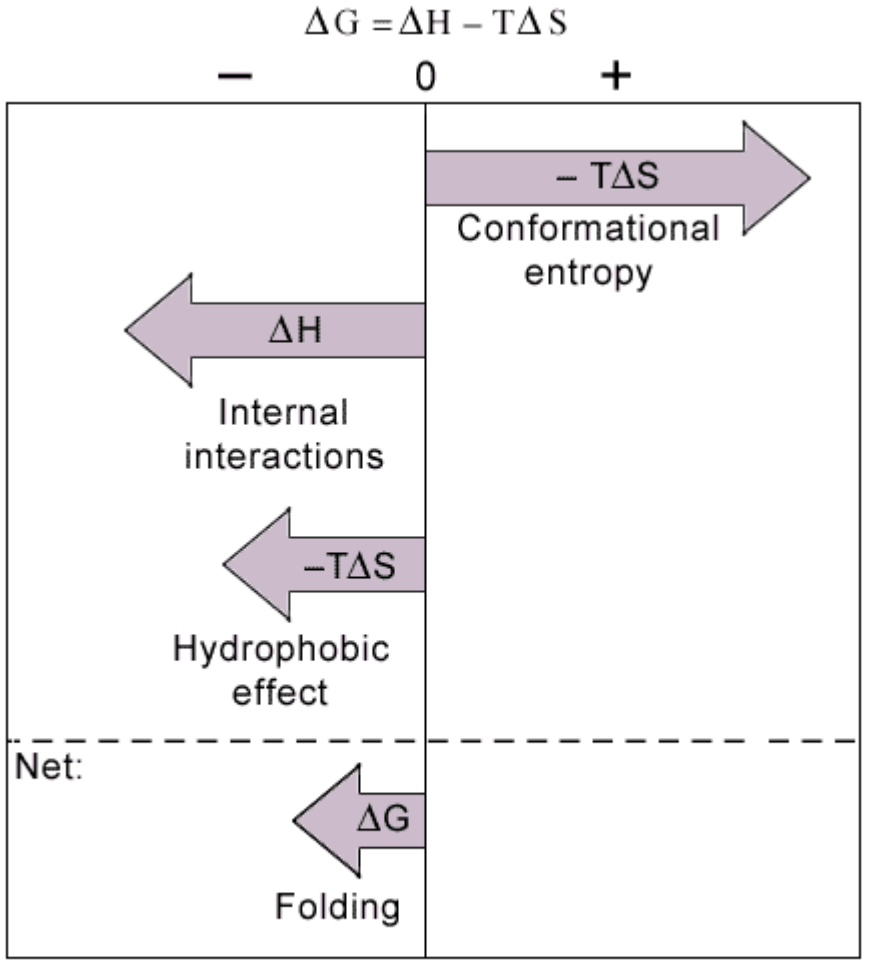
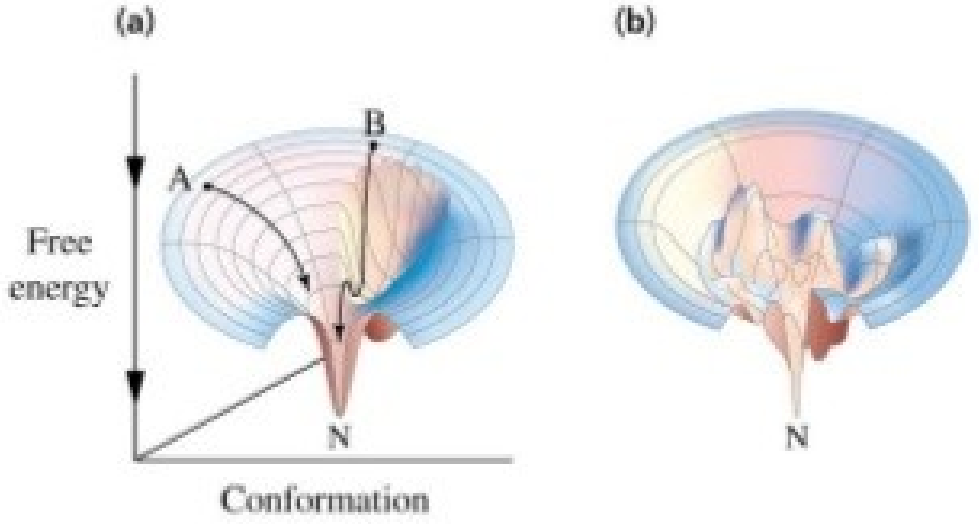


Figure 6.22, C.K. Mathews & K.E. van Holde, *Biochemistry*, 2nd edition (1996)

Protein Folding

- Structures of globular proteins are not static
- Proteins "breathing" between different conformations
- Proteins fold towards lowest energy conformation
- Multiple paths to lowest energy form
- All folding paths funnel towards lowest energy form
- Local low energy minimum can slow progress towards lowest energy form



Structure prediction based on physics

Molecular dynamics simulation

- Large number of conformations, conformational space is huge

- correct physical energy function is not known

Advances with ANTON

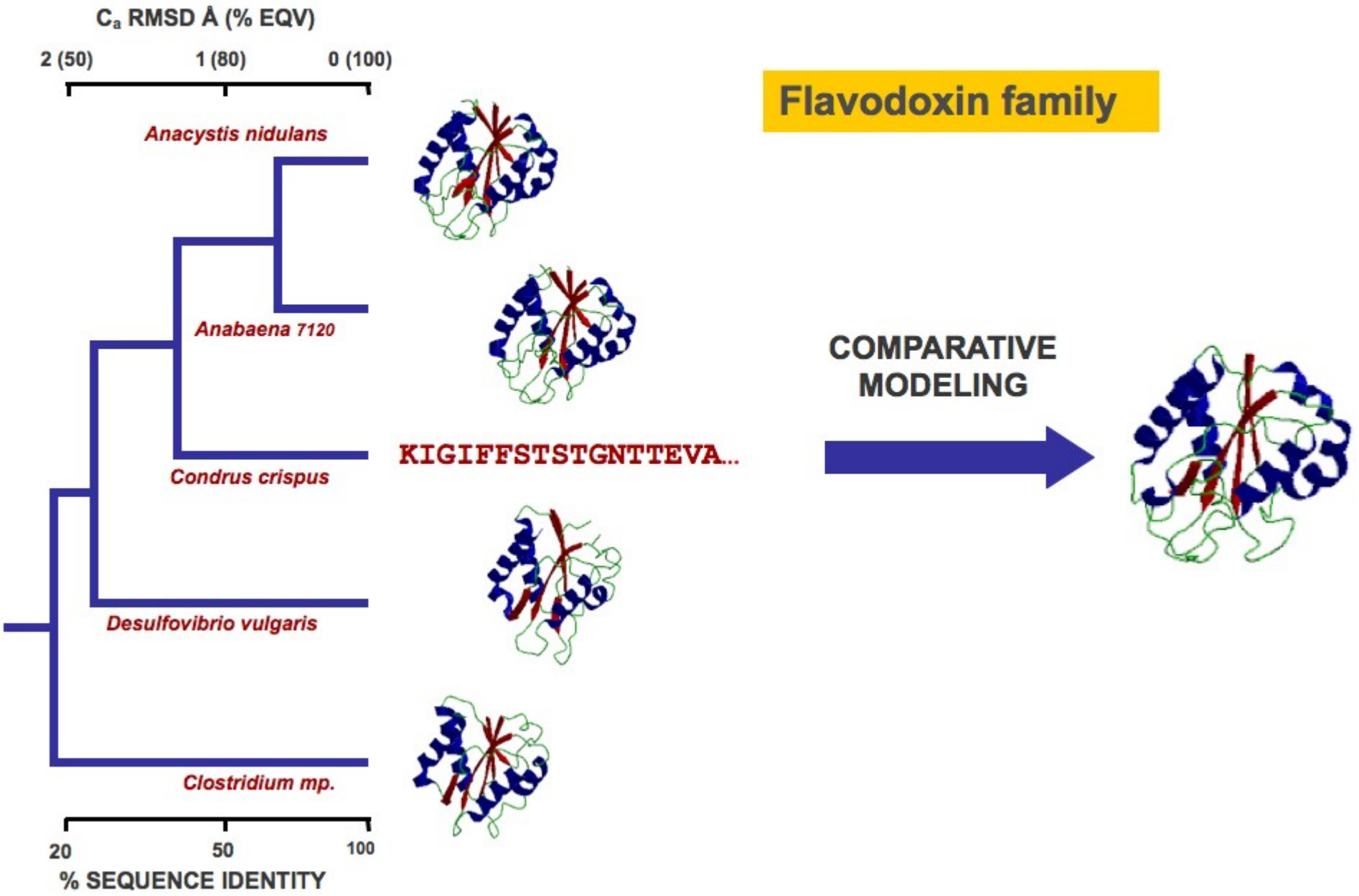
Structure Prediction

Homology modelling

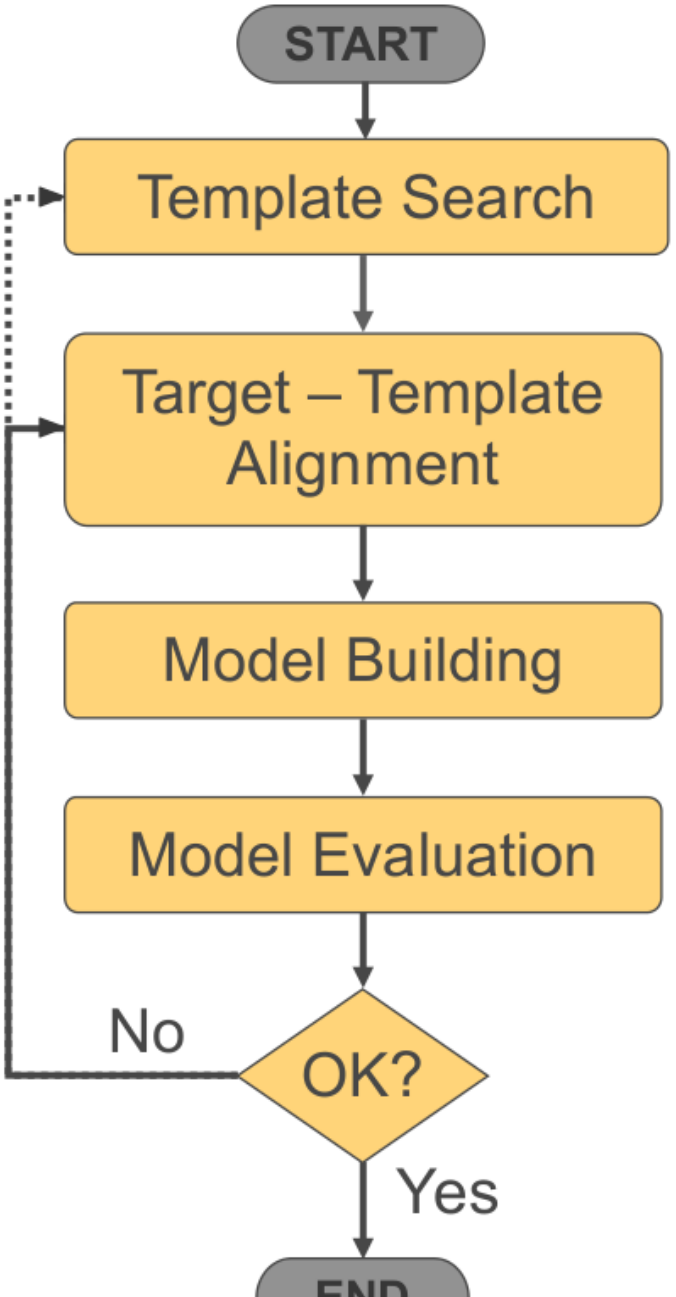
(Threading)

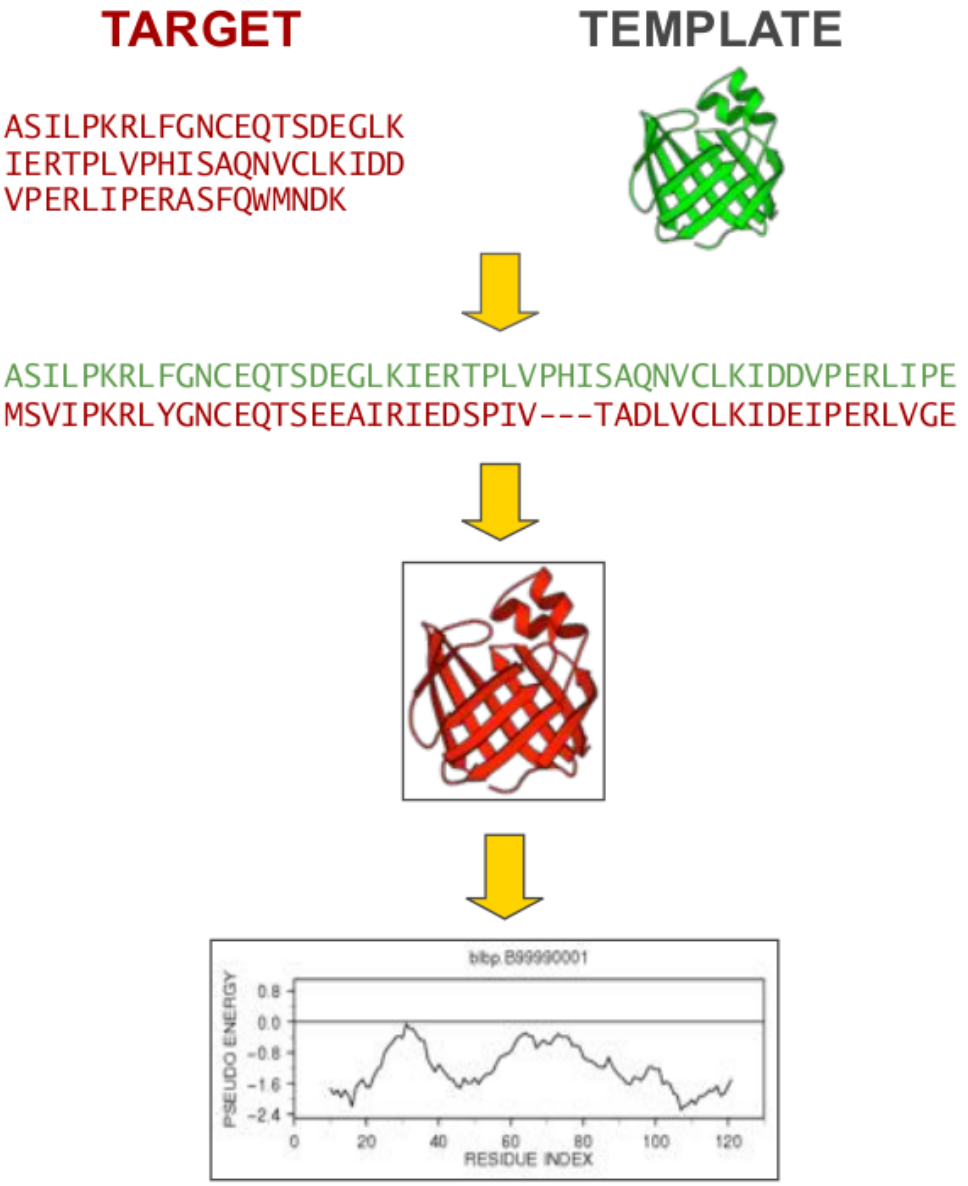
Ab-initio structure modelling

Comparative Protein Structure Modeling



Steps of homology modelling

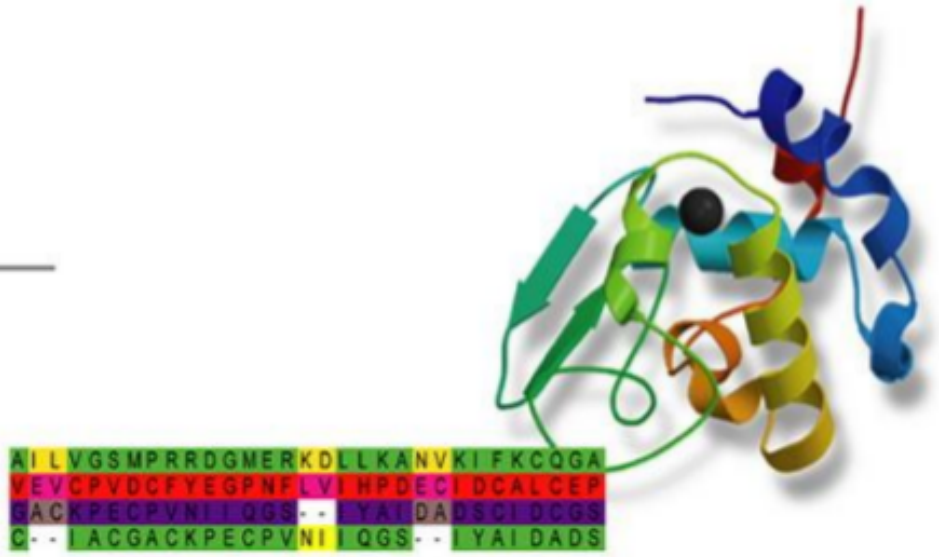




Modelling packages

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



SIB Swiss Institute of Bioinformatics

BIOZENTRUM Universität Basel

SWISS-MODEL

Modelling

myWorkspace

Automated Mode

Alignment Mode

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide.

SWISS-MODEL Team

Torsten Schwede:	Project Leader
Florian Kiefer:	SWISS-MODEL Repository
Lorenza Bordoli:	Method Development and user support
Konstantin Arnold:	SWISS-MODEL Workspace

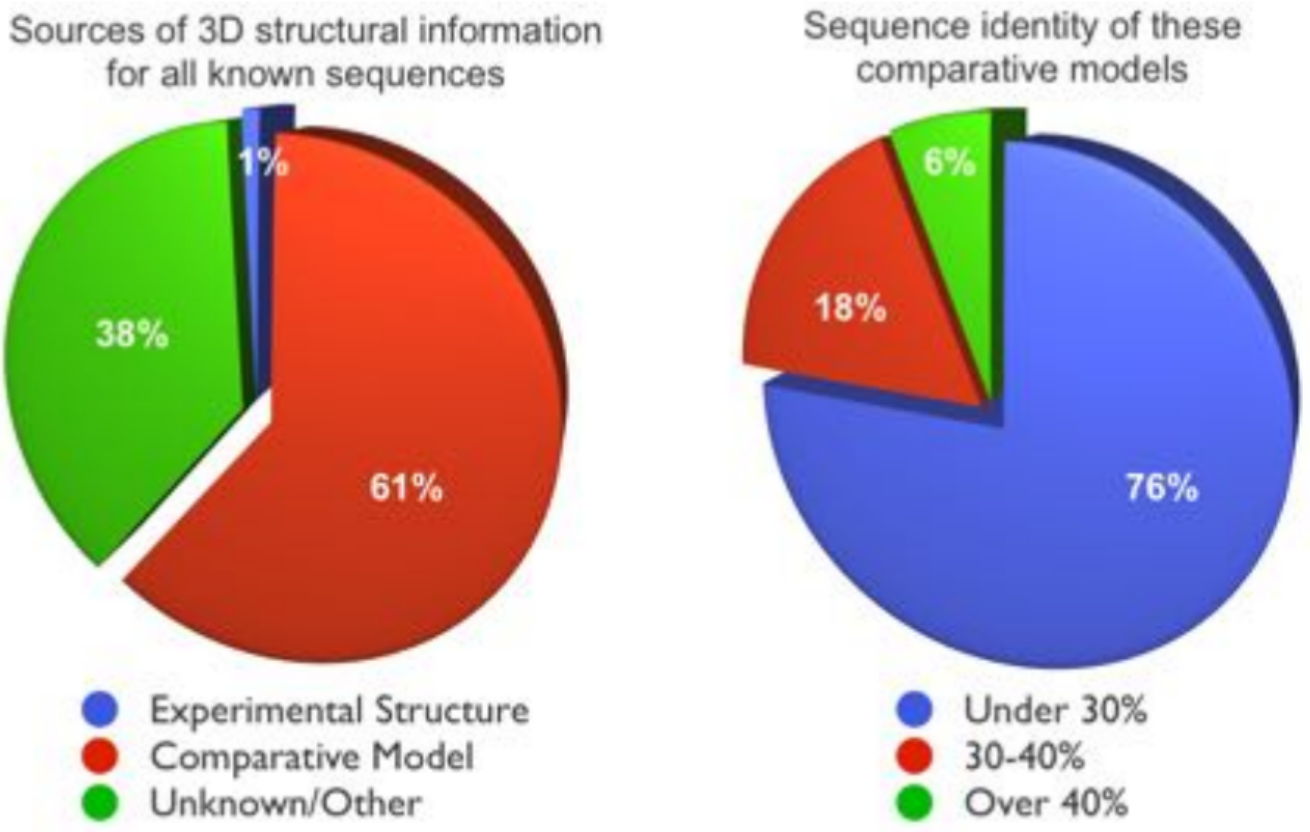
Comparative modeling of the UniProt database

Unique sequences processed: 2,130,404

Sequences with fold assignments or models: 1,273,766 (60%)

70% of models based on <30% sequence identity to template.

On average, only a domain per protein is modeled (an “average” protein has 2.5 domains of 175 aa).



Pieper *et al. Nucleic Acids Research* **34**, D291, 2006.

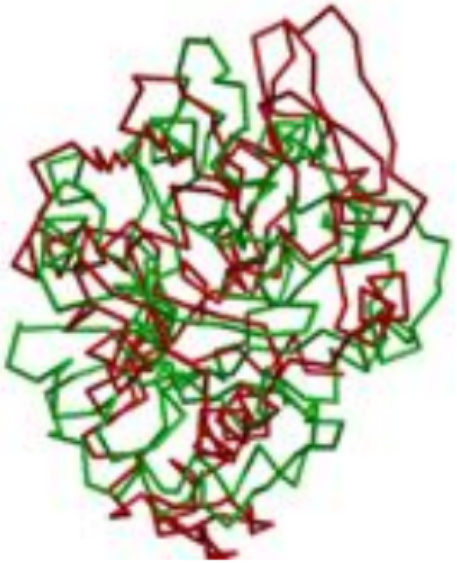
Main sources of errors

MODEL

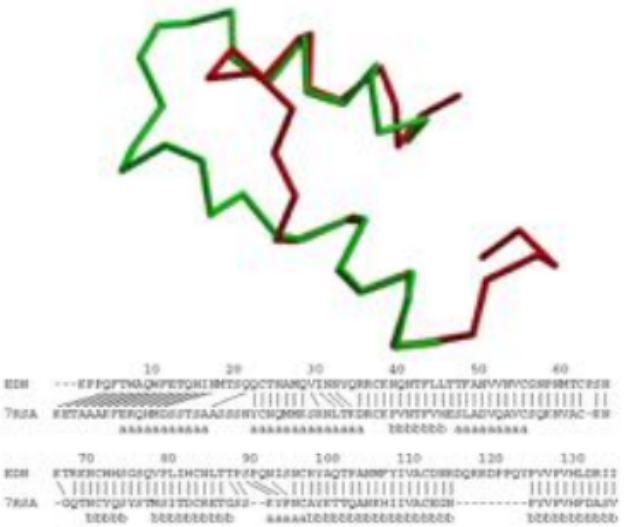
X-RAY

TEMPLATE

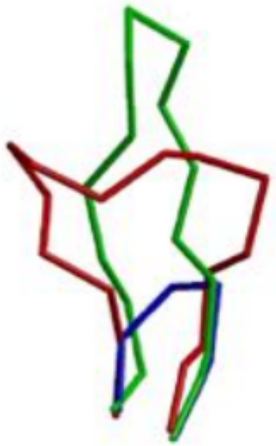
Incorrect template



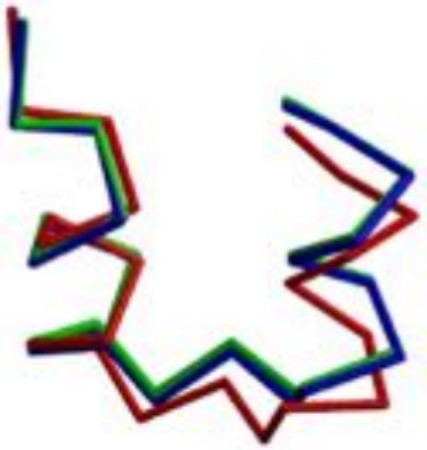
Misalignment



Region without a template



Distortion/shifts in aligned regions



Sidechain packing

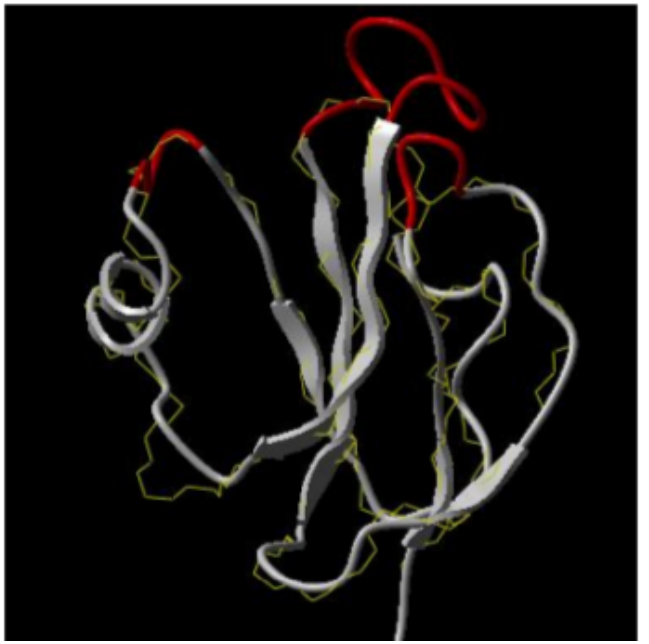
Marti-Renom *et al.* Annu.Rev.Biophys.Biomol.Struct. **29**, 291-325, 2000.

Loop modelling is more complicated

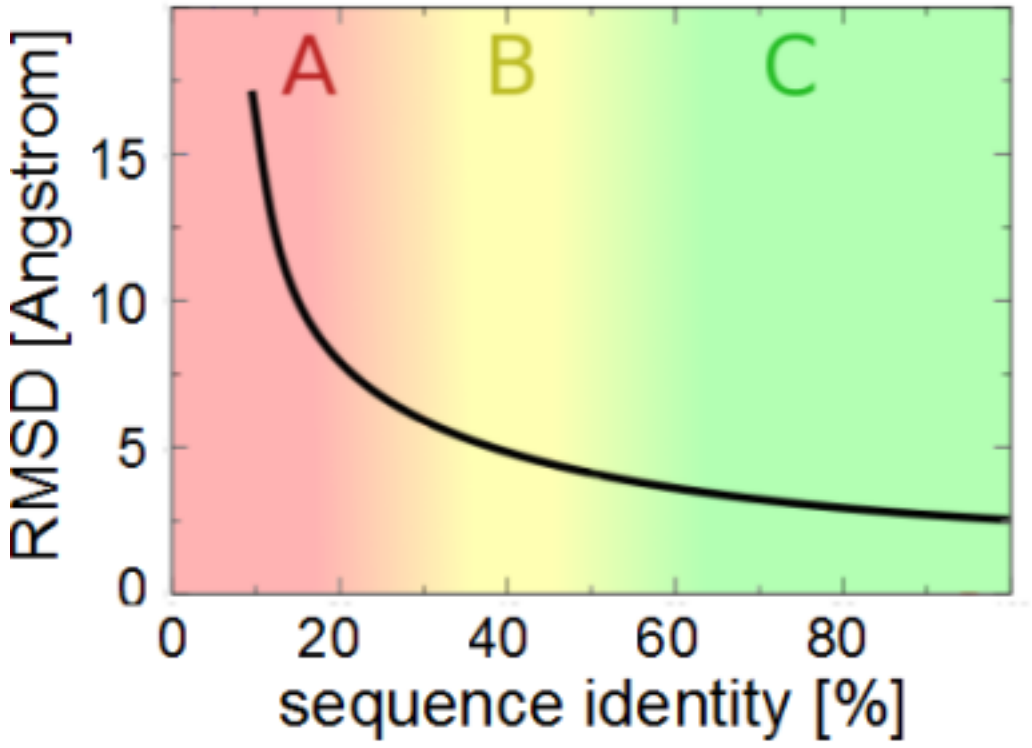
Loop regions are difficult to model:

- In general structure of core regions are conserved
- While loops vary widely.
- Then usually needed modeling insertions, which is efficient for segments of about 8-10

```
CSKPNAPWVKKAVKYLQK-------KNNPQA
CAPPQAPWVLRLREKLDT----SSARKVPNQ
CVNPKEDWVKKHLLFLSQ-------KLKRMS
CSSPTDPIVQKLIKSLDSKRKSTPQRKSKRQ
```



Model quality largely depends on the extent of sequence similarity



Alignment methods depend on identity level

> 30% sequence identity

- Automatic methods for sequence-sequence alignment are usually accurate enough

< 30% sequence identity

- Manual alignment curation required
- Use structural information (e.g. avoid gaps in secondary structure elements)
- Misalignments are critical: each mistake in buried regions is estimated to cause a ~4Å deviation in the model!!
- Therefore for this level of identity, more accurate methods are required

HHpred – Remote Homology detection & structure prediction

HHpred is a method for protein remote homology detection and 3D structure prediction based on the pairwise comparison of profile hidden Markov models (HMM-HMM alignment).

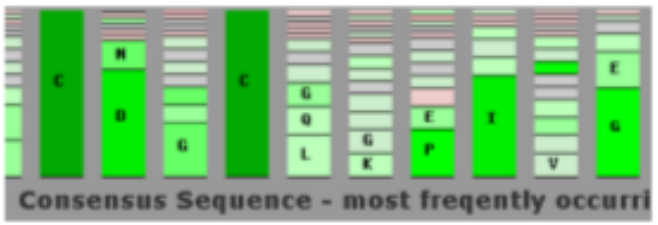
HHpred is as easy to use as BLAST or PSI-BLAST but at the same time is much more sensitive in finding remote homology.

HHpred accepts a single query sequence or a multiple alignment as input and it returns possible templates, E-value, etc.

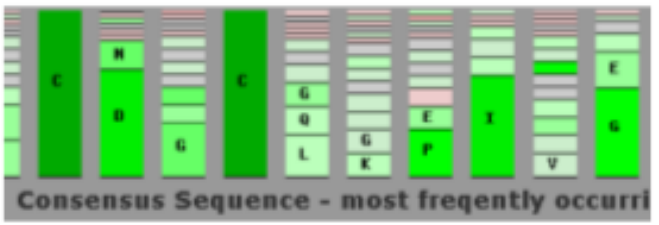
HHpred can can also produce 3D-structural models calculated by the MODELLER software.

profile-profile comparison (alignment) method

1. Calculate template & target **profiles** by constructing alignments them with sequences from a NR database
2. Align the target and the template **profiles**



Align to



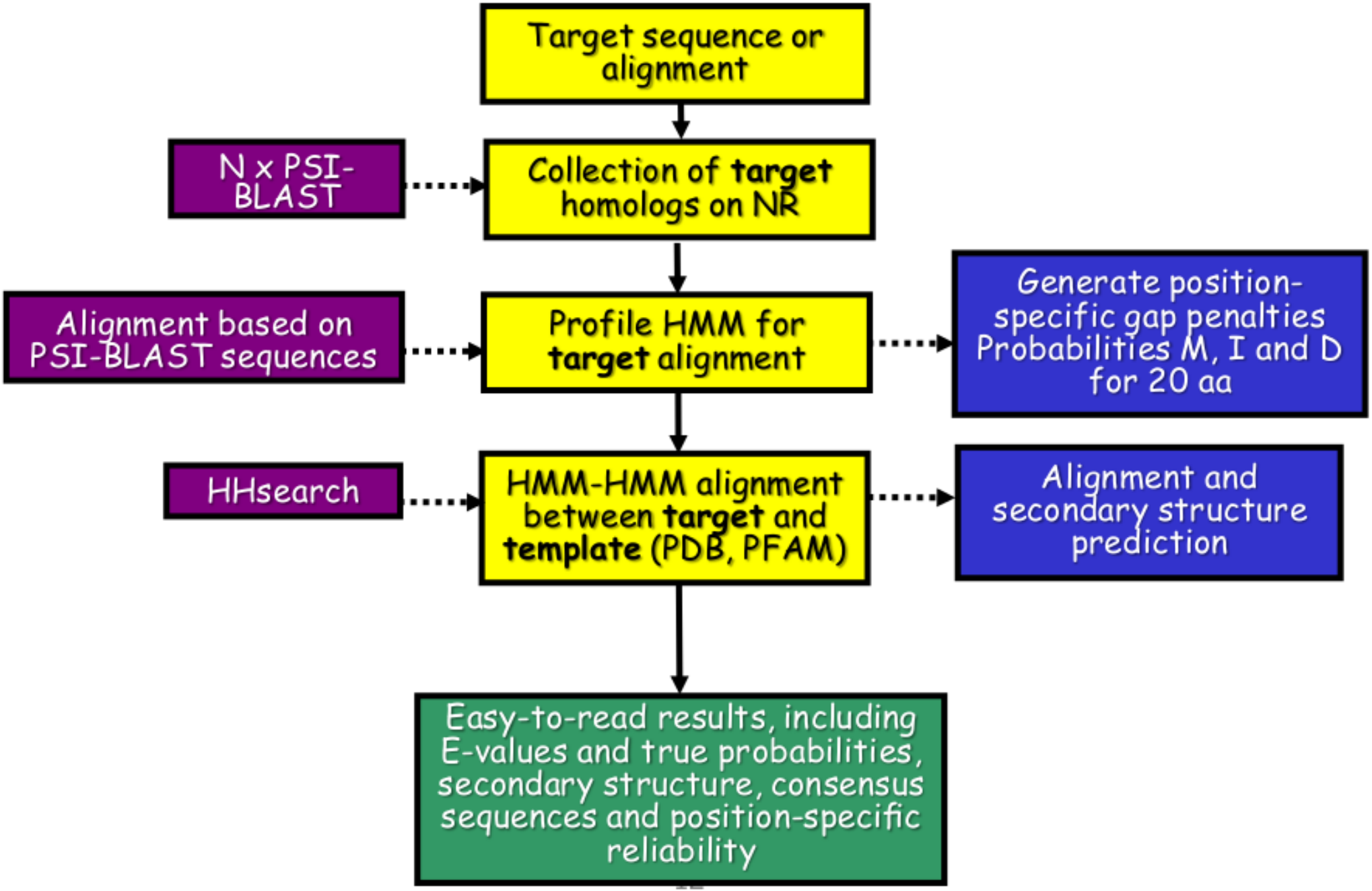
Profile 1

Profile 2

Profiles contain more evolutionary information about the family

Profile-profile alignment methods are able to provide better alignments with distant homologues

HHpred method



How to choose the template?

When we choose from multiple PDB structures

1. Higher sequence similarity
2. Close sub-family
3. Environmental similarity (solvent, pH, ligand, quaternary structure)
4. Quality of the structural template
5. The aim of the model (e.g. protein-ligand model)

How can I verify if a database distant match is really homologous?

1. Check probability and E-value
2. Check if homology is biologically suggestive or at least reasonable
3. Check secondary structure similarity
4. Check relationship among top hits
5. Check for possible conserved motifs (and their residues)
6. Check query and template alignments!
7. Try out other structure prediction servers!
8. Verify predictions experimentally

Evaluation of model

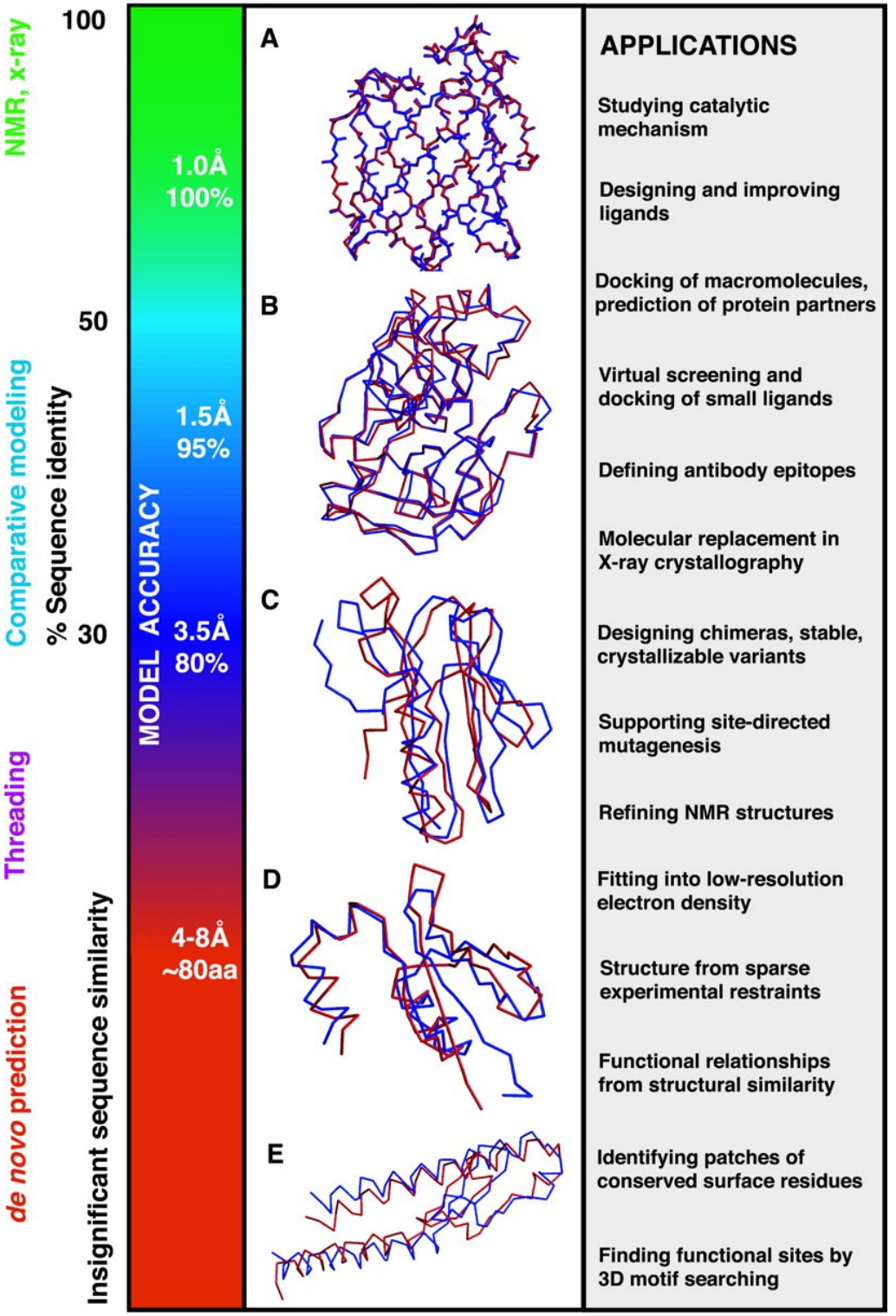
Structural consistency of the model

1) Stereo chemistry
2) Clashes
3) Angles and distances

Independent checks

1) Template checks
2) Pseudo-energy function, unreliable regions
3) Evolutionary conservations
4) Comparing to the observed angles, distances

What can you do with a structural model?



D. Baker & A. Sali.
Science 294, 93, 2001.

What if we can't identify a homolog in the PDB?

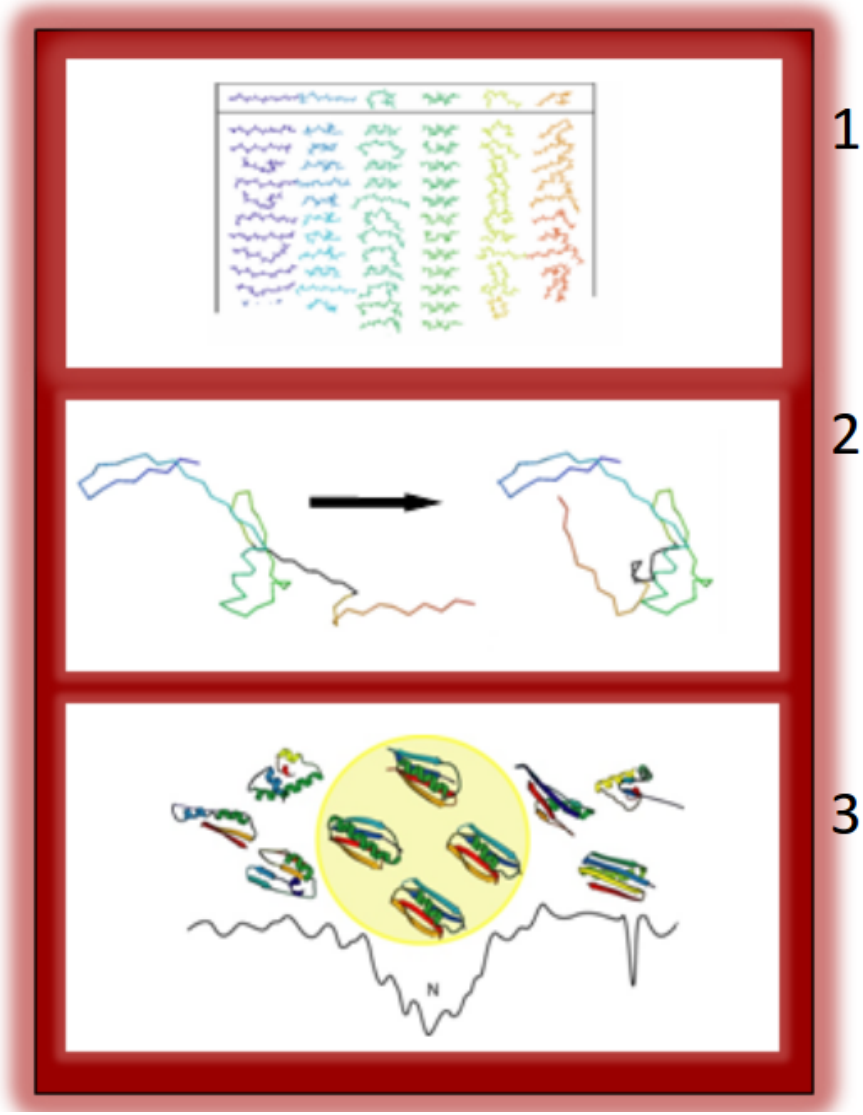
We can still use information based on known structures

– We can construct databases of observed structures of small fragments of a protein

– We can use the PDB to build empirical, "knowledge-based" energy functions

Ab – initio prediction methods

ROSETTA



1. Select fragments consistent with local sequence preferences
2. Assemble fragments into models with native-like global properties
3. Identify the best model from the population of decoys

Figures adapted from Charlie Strauss;
Protein structure prediction using ROSETTA
Rohl et al (2004) *Methods in Enzymology*, **383**:66

Knowledge-based energy functions

Coarse-grained : does not represent all atoms

Statistical potentials: Calculated from the frequency of amino acid interactions in globular proteins

For example:

L-I interaction is frequent (hydrophobic effect)

L-I interaction energy is low (favorable)

K-R interaction is rare (electrostatic repulsion)

K-R interaction energy is high (unfavorable)

Converted into energy like quantities using the Bolztmann statistics

Rosetta all-atom energy function

- Still makes simplifying assumptions:
  - Do not explicitly represent solvent (e.g., water)
  - Assume all bond lengths and bond angles are fixed

- Functional forms are a hybrid between molecular mechanics force fields and the coarse-grained energy function
  - Partly physics-based, partly knowledge-based
    - VdW, electrostatics, H-bond, solvation

I-TASSER

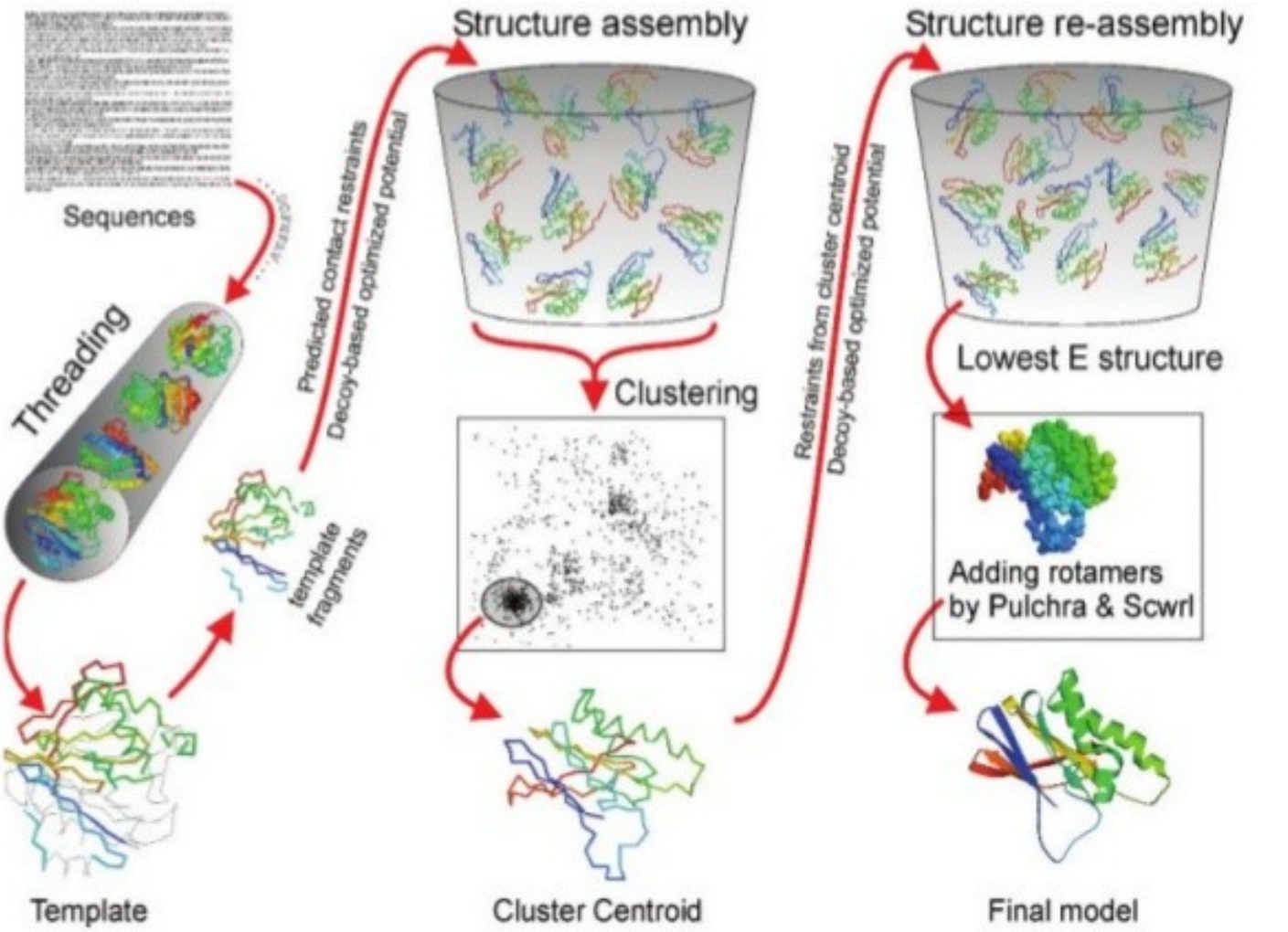


Fig.:Flowchart of I-TASSER protein structure modelling