



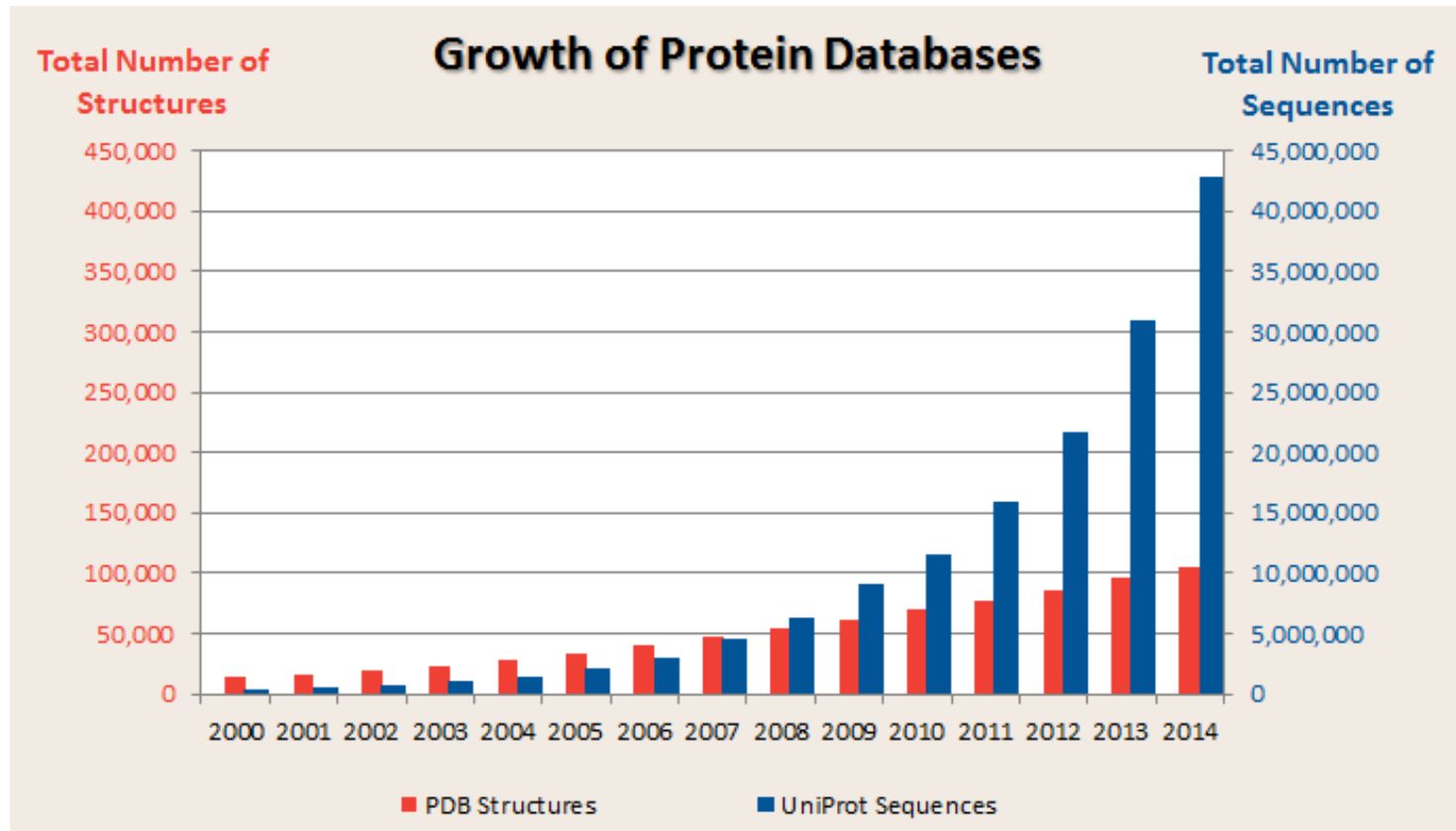
LABORATORY  
of THEORY of  
BIOPOLYMERS

# Proteins - structural bioinformatics (4)

Comparative modeling, simulations of protein dynamics, docking

<http://biocomp.chem.uw.edu.pl>

# Sequence - structure



**Protein Data Bank (PDB) - 140 000 protein structures**

**UniProtKB/TrEMBL sequence database - 133 507 323 nonredundant entries . Nov. 2018**

**Integrated Microbial Genomes & Microbiomes(IMG/M)database of 51 775 423 466 genes**

(Coding genes *E. coli* - 4000, yeast – 6000, human, about -20000)

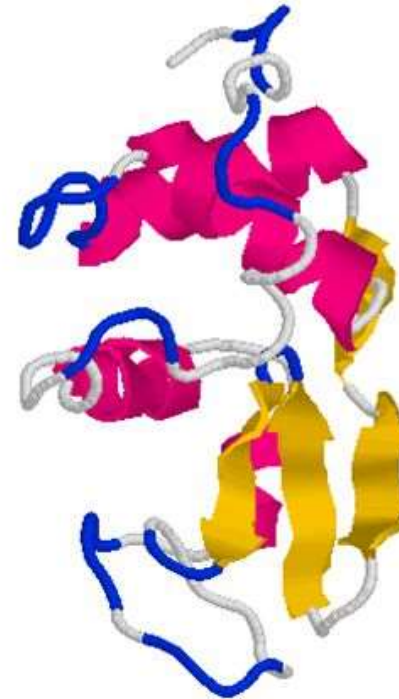
**(First high-resolution structure 1958, by John Kendrew and Max Perutz**

# Protein Folding Problem

A protein folds into a unique 3D structure under physiological conditions

## Lysozyme sequence:

```
KVFGRCELAA AMKRHGLDNY  
RGYSLGNWVC AAKFESNFNT  
QATNRNTDGS TDYGILQINS  
RWWCNDGRTP GSRNLCNIPC  
SALLSSDITA SVNCAKKIVS  
DGNGMNAWVA WRNRCKGTDV  
QAWIRGCRL
```



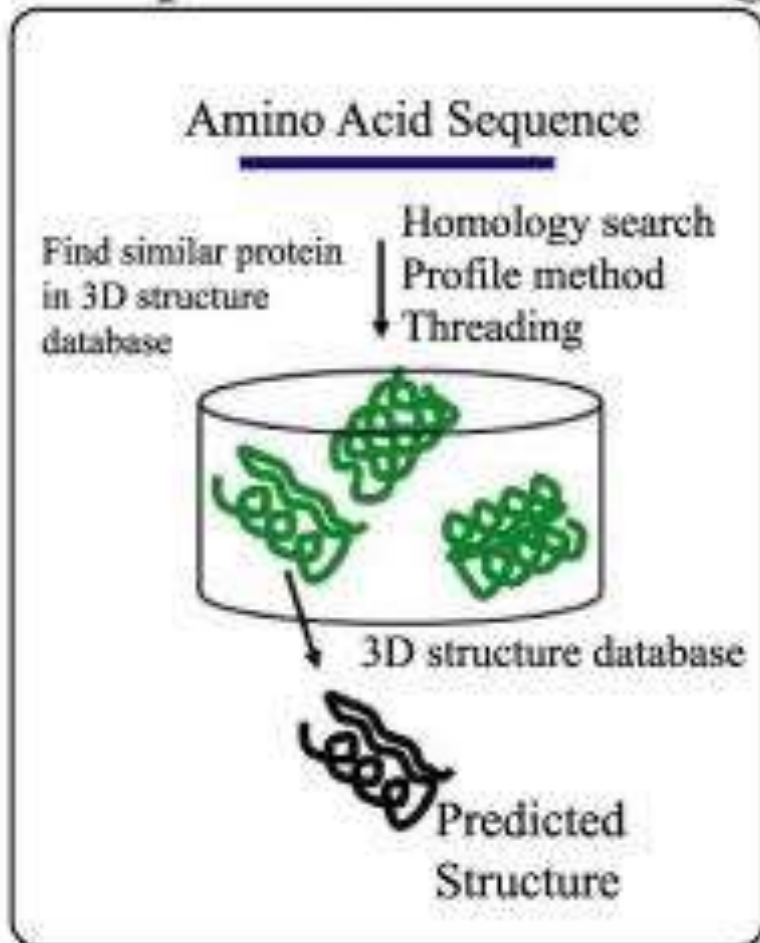
Anfinsen, 1960: denatured proteins can refold to active enzymes

# Structure Prediction

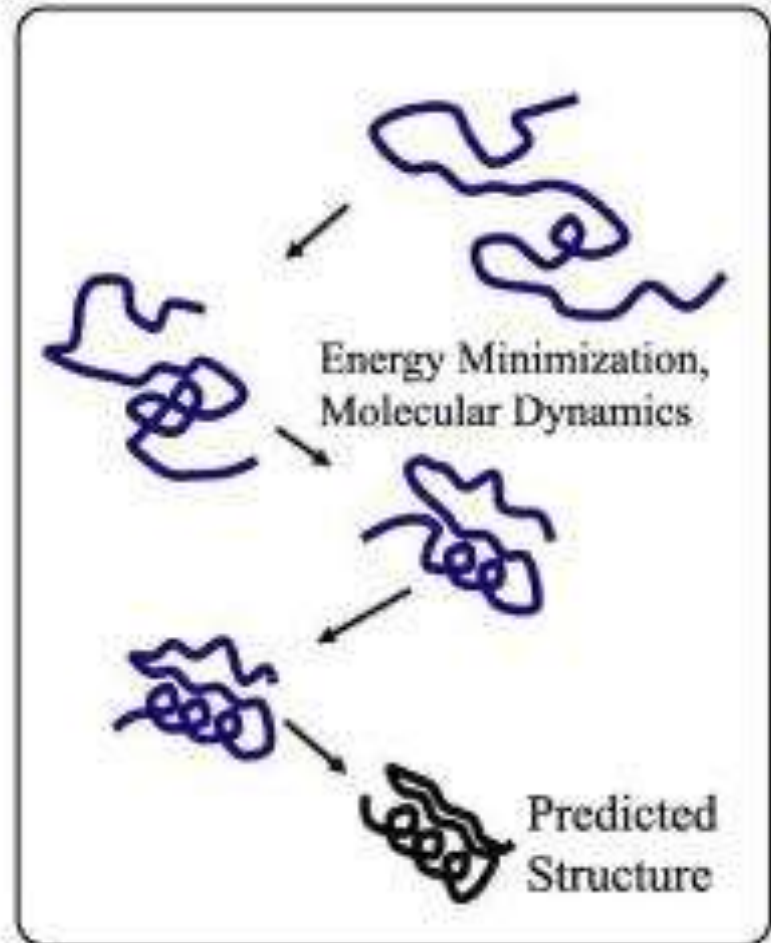
- Experimentally solved structures (130 000) – about 0.11% of the number of protein sequences deposited in UniprotKB and TrEMBL
- Theoretical predictions (we know about 1500 folds from 5000 – 20,000 of possible)
  - *de novo* prediction (Protein folding problem)
  - comparative modeling (Most of newly identified protein structures are similar to already known)

# Structure Prediction

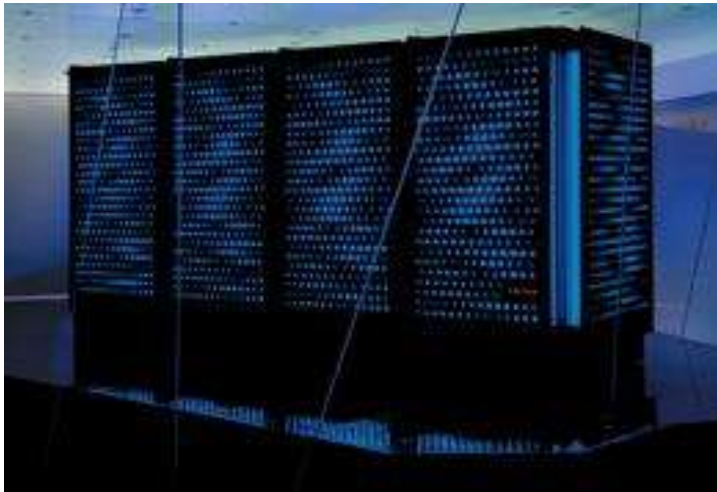
## Comparative Modelling



## *Ab initio* Prediction



# Protein folding problem - the Holy Grail of the structural biology



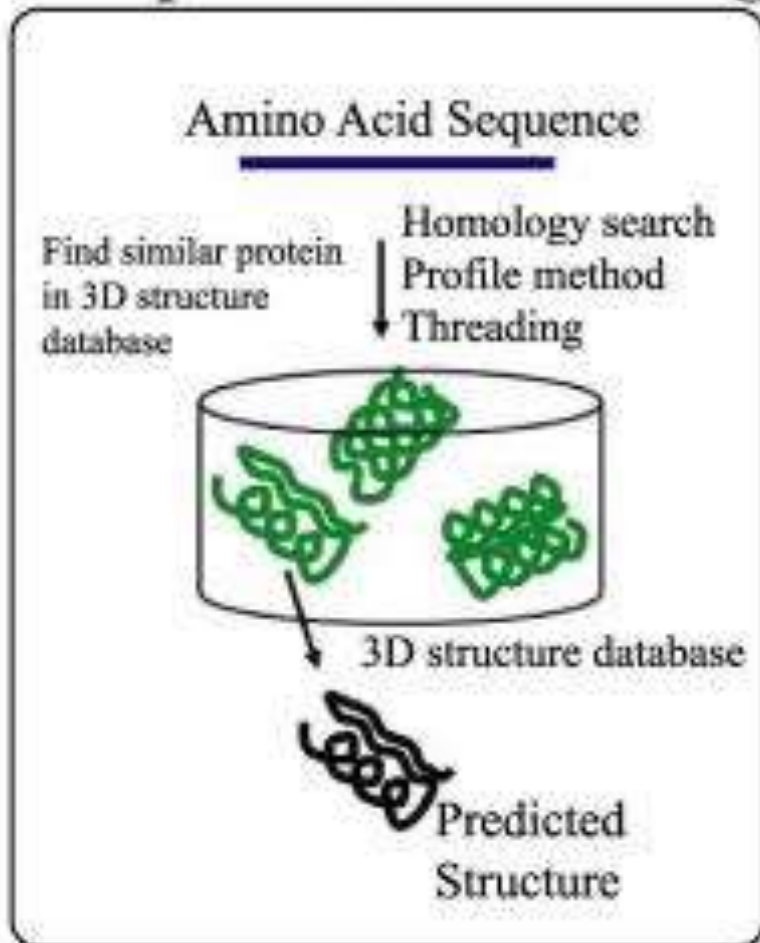
Anton  
David E. Shaw Research

All-atom MD with explicit water  
- milliseconds of folding process  
of a small protein.

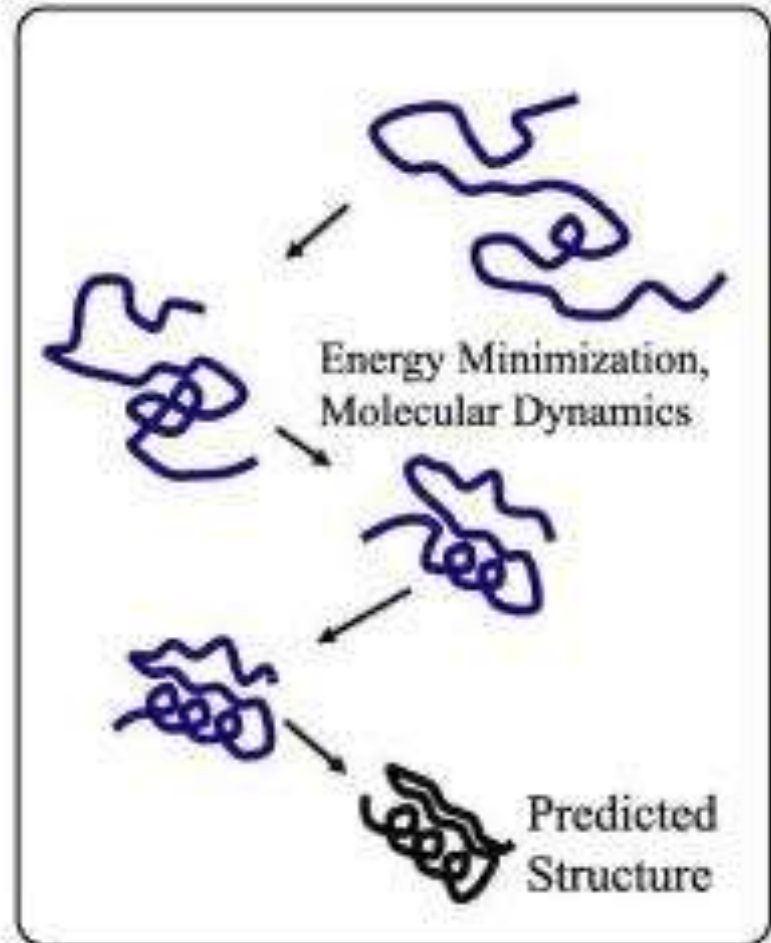
For realistic modeling of larger  
biomolecular systems, including  
flexible protein-protein docking, **we  
need much faster simulations.**

# Structure Prediction

## Comparative Modelling



## *Ab initio* Prediction



# Local Alignment

## Pairwise Sequence Alignment

Target Sequence

5' ACTACTAGATTACTTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||

Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'

# Global Alignment

Target Sequence

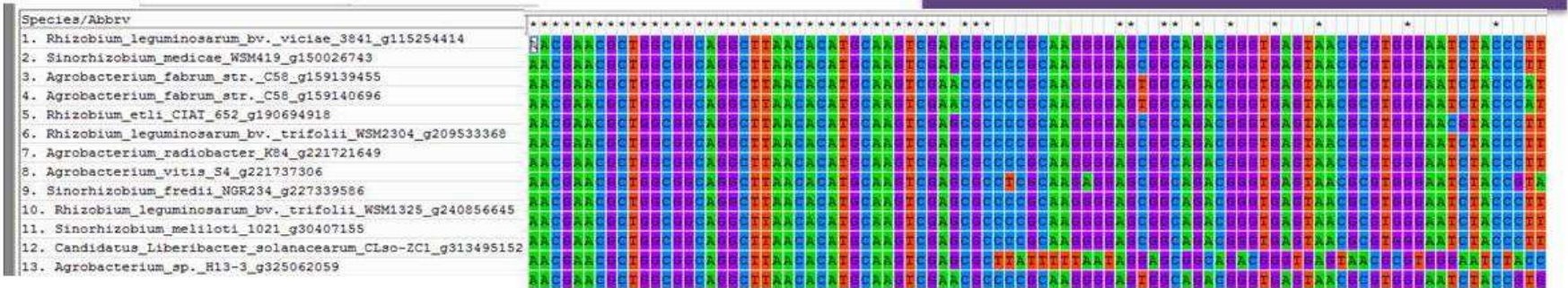
5' ACTACTAGATTACTTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||

5' ACTACTAGATT - - - -ACGGATC - -GTACTTTAGAGGCTAGCAACCA 3'

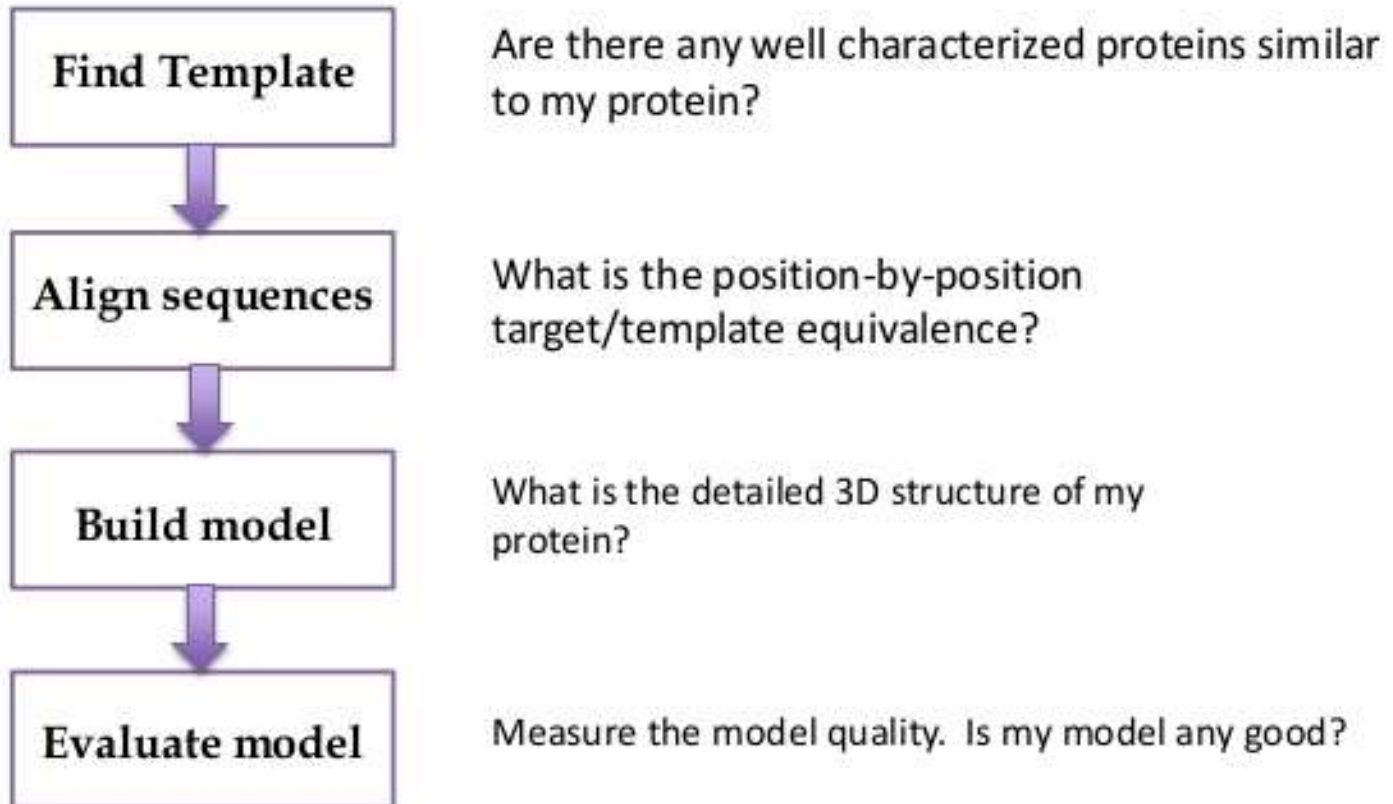
Query Sequence

## Multiple Sequence Alignment (MSA)





# Homology modeling workflow

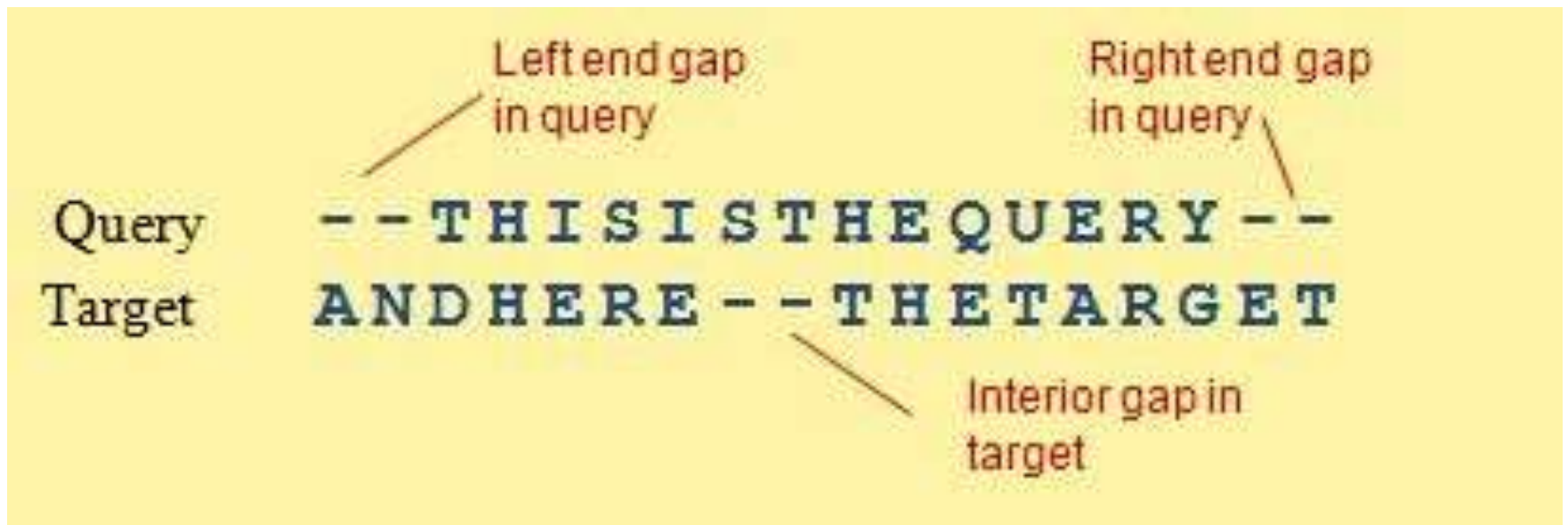


# Comparative Modeling--Basic Protocol

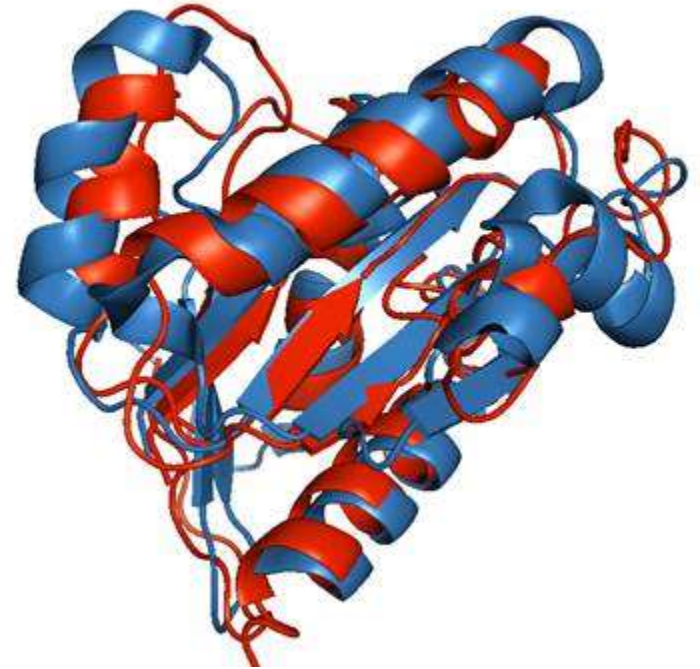
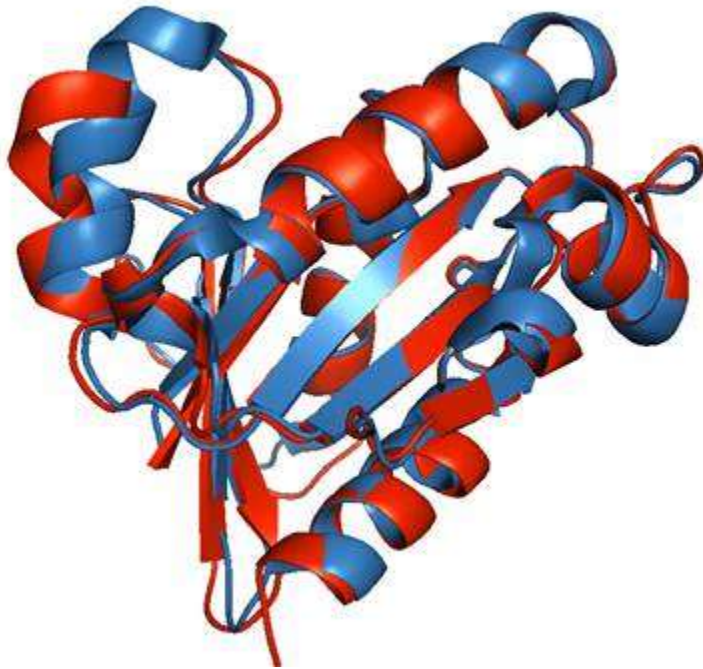
42

1. **Identification** of homologue for target sequence
2. **Alignment** of target sequence to template sequence and structure
3. **Side-chain modeling**, copy the backbone of the template and model the new side chains onto this backbone
4. **Loop modeling**, for insertions and deletions in the alignment
5. **Refinement of model** -- moving template closer to target
6. **Assessment** of (predicted) model quality
7. **Using the model** to explain experiments and guide new ones

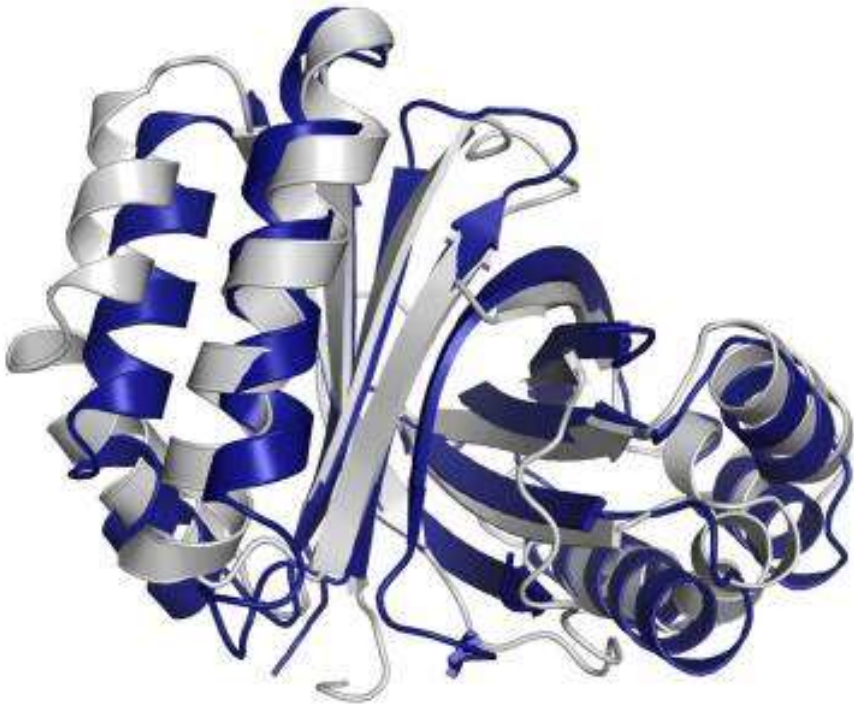
# Structure – Comparative modeling – alignment gaps



# Comparative (homology) modeling

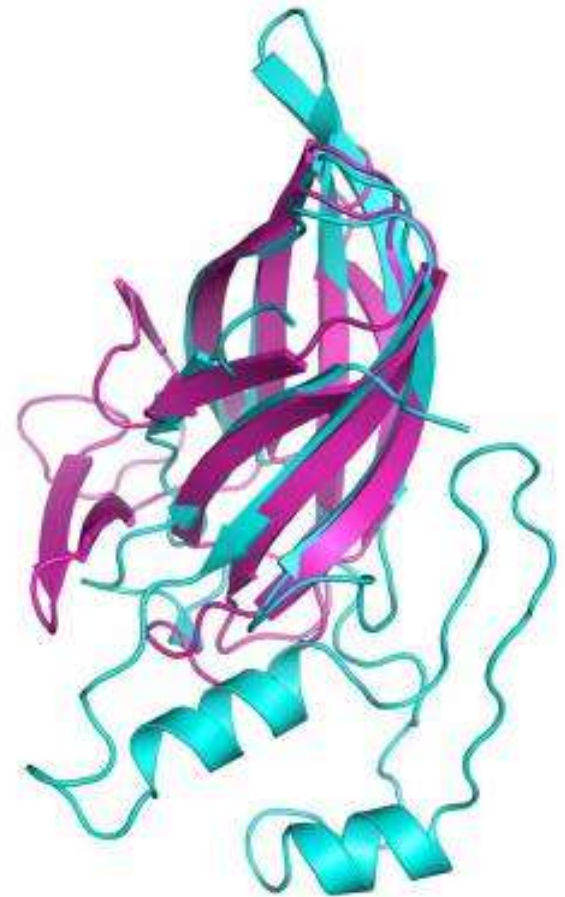


# Comparative (homology) modeling



A

Both cases (A,B) represent extremely distant homologies with sequence identity on the level of 10–12%



B

## Sources of errors

- experimental errors and uncertainties in X-ray, NMR

1Å  
100%



## Applications

- studying catalytic mechanism / function

- side-chain packing  
- mis-placed side-chains

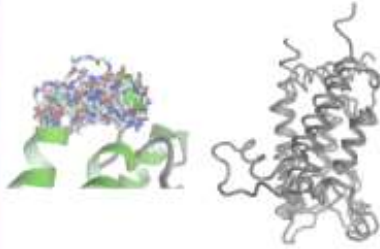
1.5Å  
95%



- structure-based drug design, ligand docking

- modeling of loop regions (insertions and deletions)

60%



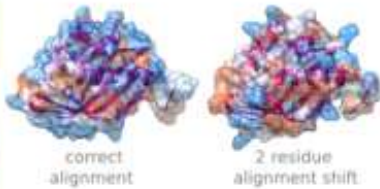
- structural support for mutagenesis studies

- distortions of aligned regions

- molecular replacement

- alignment errors

3Å  
40%



- integrative modeling

- modeling into low-resolution density maps

- sub-optimal template selection

>3Å  
<30%



- domain boundaries

- model may even have the wrong fold

- identification of structural motives

# Comparative (homology) modeling

## MODELLER

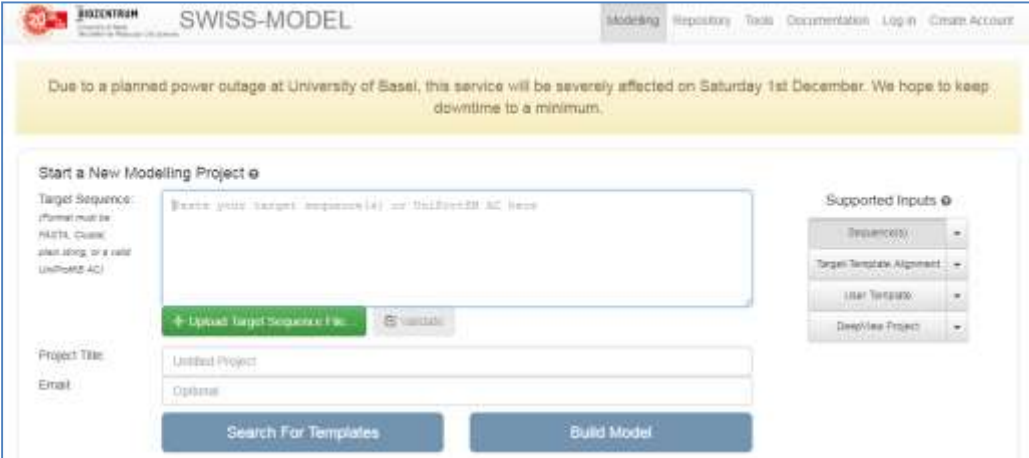
<https://salilab.org/modeller/>



The screenshot shows the MODELLER website homepage. At the top, the word "Modeller" is written in a large, red, serif font. Below it, the text reads "Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints". To the right of this text is a 3D ribbon diagram of a protein structure, colored in various colors (green, blue, yellow, red). Below the ribbon diagram is a sequence logo or alignment visualization. On the left side, there is a vertical navigation menu with links: "About MODELLER", "MODELLER News", "Download & Installation", "Release Notes", "Data File Downloads", "Registration", "Biopython Wrappers", "Discussion Forum", and "Subscribe". The main content area is titled "About MODELLER" and contains a paragraph of text describing the program's capabilities: "MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is available for download for most Unix/Linux systems, Windows, and Mac. Several graphical interfaces to MODELLER are commercially available. There are also many other resources and people using Modeller in".

## SWISS-MODEL

<https://swissmodel.expasy.org>



The screenshot shows the SWISS-MODEL website interface. At the top, there is a navigation bar with the SWISS-MODEL logo and the text "SWISS-MODEL". To the right of the logo are links for "Modeling", "Repository", "Tools", "Documentation", "Login", and "Create Account". Below the navigation bar is a yellow banner with a message: "Due to a planned power outage at University of Basel, this service will be severely affected on Saturday 1st December. We hope to keep downtime to a minimum." Below the banner is a form titled "Start a New Modelling Project". The form has several input fields: "Target Sequence:" (with a placeholder text: "Enter your target sequence(s) or UniProtKB AC here"), "Project Title:" (with a placeholder text: "Untitled Project"), and "Email:" (with a placeholder text: "Optional"). There are two buttons at the bottom of the form: "Search For Templates" and "Build Model". To the right of the form is a "Supported Inputs" section with a dropdown menu showing "Sequences", "Target Template Alignment", "User Template", and "Download Project".

# MODELLER (Sali)

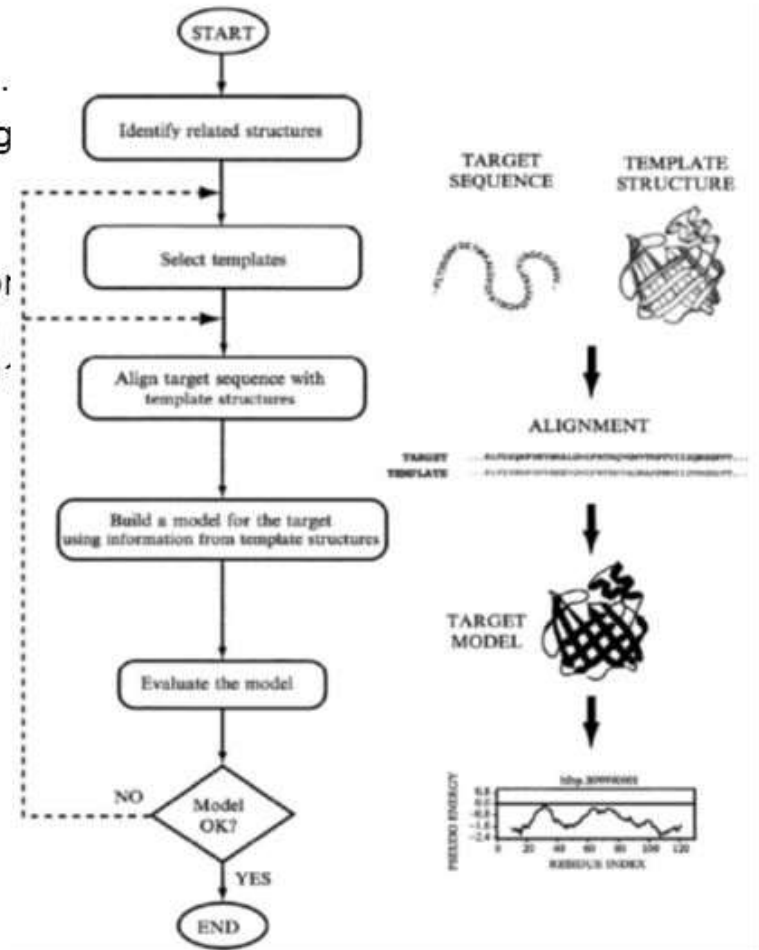
- references

- A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815, 1993.
- A. Fiser, R. K. G. Do and A. Šali. Modeling of loops in protein structures. *Protein Science* 9, 1753-1773, 2000.
- Fiser A, Sali A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enz.* 374:461-9

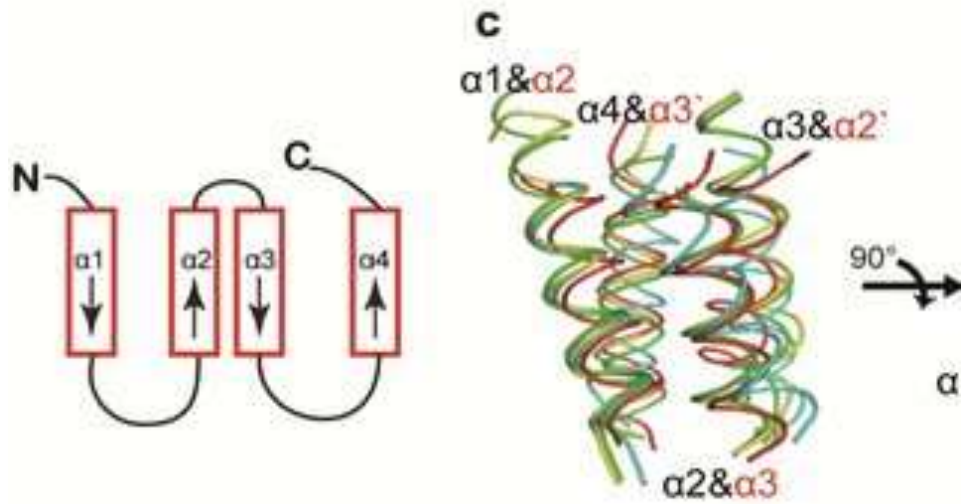
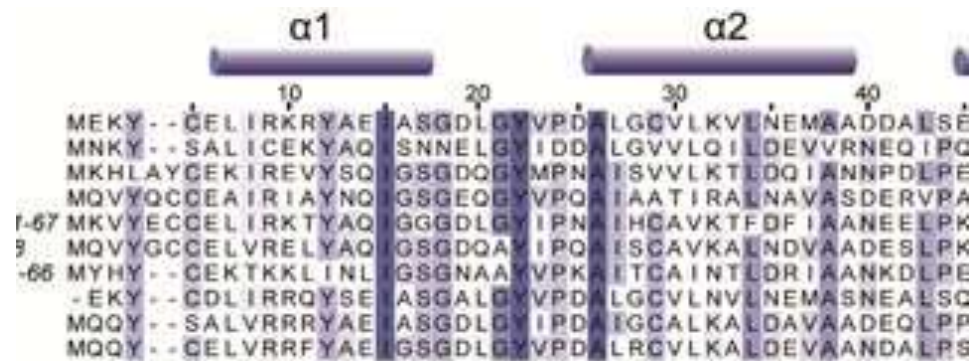
- loop-modeling via dynamics

- evaluation:

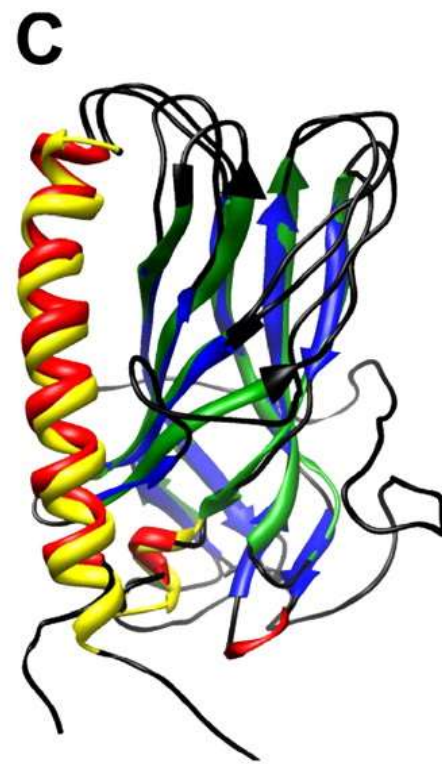
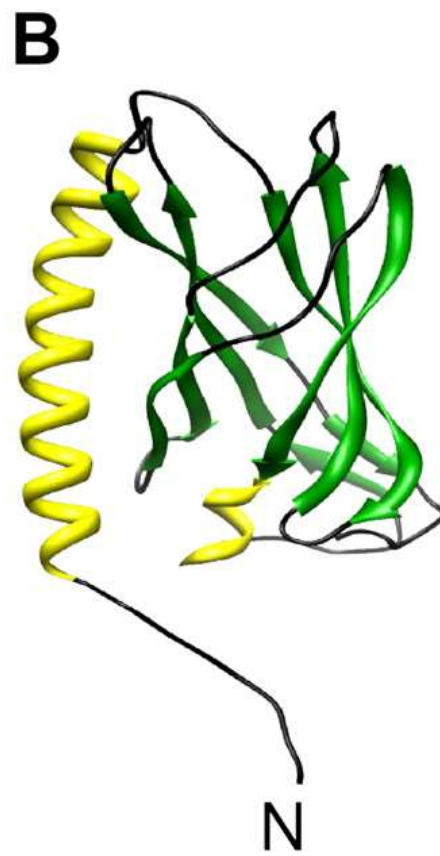
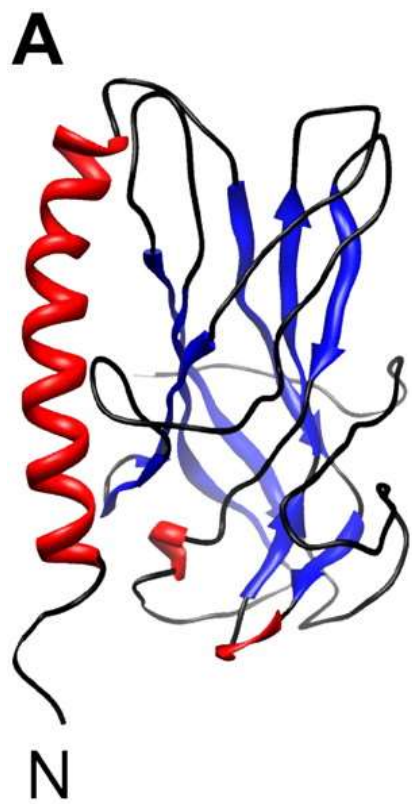
- >30% identity?
- stereochemistry: Procheck
- contacts/exposure: ProSA (Sippl, 1993) – distance-based pair potentials





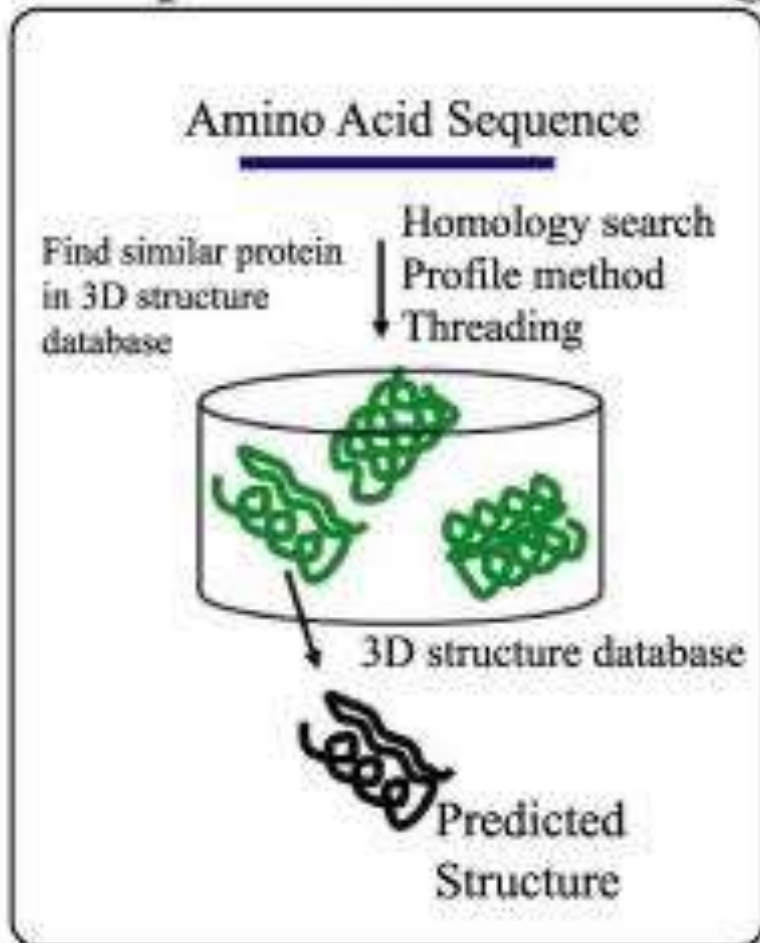


Note problems with sequence alignments

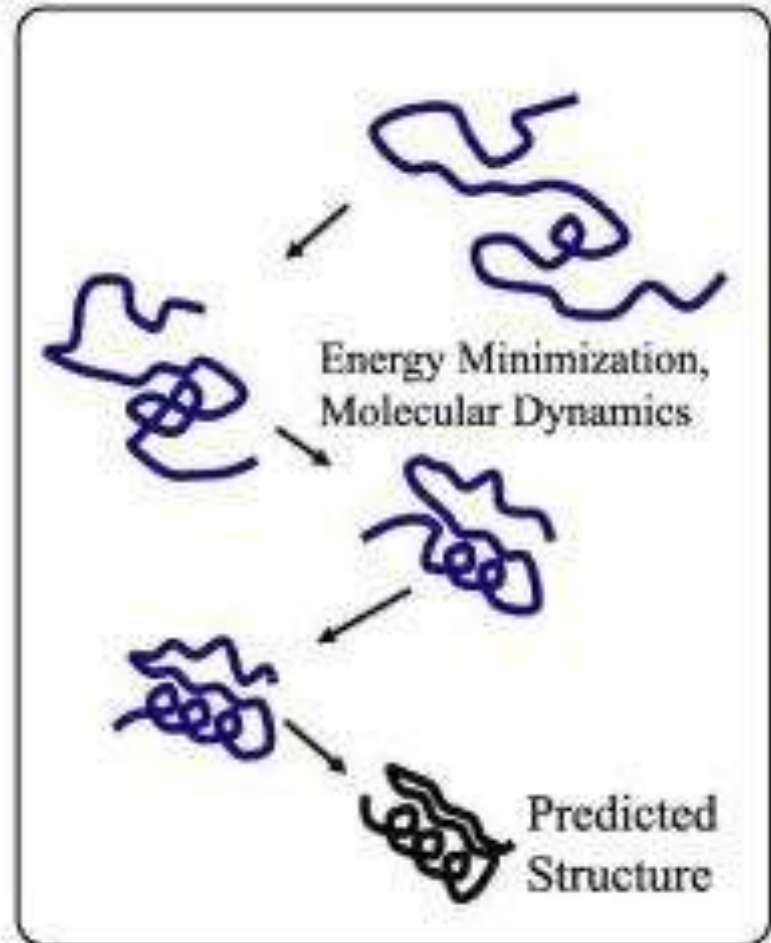


# Structure Prediction

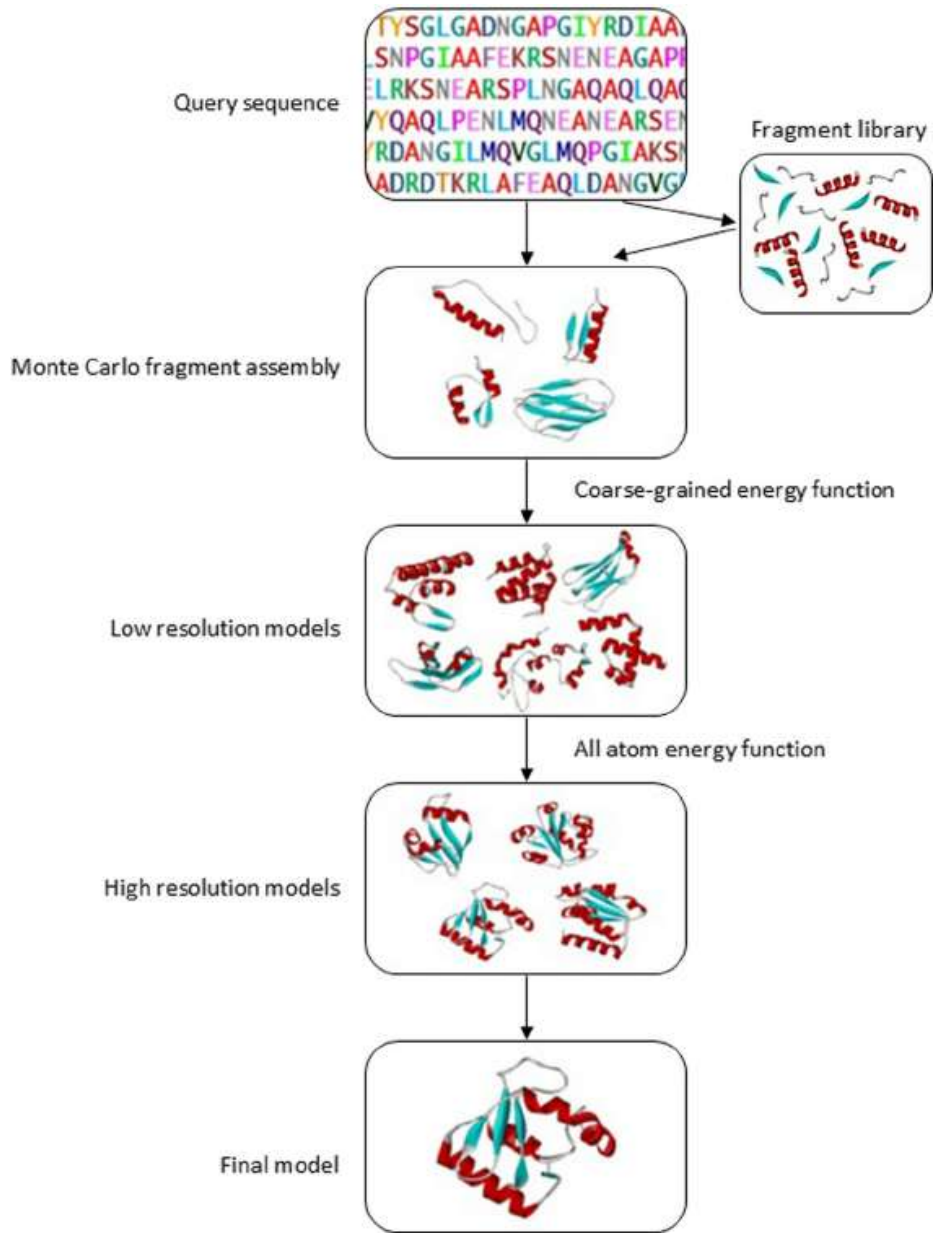
## Comparative Modelling



## *Ab initio* Prediction

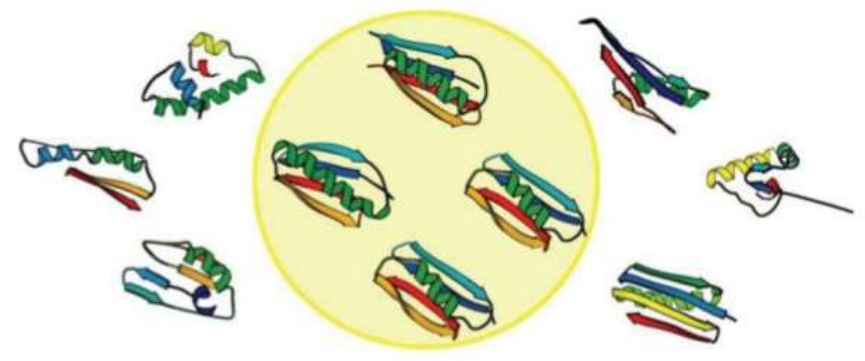


# Rosetta Dawid Baker



Rosetta protein modeling consists of sampling and scoring

15



# Rosetta in CASP4

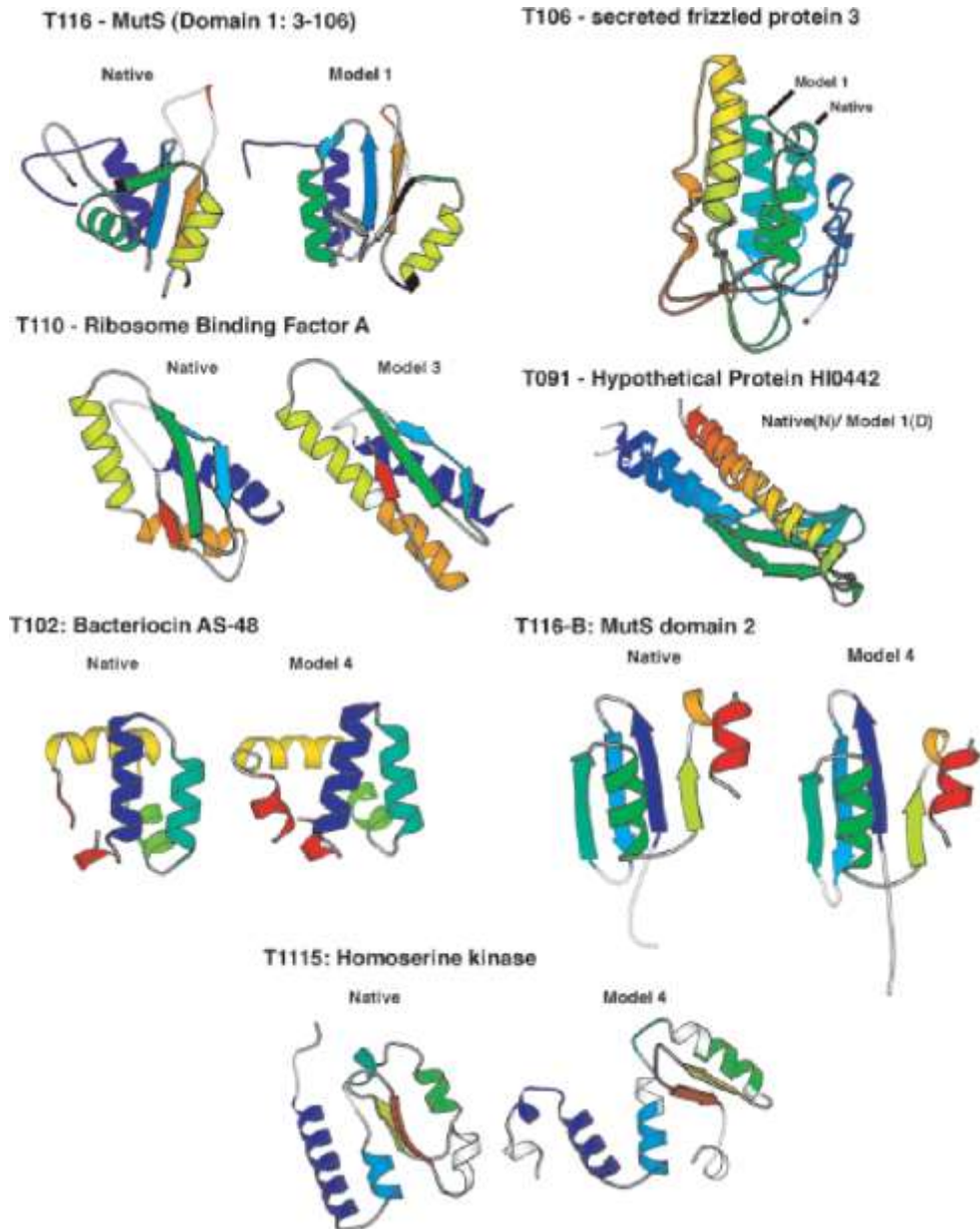


FIG. 3. Comparison of predicted and native structures. Corresponding sequence regions are colored

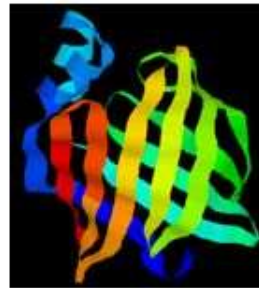
# Concept of Threading

- Thread (*align* or *place*) a query protein sequence onto a template structure in “optimal” way
- Good alignment gives approximate backbone structure

## Query sequence

MTYKLLILNGKTKGETTTEAVDAATAEKVVFQYANDNGVDGEWTYTE

## Template set



---

# Protein threading

**Structure is better conserved than sequence**

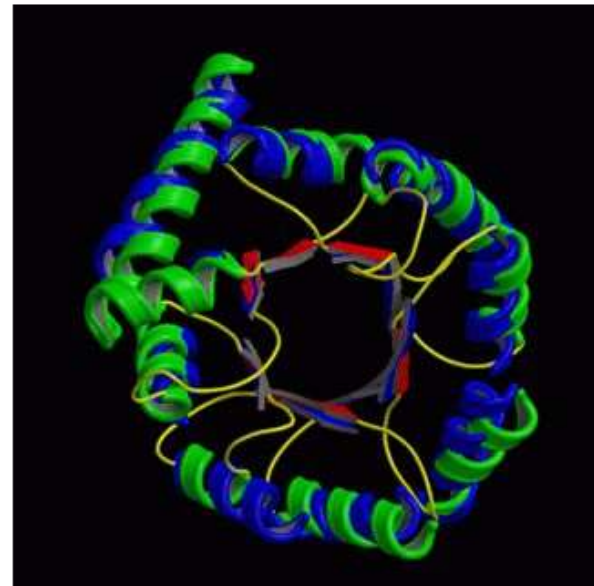
Structure can adopt a wide range of mutations.

Physical forces favor certain structures.

Number of folds is limited.

Currently ~700

Total: 1,000 ~10,000

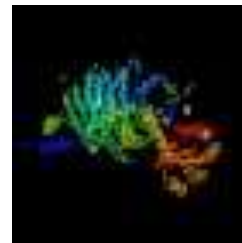
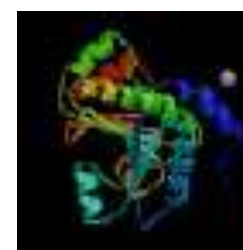
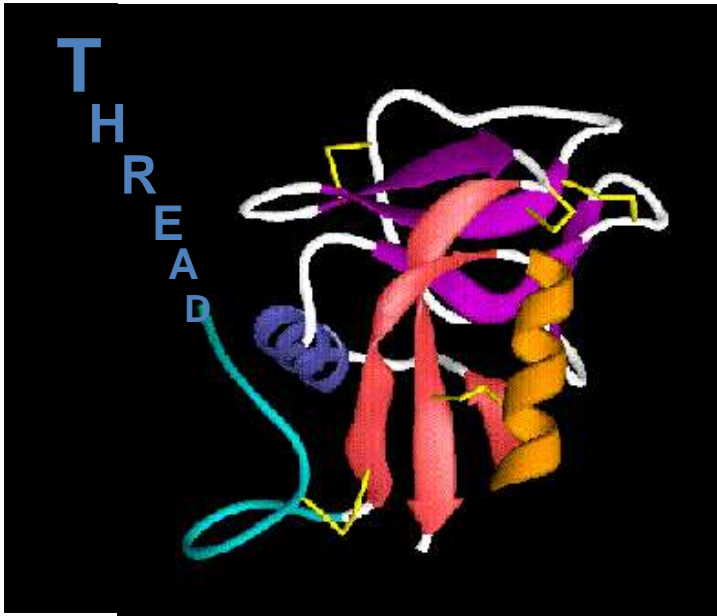


TIM barrel

---

# Visualizing Threading

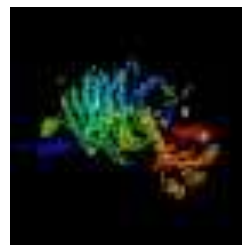
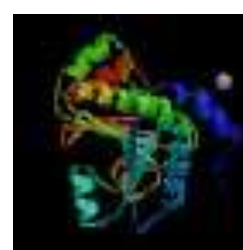
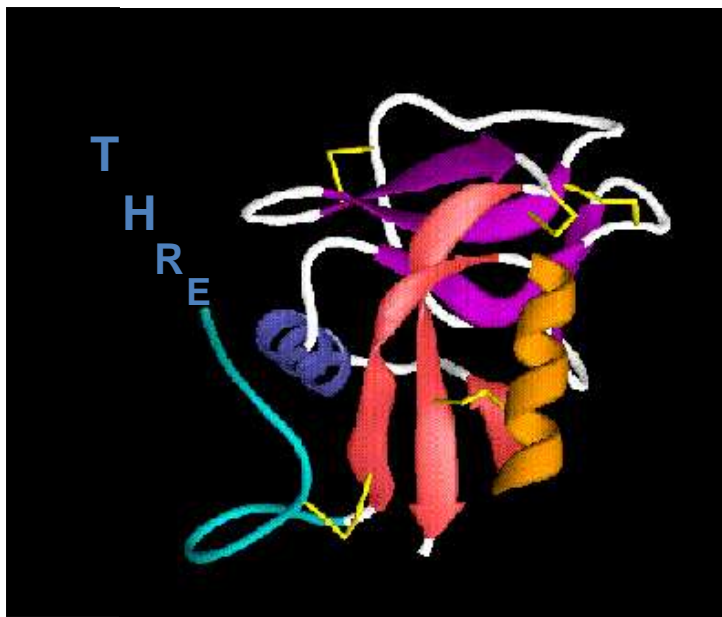
THREADINGSEQNCEECSGNIE  
RHTHREADINGSEQNCETHREAD  
GSEQNCEQCQESGIDAERTHR...





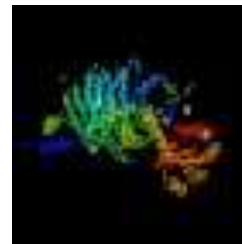
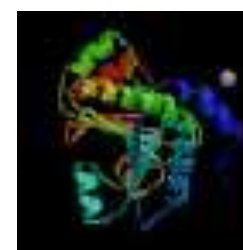
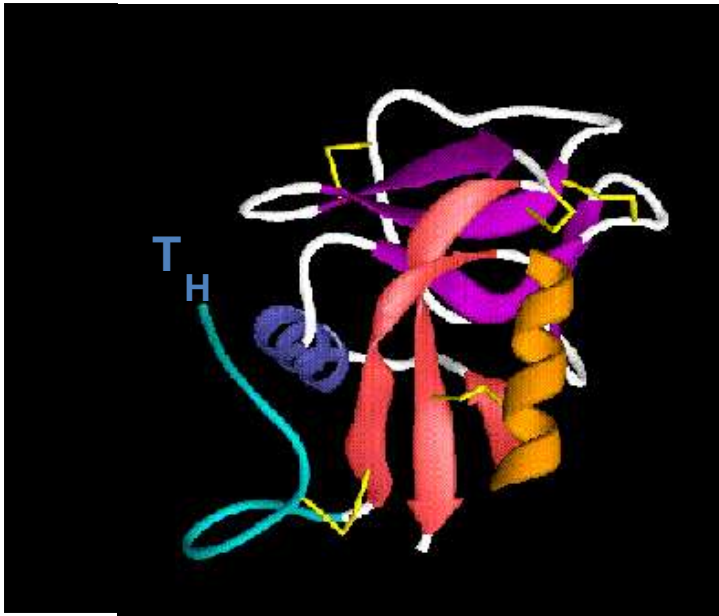
# Visualizing Threading

THREADINGSEQNCEECSGNI  
ERHTHREADINGSEQNCETHREAD  
GSEQNCEQCQESGIDAERTHR...



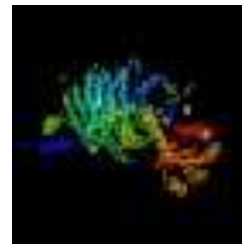
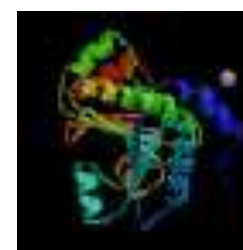
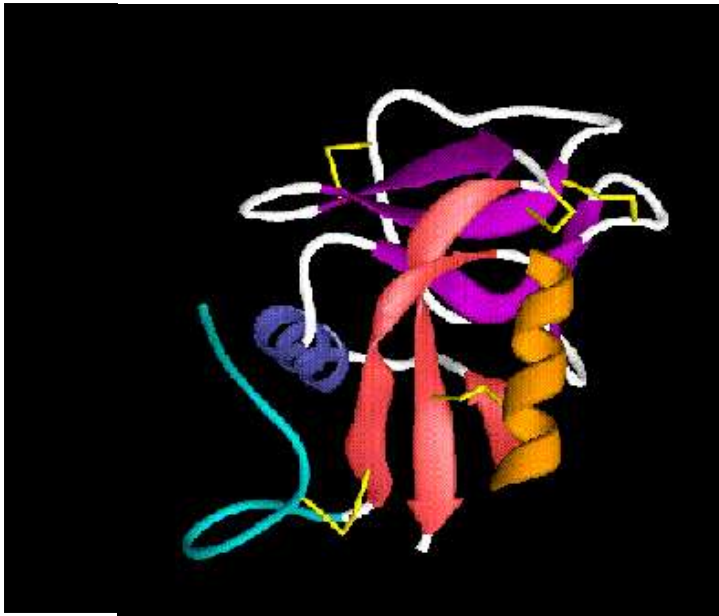
# Visualizing Threading

THREADINGSEQNCEECSGNI  
ERHTHREADINGSEQNCETHREAD  
GSEQNCEQCQESGIDAERTHR...



# Visualizing Threading

THREADINGSEQNCEECSGNI  
ERHTHREADINGSEQNCETHREAD  
GSEQNCEQCQESGIDAERTHR...

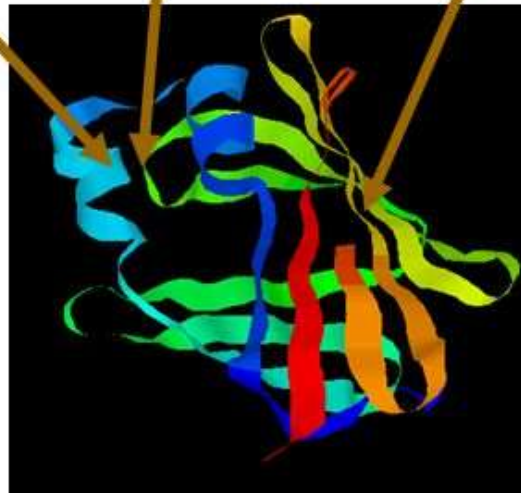


# Protein Threading – energy function

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

how preferable to put  
two particular residues  
nearby:  $E_p$

alignment gap  
penalty:  $E_g$



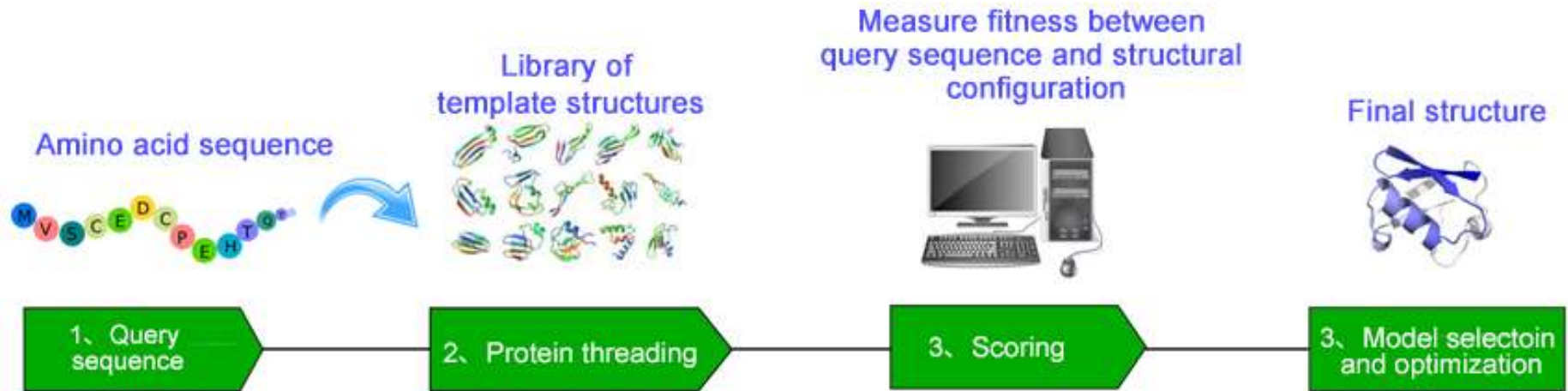
how well a residue fits  
a structural  
environment:  $E_s$

total energy:  $E_p + E_s + E_g$

**find a sequence-structure alignment  
to minimize the energy function**

# Comparative (homology) modeling

## Threading instead of sequence alignment



# I-TASSER (Y. Zhang)

