



LABORATORY  
of THEORY of  
BIOPOLYMERS

# Proteins - structural bioinformatics (3)

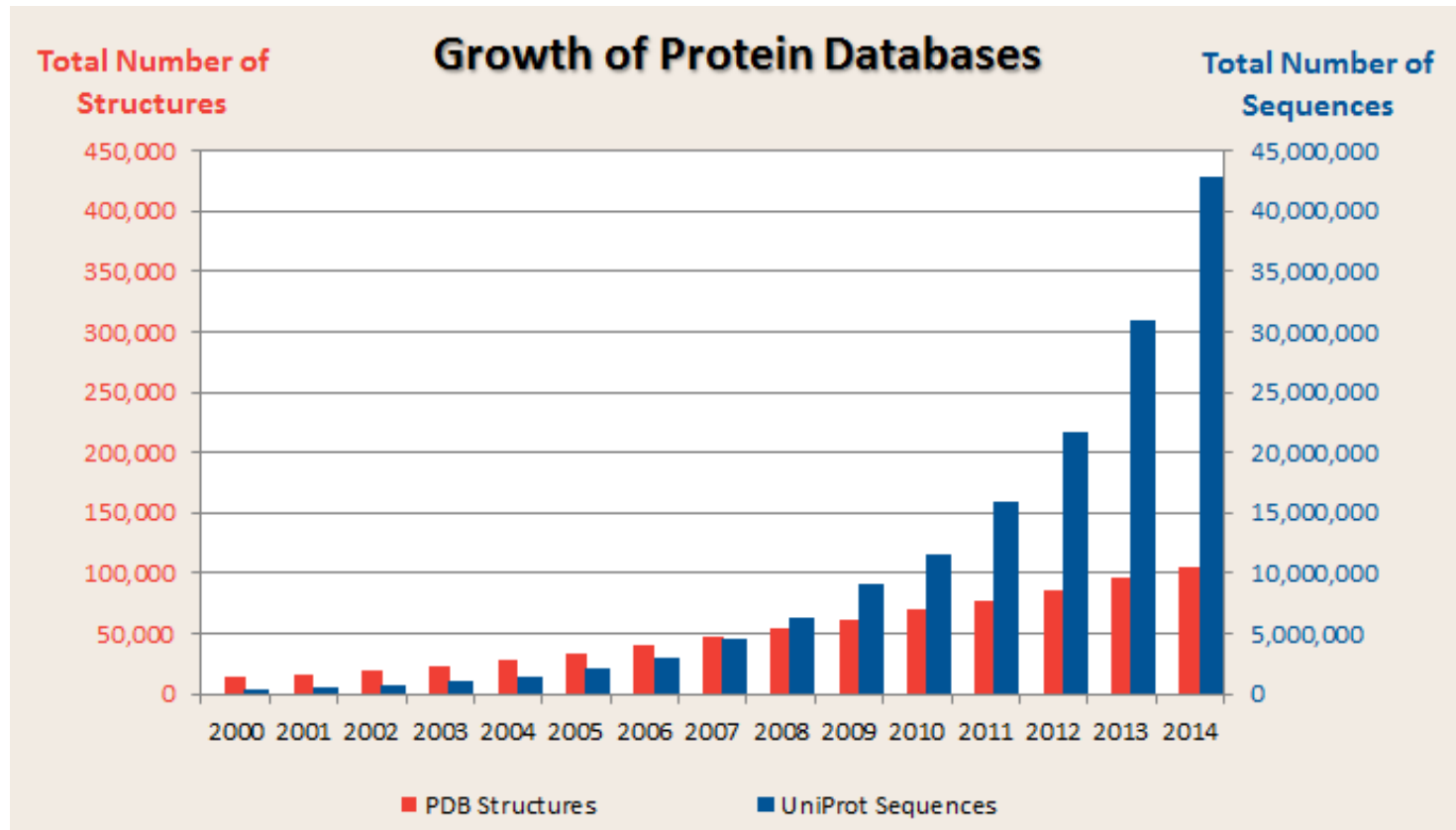
## Comparative modeling (homology modeling)

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

# How many proteins?

- Just 150 AA protein -  $10^{195}$  sequences
- Eukaryotic protein universe  $\sim 10^{12}$
- Prokaryotic – much more, difficult to estimate
- 12-14 thousand of known protein families cover about 60% of known proteins
- 5000 – 20,000 of possible folds ( about 1500 currently known)

# 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)

# Structure Prediction

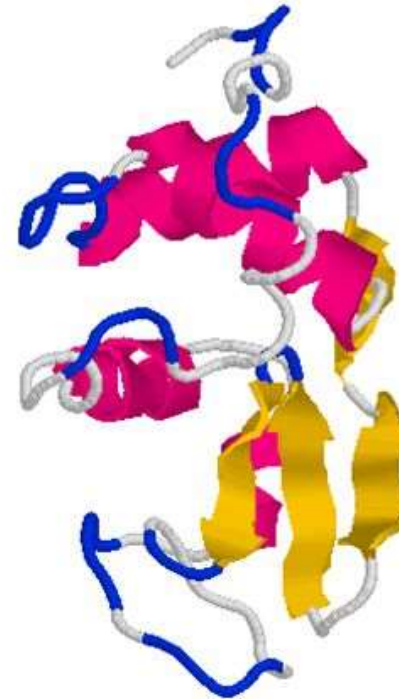
- Experimentally solved structures (130 000) – about 0.11% of the number of protein sequences deposited in UniprotKB and TrEMBL
- Theoretical predictions (we know about 1500 folds from 5000 – 20,000 of possible)
  - *de novo* prediction (Protein folding problem)
  - comparative modeling (Most of newly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

# Protein Folding Problem

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

## Lysozyme sequence:

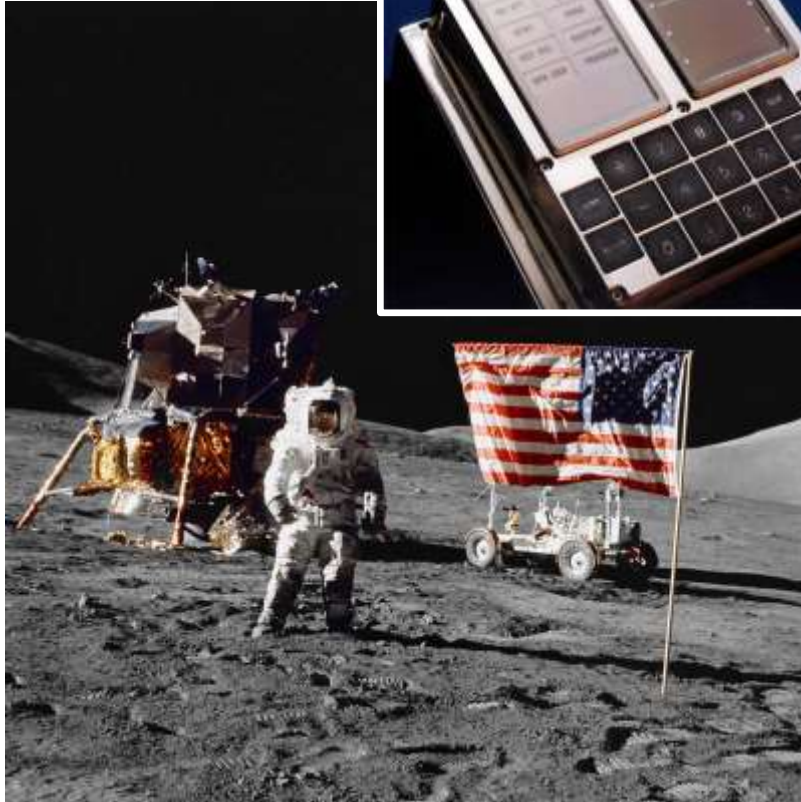
```
KVFGRCELAA AMKRHGLDNY  
RGYSLGNWVC AAKFESNFNT  
QATNRNTDGS TDYGILQINS  
RWWCNDGRTP GSRNLCNIPC  
SALLSSDITA SVNCAKKIVS  
DGNGMNAWVA WRNRCKGTDV  
QAWIRGCRL
```



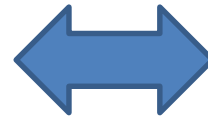
Anfinsen, 1960: denatured proteins can refold to active enzymes

# Computing power

1969



2018

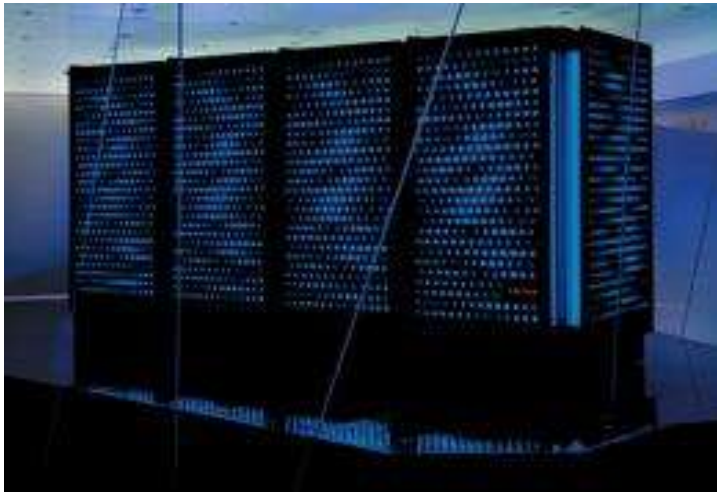


APOLLO MISSION  
120,000,000

/

smartphone  
1

# 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.

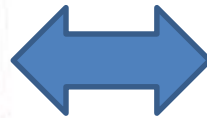
For realistic modeling of larger  
biomolecular systems, including  
flexible protein-protein docking, **we  
need much faster simulations.**

# Computing power

5 MB hard drive in 1956



128 GB pen drive in 2017





# CASP and CAPRI



## CASP Competition

---

- CASP competition (Critical Assessment of Techniques for Protein Structure Prediction)  
<http://predictioncenter.llnl.gov/>
- Their goal is to help advance the methods of identifying protein structure from sequence.

## CASP Experiment

- Experimentalists are solicited to provide information about structures expected to be soon solved
- Predictors retrieve the sequence from prediction center (predictioncenter.llnl.gov)
- Deposit predictions throughout the season
- Meeting held to assess results

29

Polish scientists in CASP: Ginalski, Rychlewski, Bujnicki, Kolinski, Liwo, and others

# CASP –every 2 years since 1994

## Leading trends:

- Art of modeling (knowledge-based homology modeling) by Alexey Murzin
- Careful alignment + Modeller by Krzysztof Ginalski
- Rosetta fragment assembly (comparative and de novo) by David Baker and co-workers
- Refined alignments and Coarse-Grained modeling using CABS tools by Janusz Bujnicki and Andrzej Kolinski
- Sophisticated ranking of alignments and fragment modeling using CAS (a version of CABS) by Yang Zhang
- Computer deep-learning and fragment assembly (Rosetta) Lee Sedol

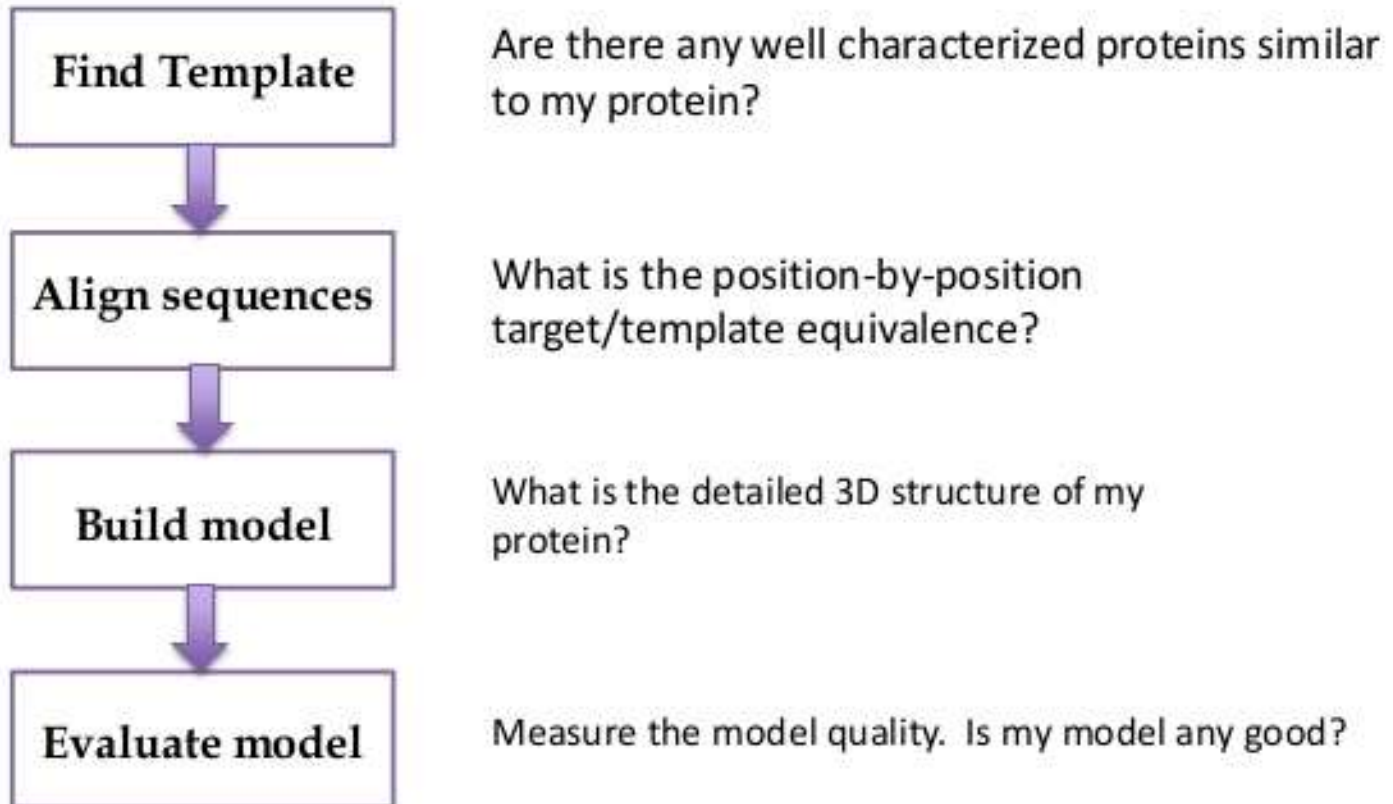
# CASP6: Average scoring, all categories

(New Folds, Fold Recognition, Comparative Modeling)

- 1 Ginalski (ICM, POLAND)
- 2 Kolinski & Bujnicki (UW-IIMCB, POLAND)
- 3 Baker (USA)
- 4 Skolnick\_Zhang (USA)
- 5 GeneSilico (IIMCB, POLAND)

A. Kolinski and J. M. Bujnicki, "Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models", *Proteins* **61**(S7):84-90 (2005)

# Homology modeling workflow

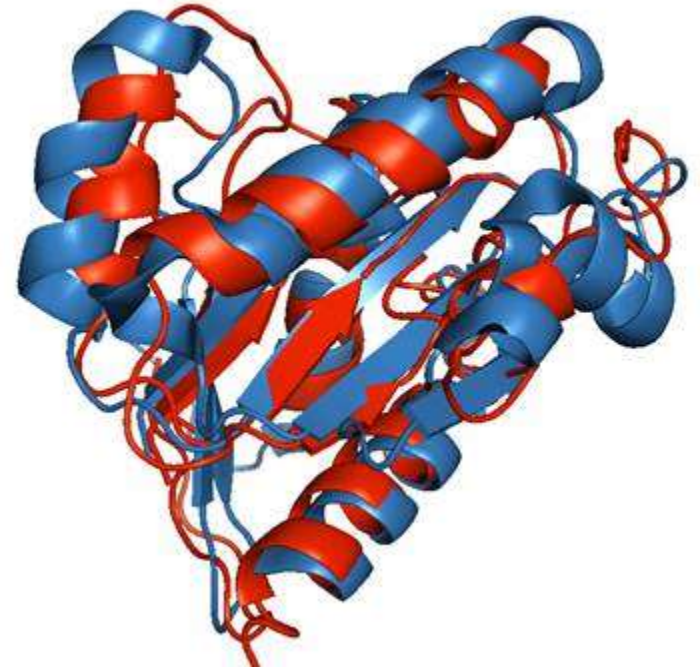
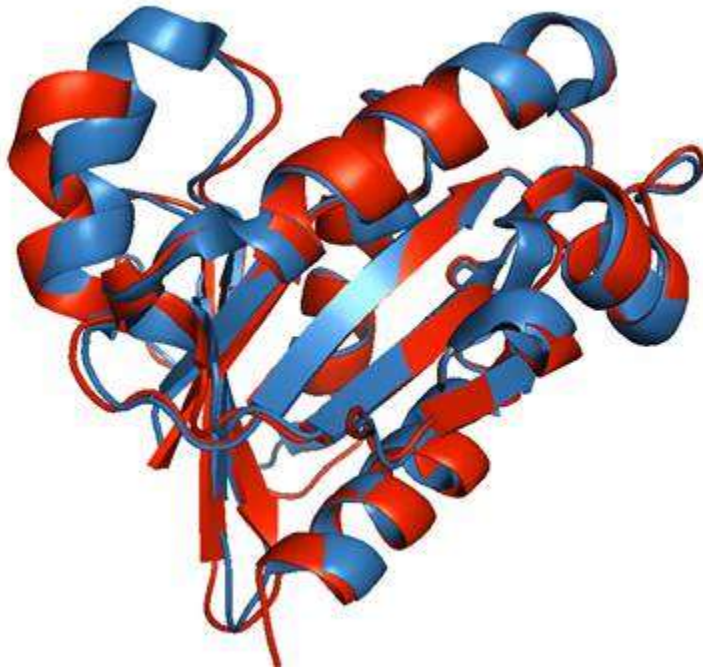


# Comparative Modeling--Basic Protocol

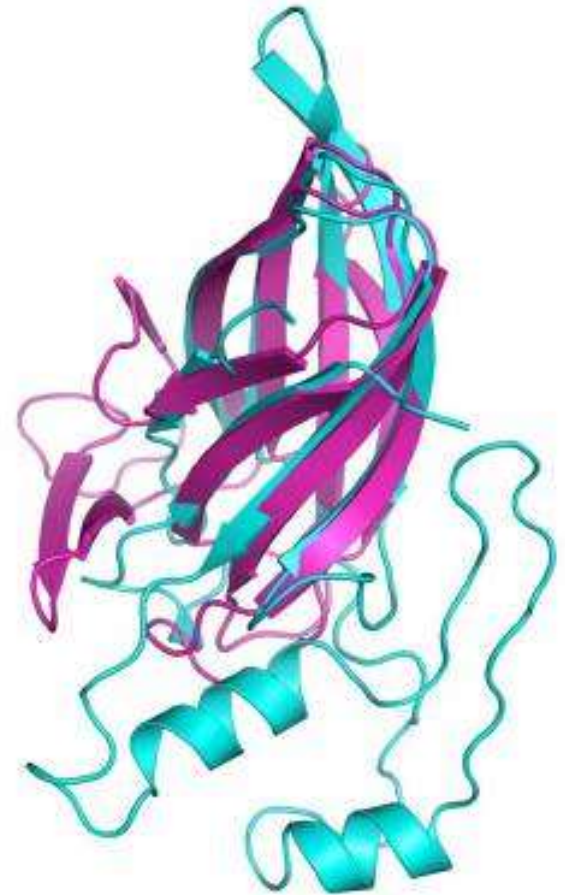
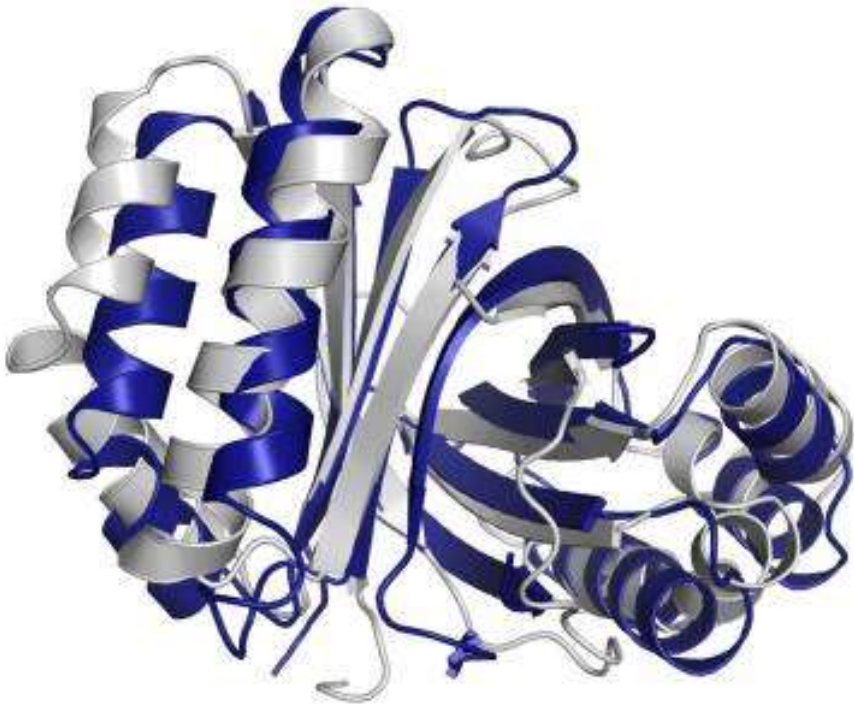
42

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

# Comparative (homology) modeling



# Comparative (homology) modeling



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

A

B

# Comparative (homology) modeling

## MODELLER

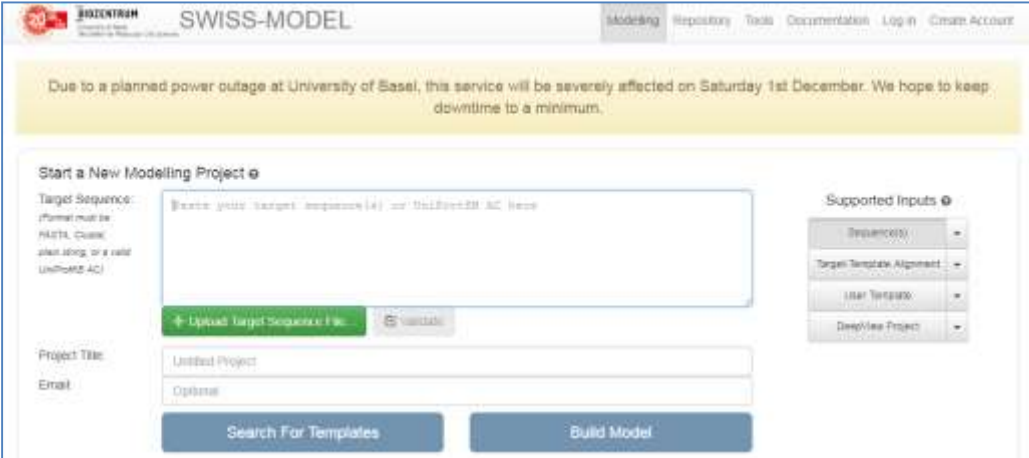
<https://salilab.org/modeller/>



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 buttons for "About MODELLER", "MODELLER News", "Download & Installation", "Release Notes", "Data File Downloads", "Registration", "Biopython Wrapping", "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 links for "Modeling", "Repository", "Tools", "Documentation", "Login", and "Create Account". A yellow banner at the top of the main content area contains a notice: "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, the main heading is "Start a New Modelling Project". The interface includes a "Target Sequence" input field with a placeholder text: "Enter your target sequence(s) or UniProtKB AC here". Below this field are two buttons: "Upload Target Sequence File" and "Paste". To the right of the input field is a "Supported Inputs" section with a dropdown menu showing options: "Sequences", "Target Template Alignment", "User Template", and "Download Project". Below the input field and buttons are two more input fields: "Project Title" (with the text "Untitled Project") and "Email" (with the text "Optional"). At the bottom of the form are two large buttons: "Search For Templates" and "Build Model".



# MODELLER (Sali)

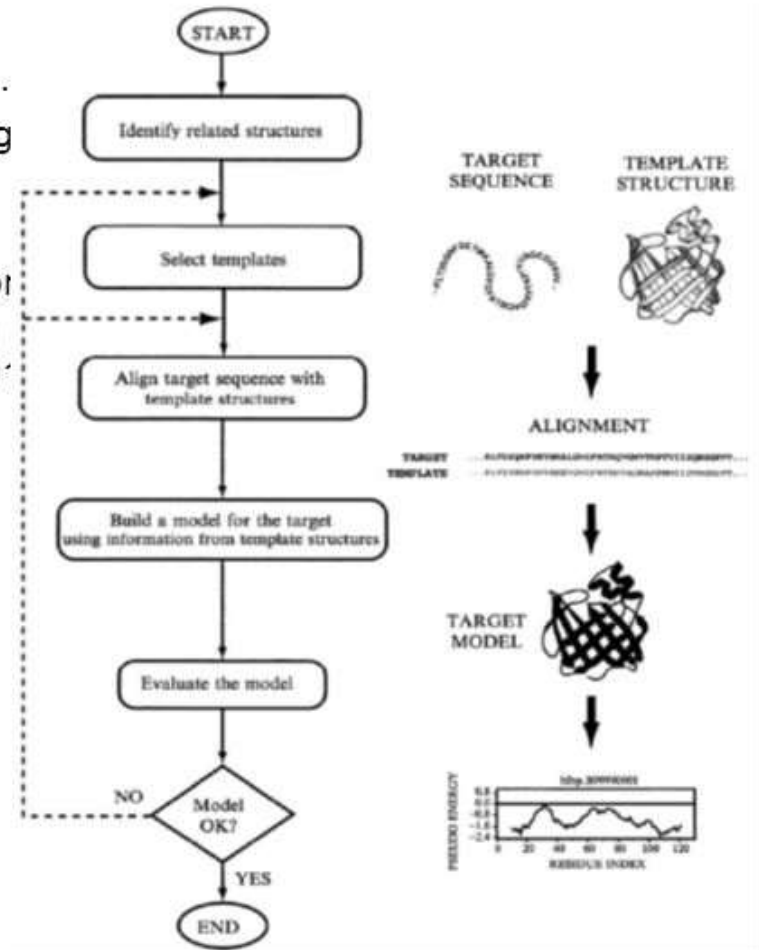
- references

- 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

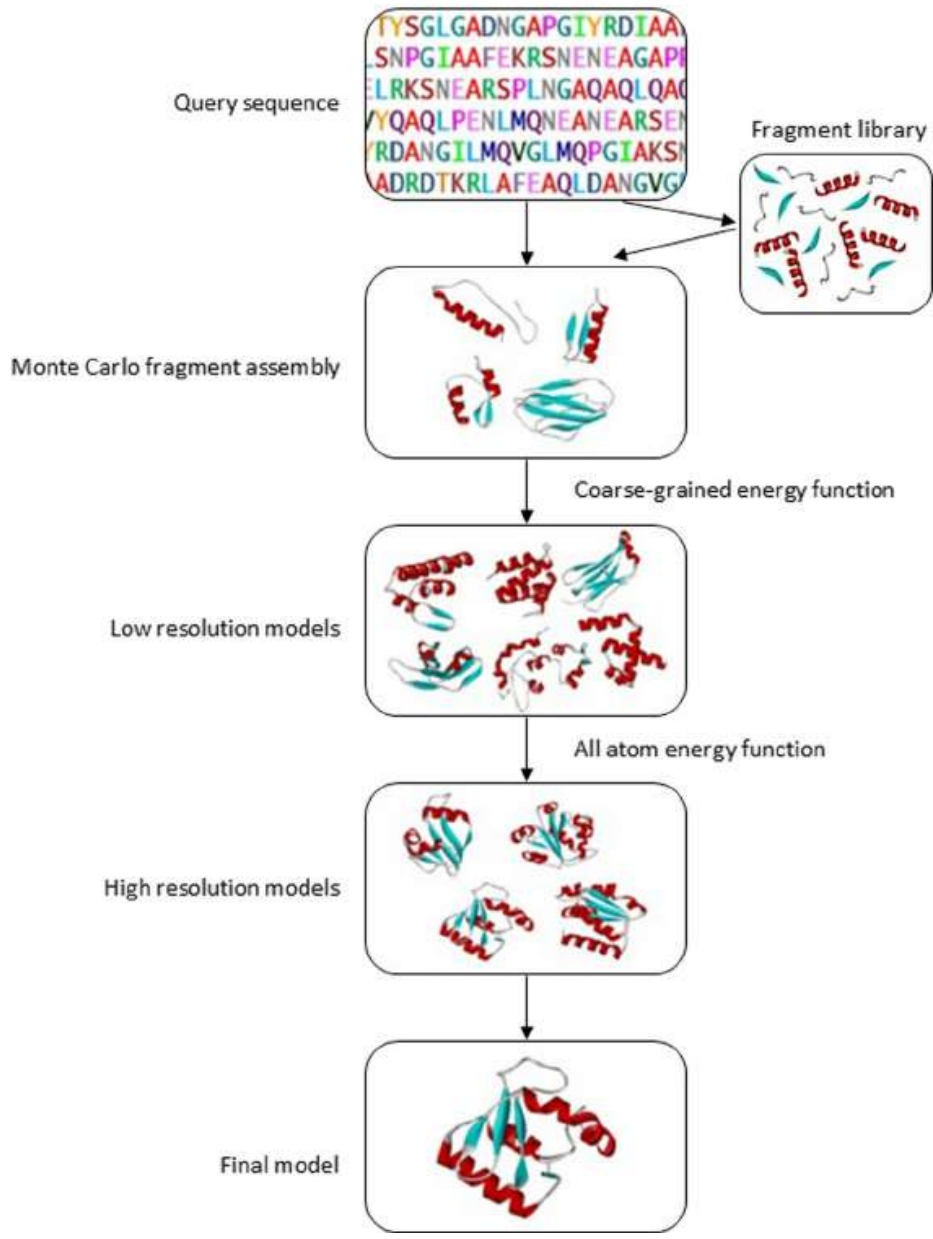
- evaluation:

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



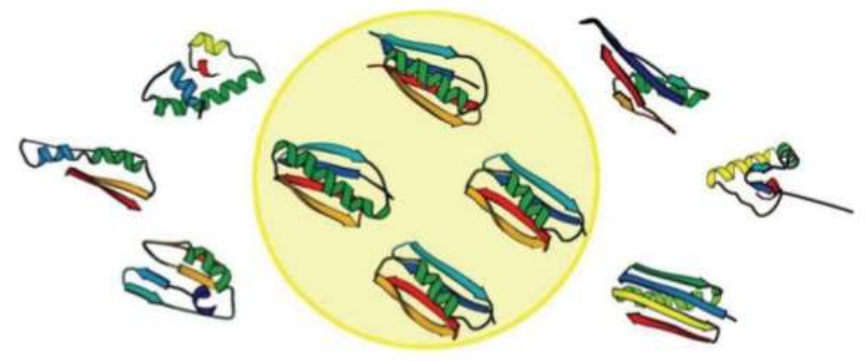
# Rosetta

## Dawid Baker



Rosetta protein modeling consists of sampling and scoring

15



# Rosetta in CASP4

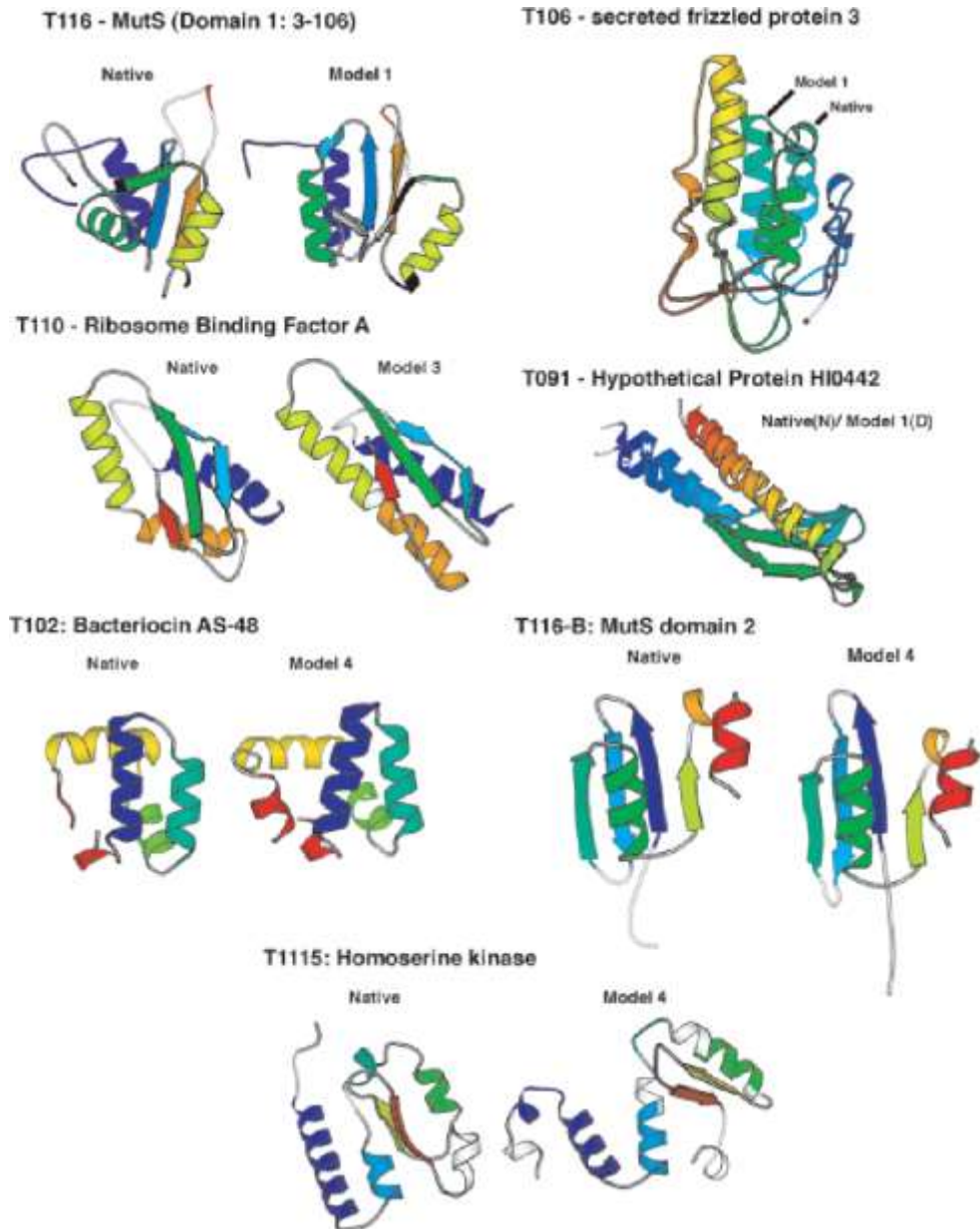


FIG. 3. Comparison of predicted and native structures. Corresponding sequence regions are colored

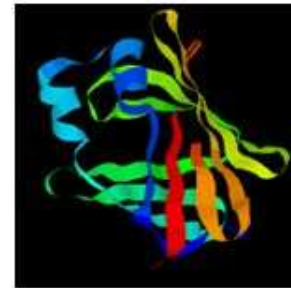
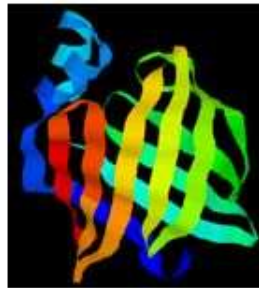
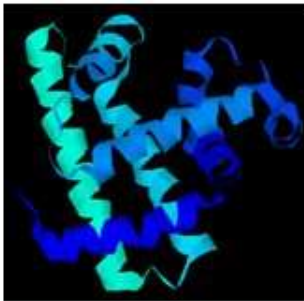
# Concept of Threading

- 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



---

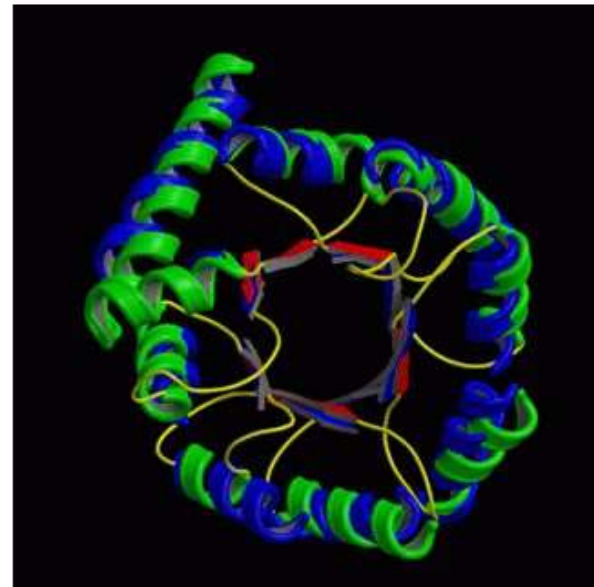
# Protein threading

**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

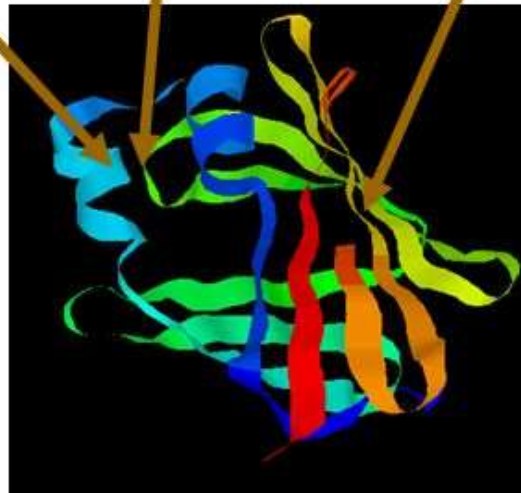
---

# Protein Threading – energy function

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

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

alignment gap  
penalty:  $E_g$



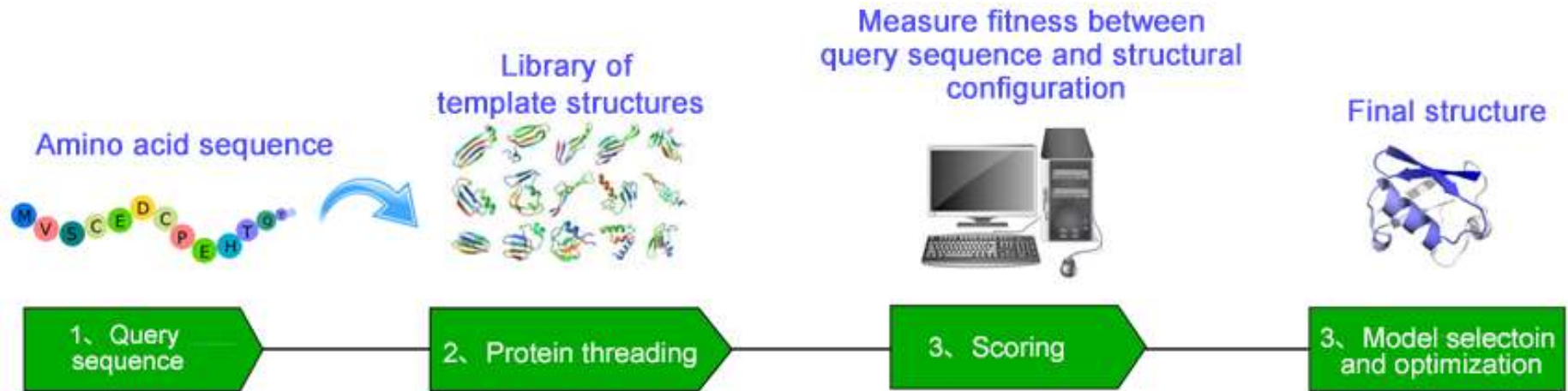
how well a residue fits  
a structural  
environment:  $E_s$

total energy:  $E_p + E_s + E_g$

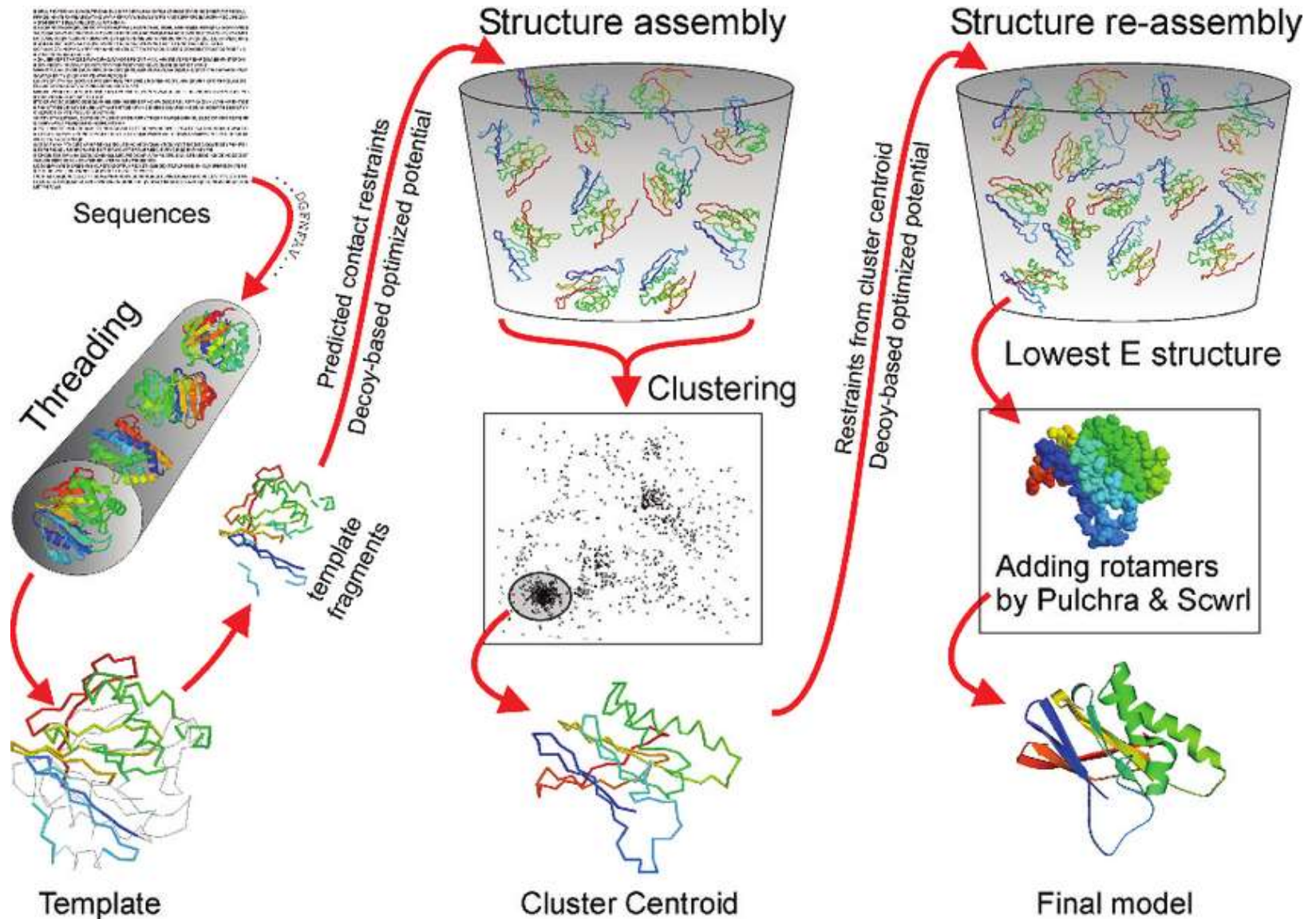
**find a sequence-structure alignment  
to minimize the energy function**

# Comparative (homology) modeling

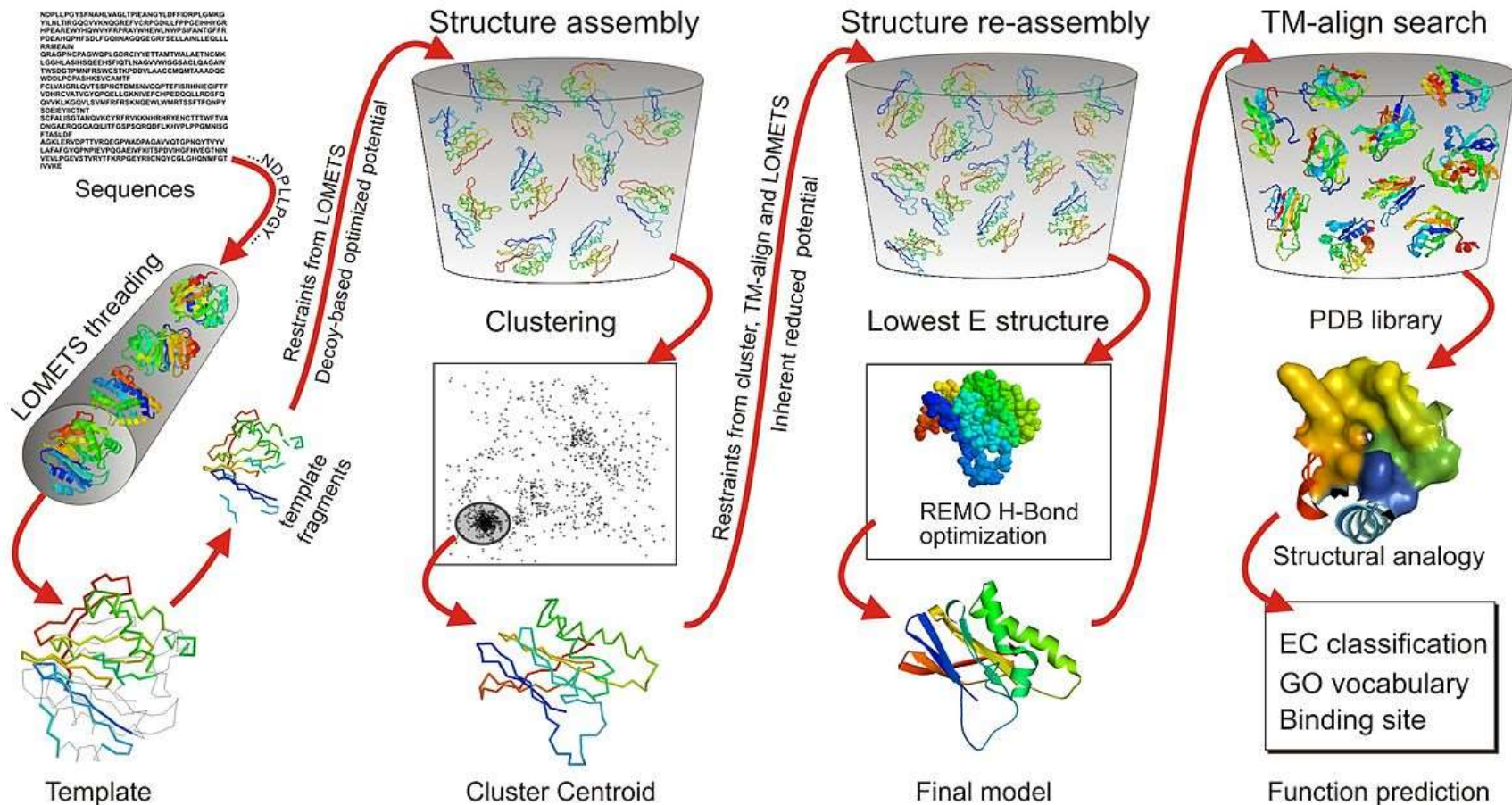
## Threading instead of sequence alignment



# I-TASSER (Y. Zhang)





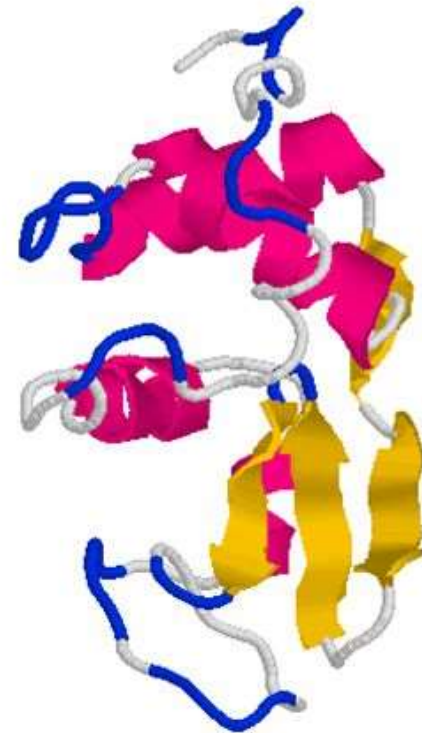


# Protein Folding Problem

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

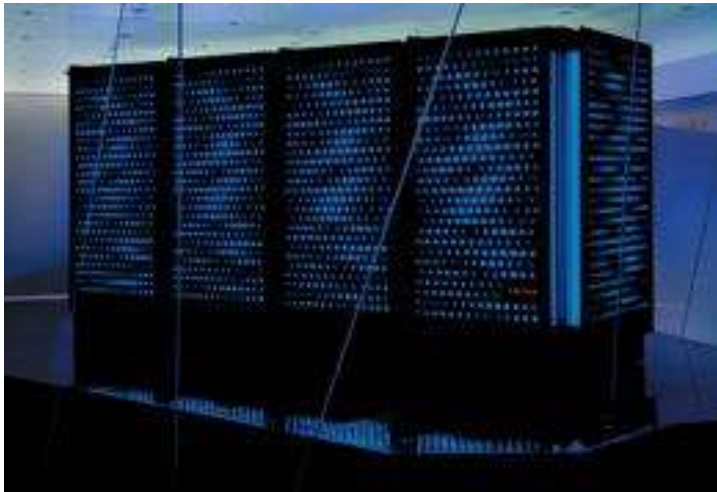
## Lysozyme sequence:

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



Anfinsen, 1960: denatured proteins can refold to active enzymes

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

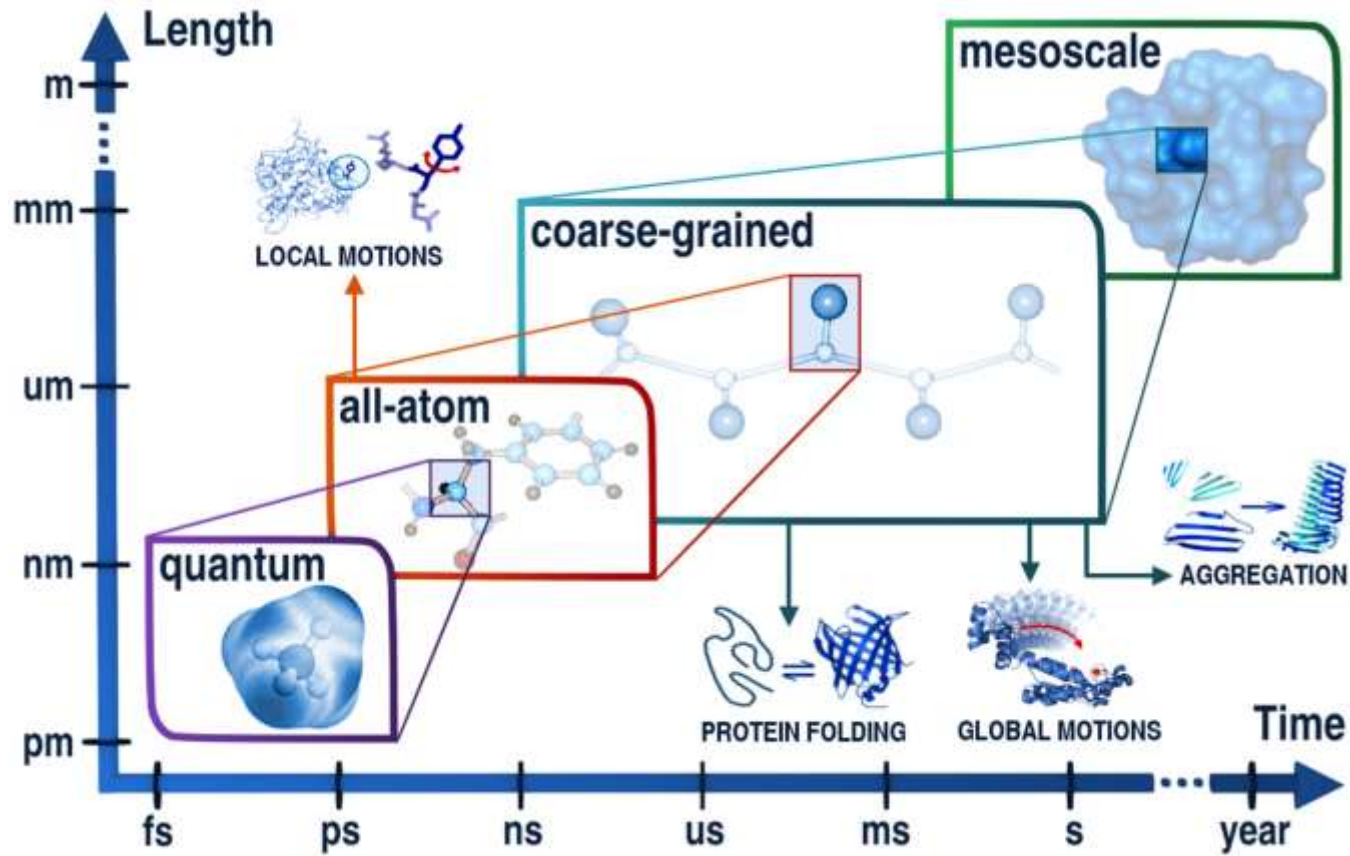


Anton  
David E. Shaw Research

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

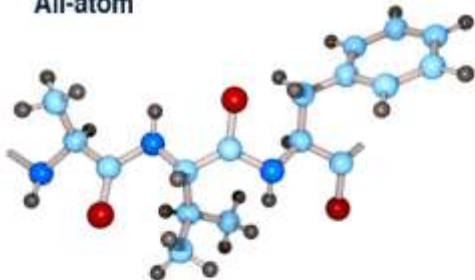
For realistic modeling of larger  
biomolecular systems, including  
flexible protein-protein docking, **we  
need much faster simulations.**

# How to solve the Holy Grail problem

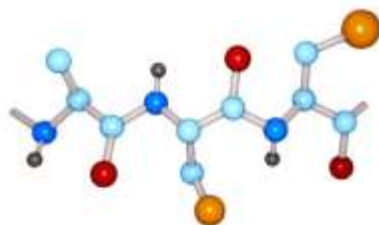


# How to solve the Holy Grail problem – Multiscale Modeling

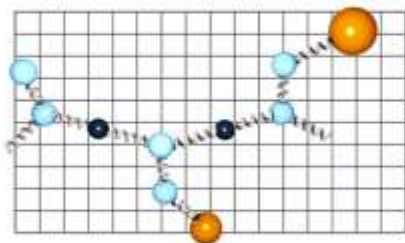
All-atom



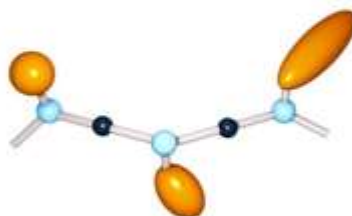
Rosetta CEN



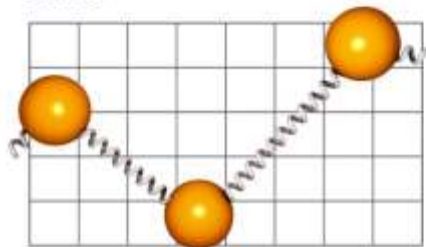
CABS



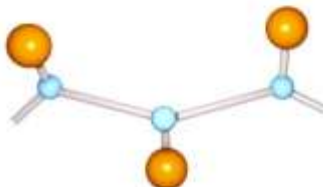
UNRES



SICHO




Levitt-Warshel




Nobelpriset 2013


The Nobel Prize in Chemistry 2013



**Martin Karplus**  
Université de Strasbourg,  
France and Harvard  
University, Cambridge,  
MA, USA



**Michael Levitt**  
Stanford University School of  
Medicine, CA, USA



**Arieh Warshel**  
University of Southern  
California, Los Angeles, CA,  
USA

for "the development of multiscale models for complex chemical systems"

# CABS model

$C_{\alpha}$ - $C_{\beta}$ -Side chain

High-coordination lattice

Statistical force-field

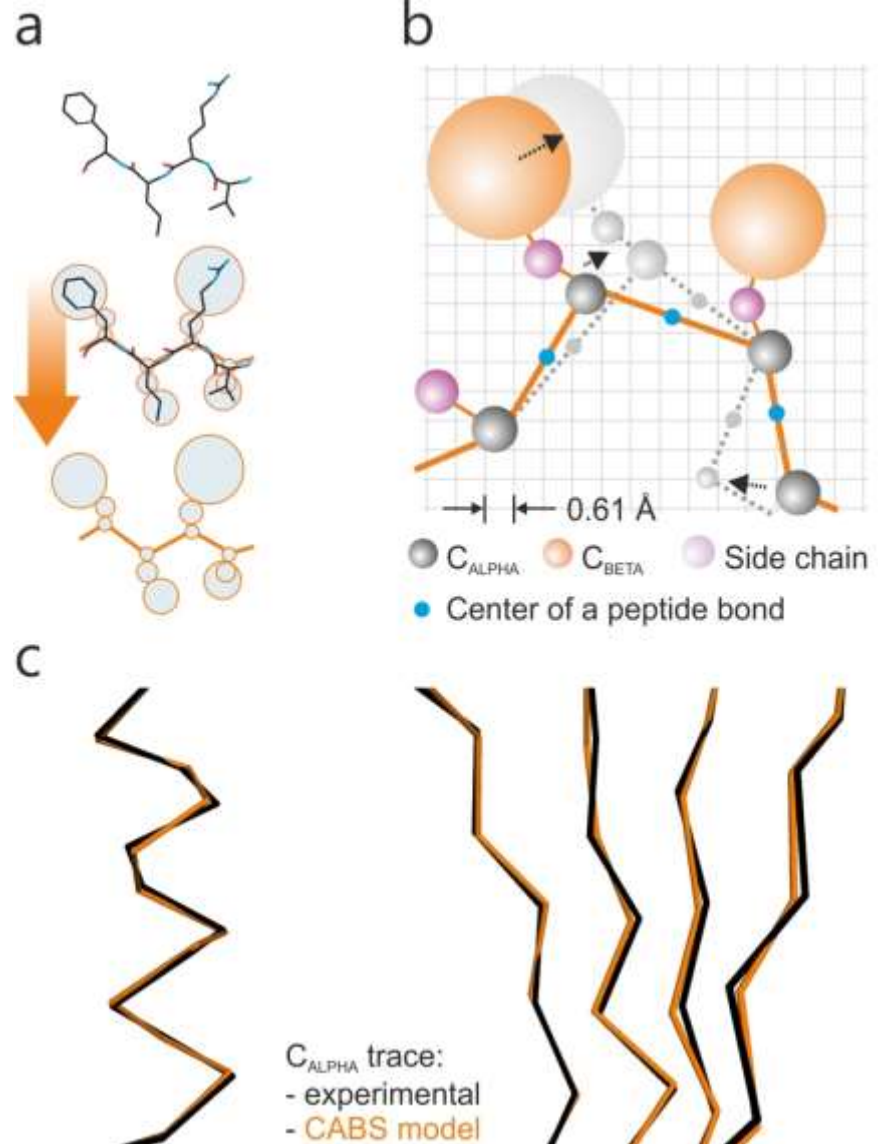
Monte Carlo dynamics

Figures:

a) Building reduced model

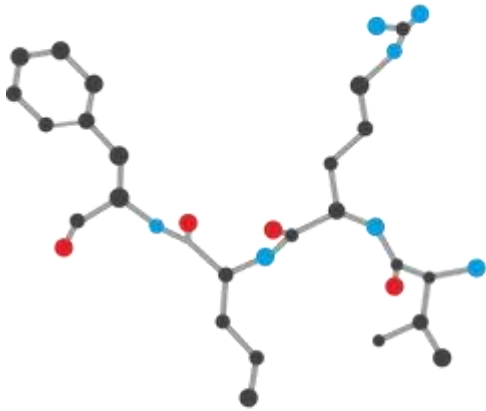
b) MC moves on the high-coordination lattice

c) Accuracy ( $C_{\alpha}$ -traces)



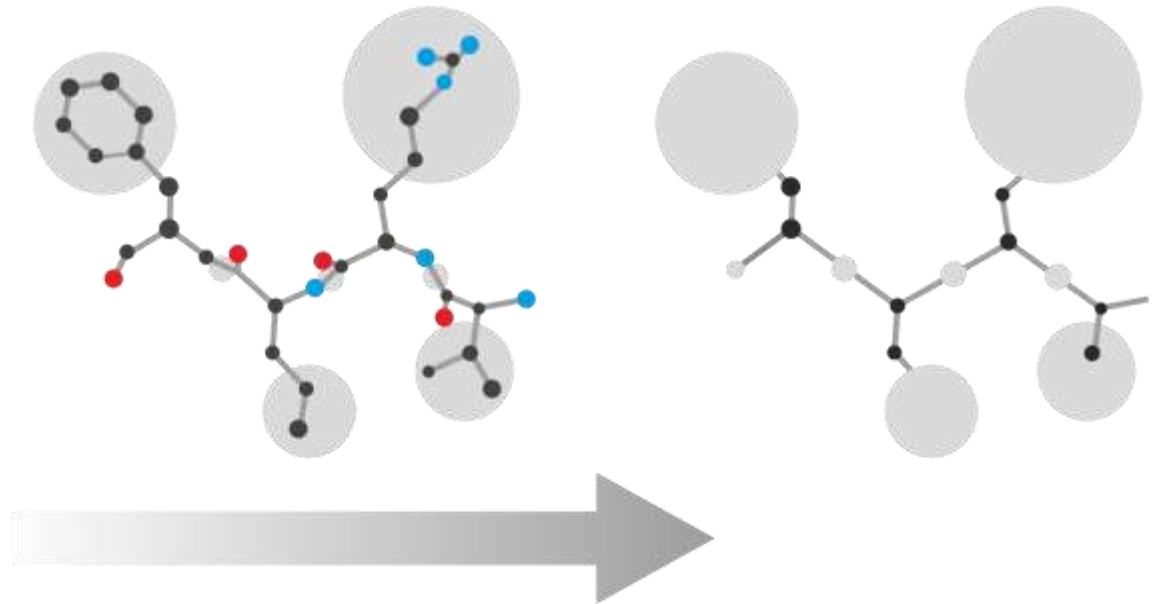
# Time scales MD vs. CABS

All-atom molecular dynamics (MD)



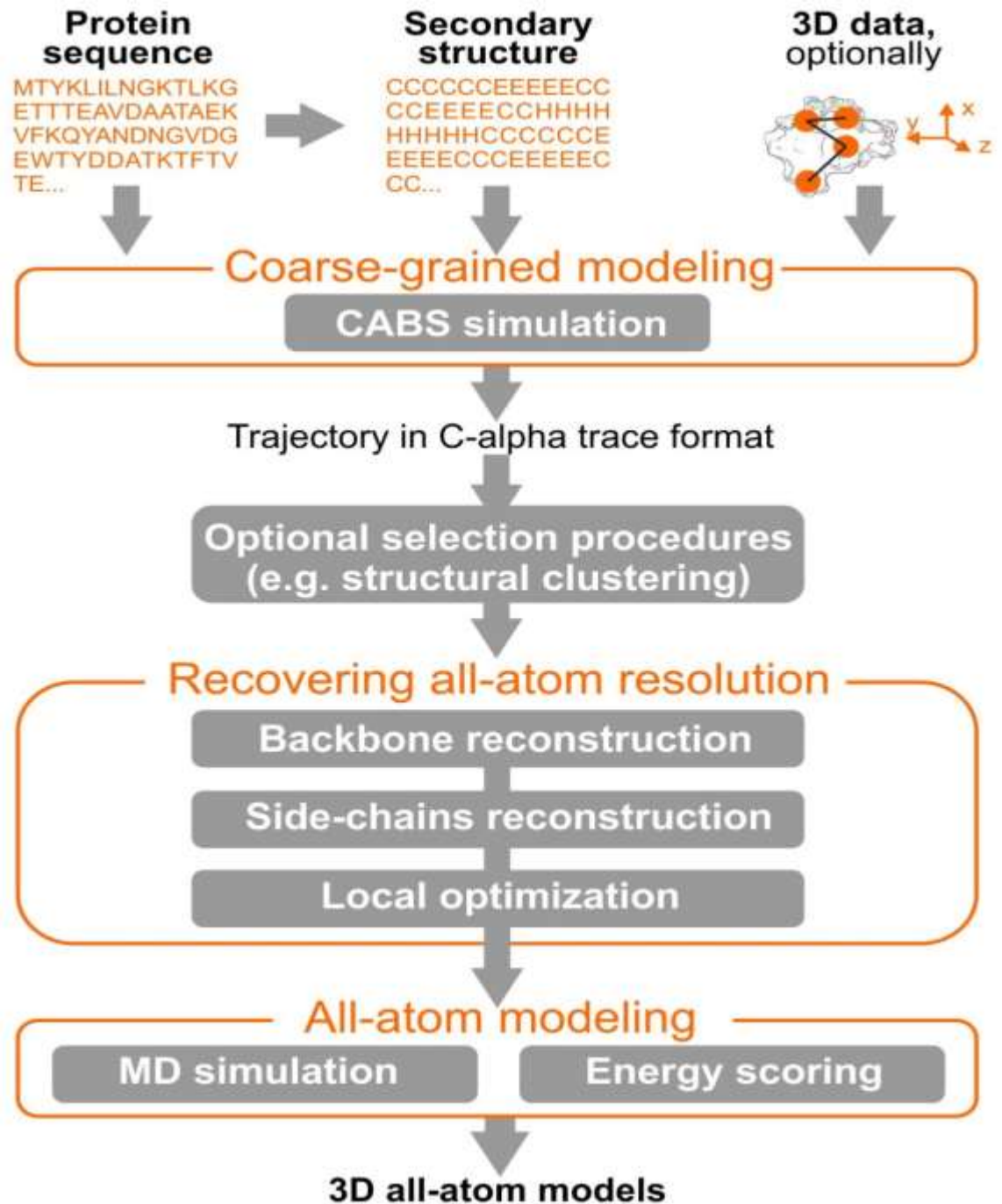
Max ~ 1 millisecond

CABS Monte Carlo dynamics



~  $10^3 / 10^4$  faster

# Multiscale modeling with CABS



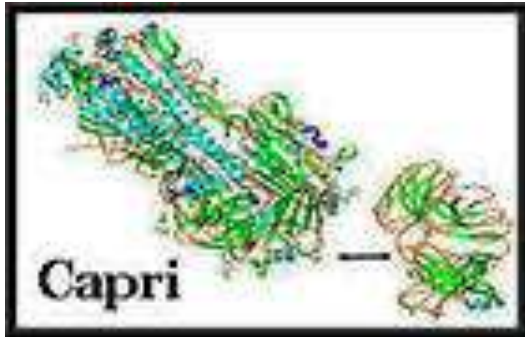
## APPLICATINS

Structure prediction

Protein dynamics

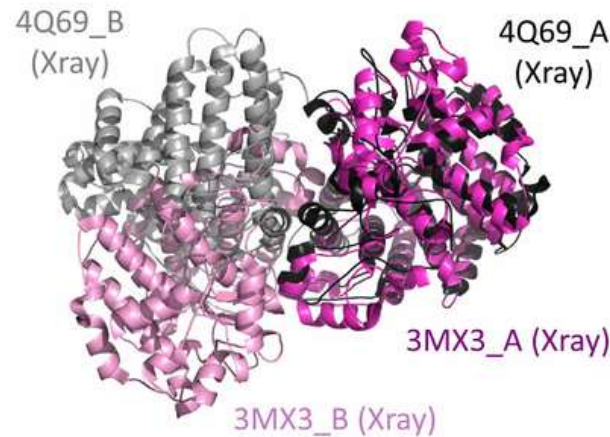
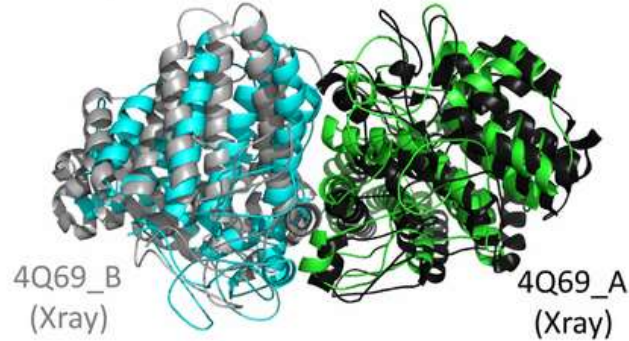
Protein docking



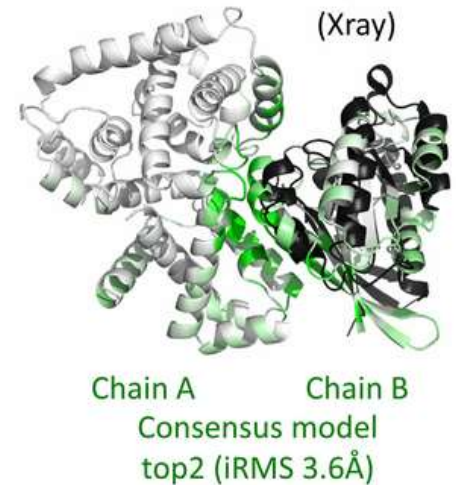


# CAPRI: Critical Assessment of PRediction of Interactions

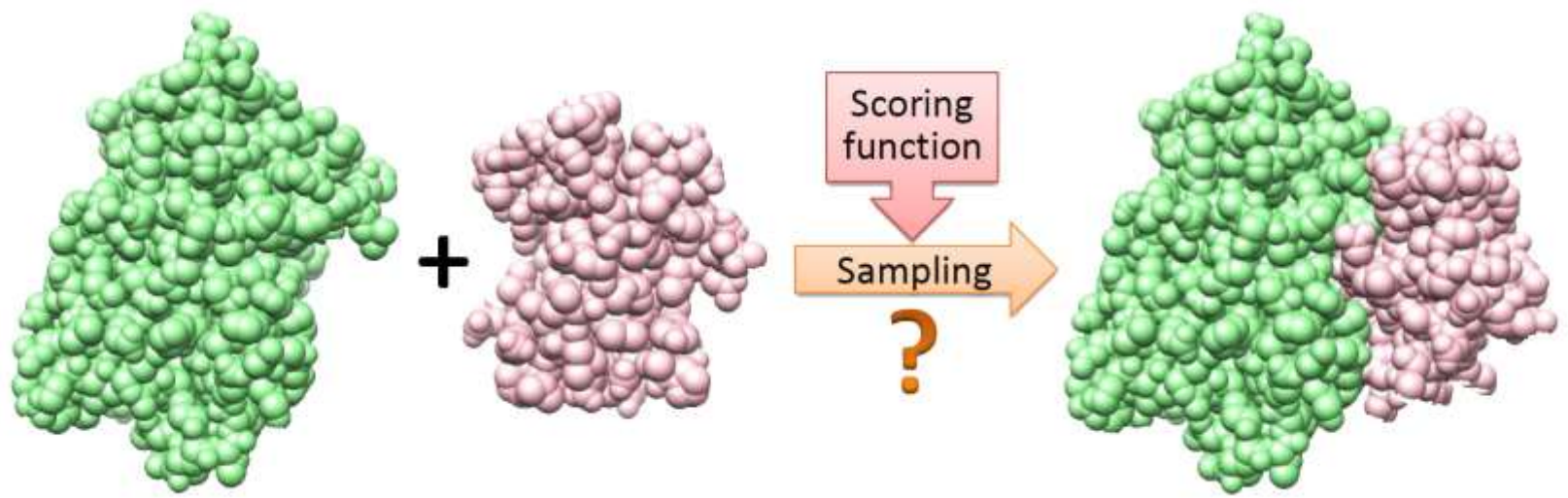
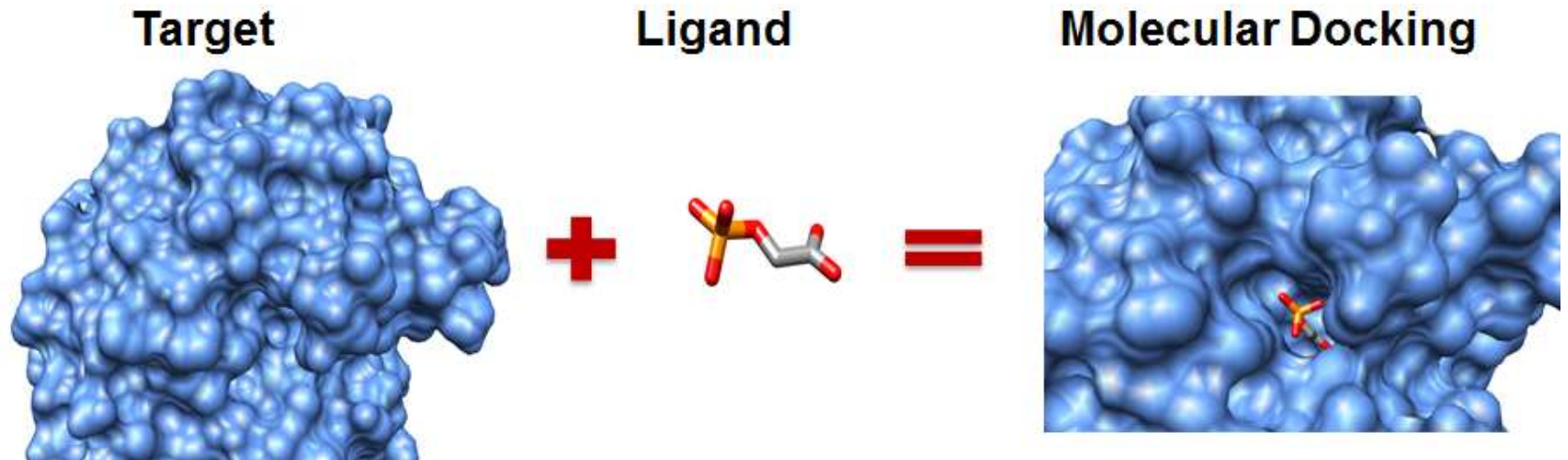
A CAPRI T72 Consensus Top5 Chain B CAPRI T72 Consensus Top5 Chain A



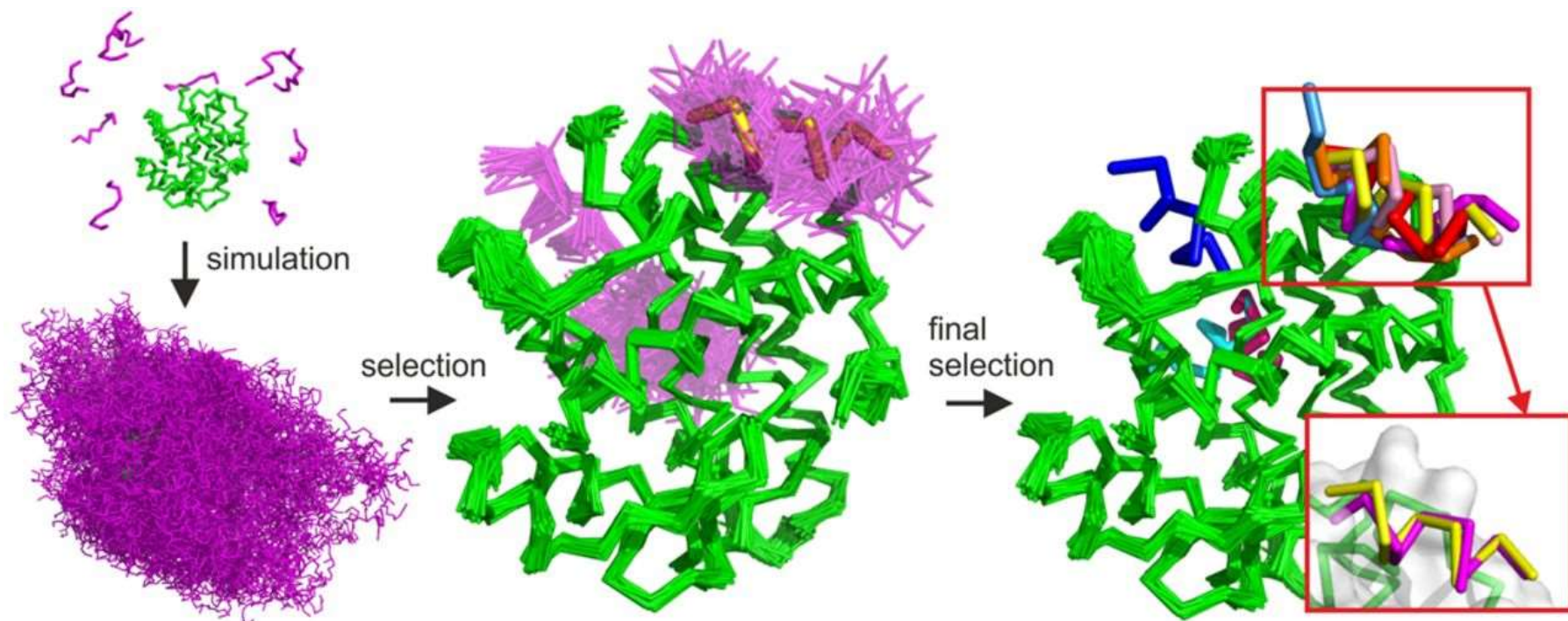
B 2G77 (Xray)



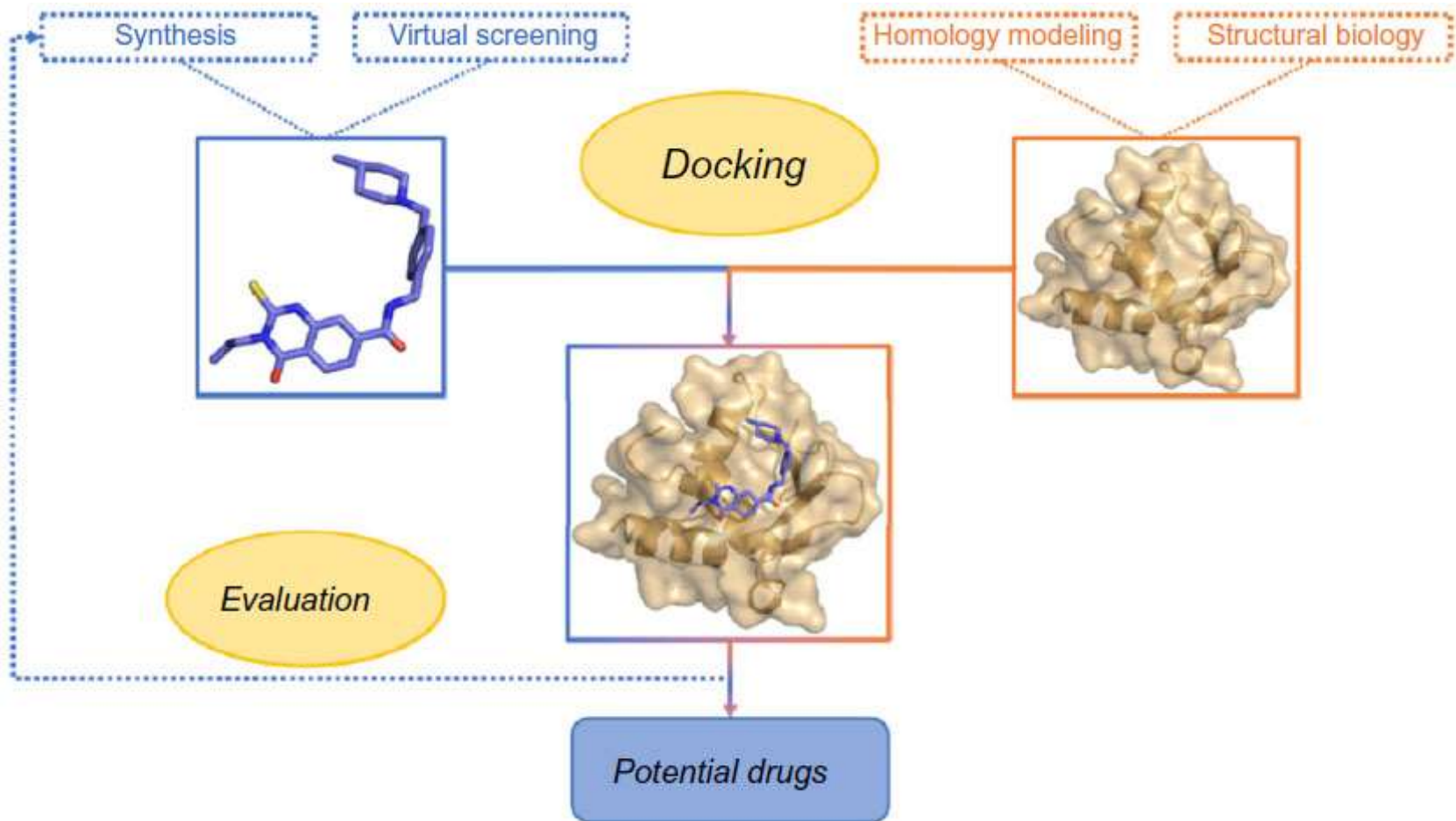
# Molecular docking



# Peptide docking with CABS model



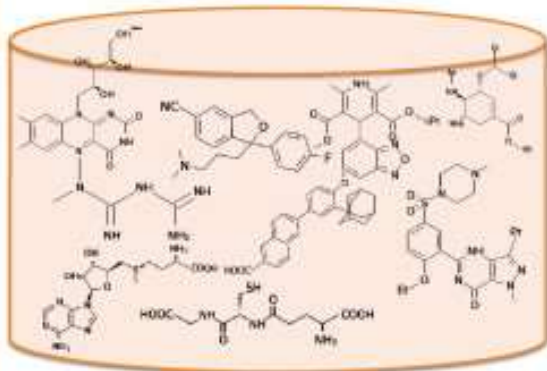
M. Kurcinski, M. Jamroz, M. Blaszczyk, A. Kolinski & S. Kmiecik, “CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site”, *Nucleic Acids Research*, 2015



**(A) Docking**



Protein of interest



Chemical database

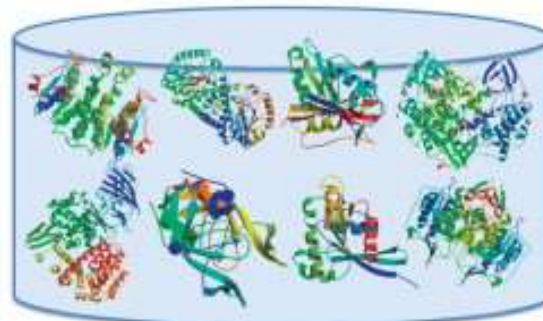


Possible binding ligand

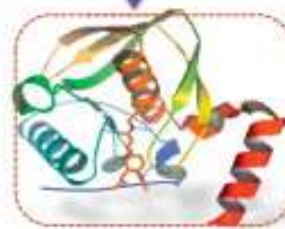
**(B) Reverse docking**



Active compound or existing drug



Protein target database



Possible binding protein