LABORATORY of THEORY of BIOPOLYMERS

Proteins - structural bioinformatics (3)

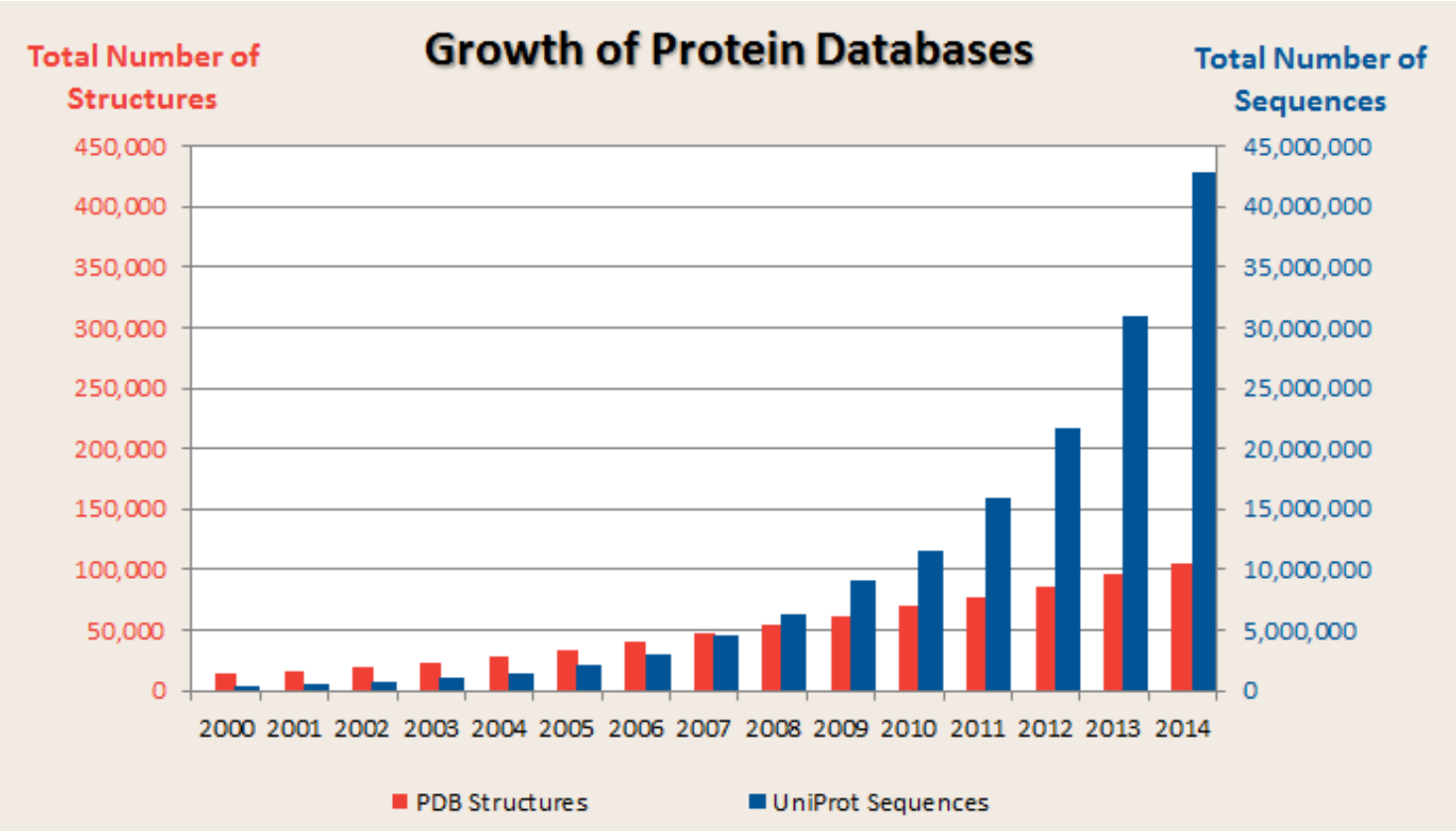
Comparative modeling (homology modeling)

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

How many proteins?

- Just 150 AA protein - 10^{195} sequences
- Eukaryotic protein universe ~ 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)

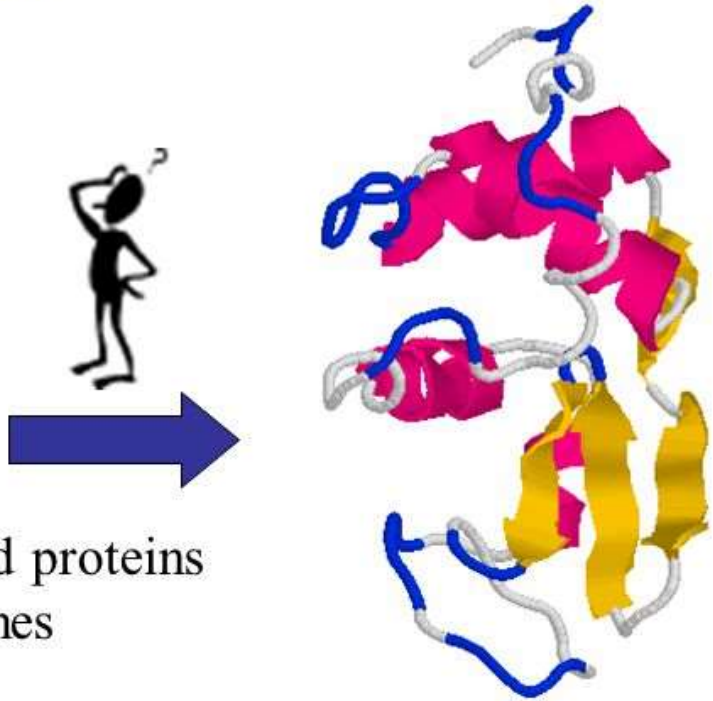
  - comparative modeling (Most of newly identified protein structures are similar to already known)

Protein Folding Problem

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

Lysozyme sequence:

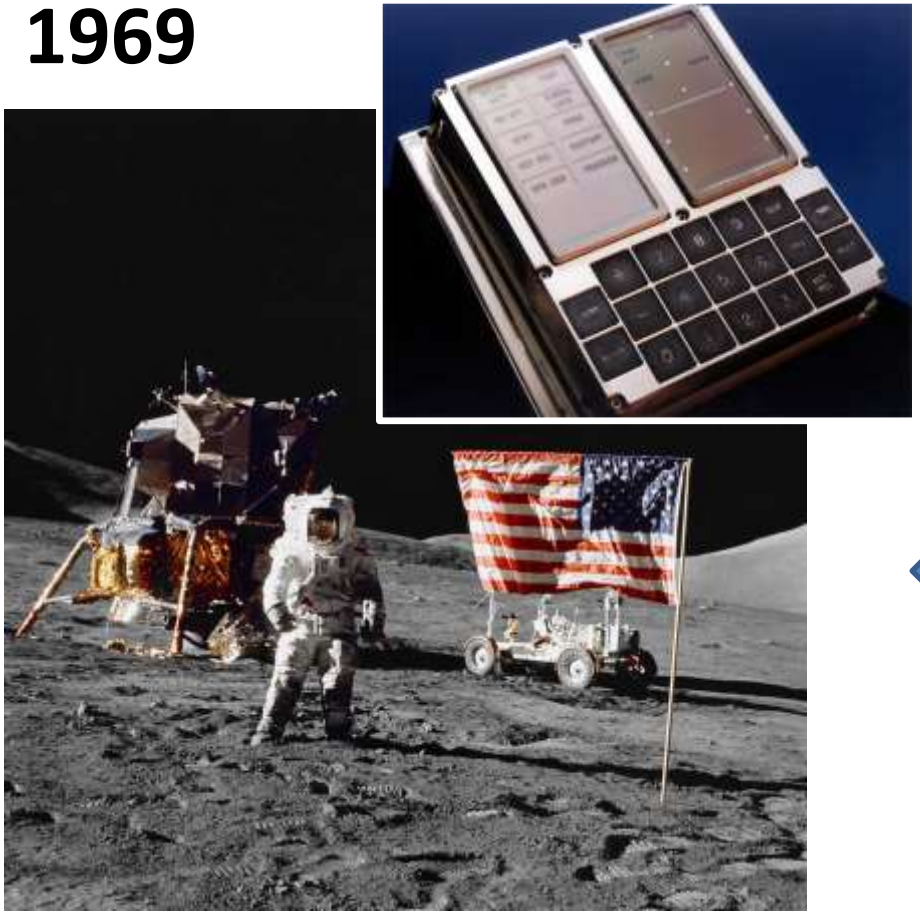
```
KVFGRCELAA AMKRHGLDNY
RGYSLGNWVC AAKFESNFNT
QATNRNTDGS TDYGILQINS
RWWCNDGRTP GSRNLCNIPC
SALLSSDITA SVNCAKKIVS
DGNGMNAWVA WRNRCKGTDV
QAWIRGCRL
```



Anfinsen, 1960: denatured proteins can refold to active enzymes

Computing power

1969

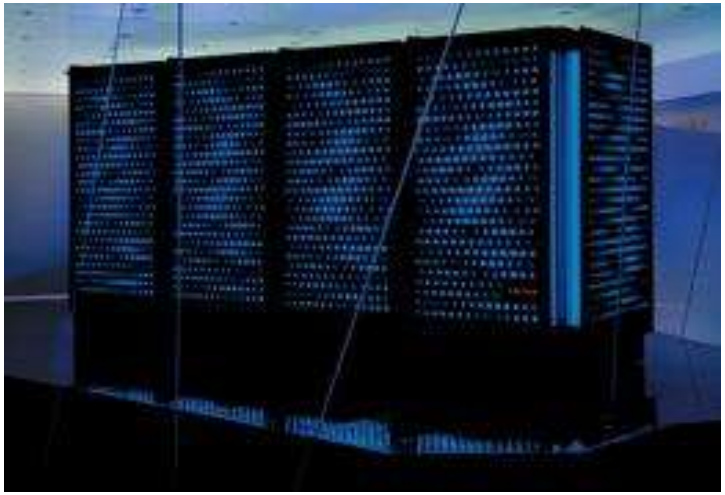


2018

APOLLO MISSION | **smartphone**

120,000,000 / 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.

29

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

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 (copmpartive and de novo) by David Baker and co-workers
- Refined allignments and Coarse-Grained modeling using CABS tools by Janusz Bujnicki and Andrzej Kolinski
- Sofisticated ranking of allignments and fragment modeling using CAS (a version of CABS) by Yang Zhang
- Computer deep-learning and fragment assembly (Rosetta) Lee Sedol

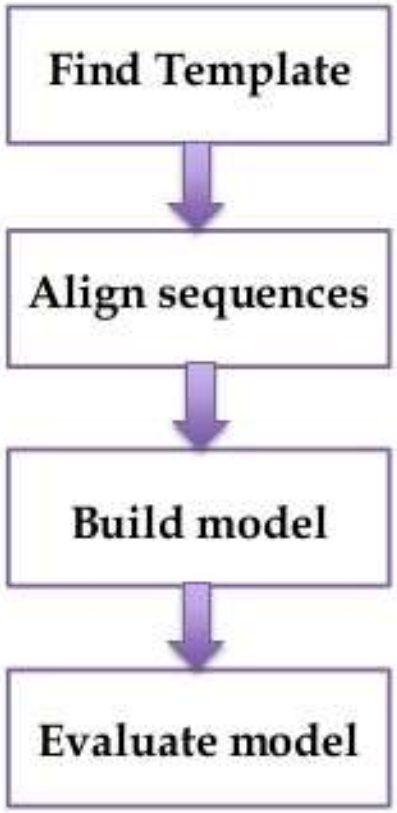
CASP6: Average scoring, all categories

(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



Are there any well characterized proteins similar to my protein?

What is the position-by-position target/template equivalence?

What is the detailed 3D structure of my protein?

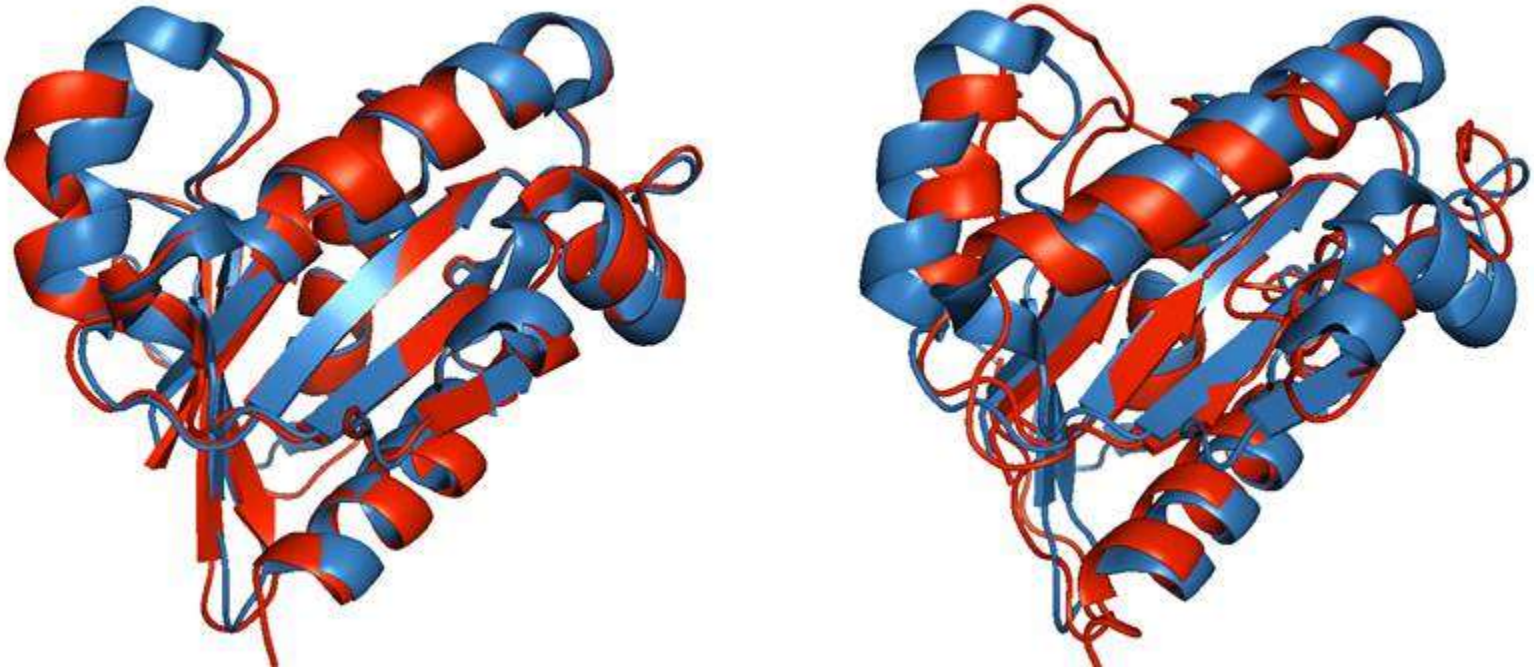
Measure the model quality. Is my model any good?

Comparative Modeling--Basic Protocol

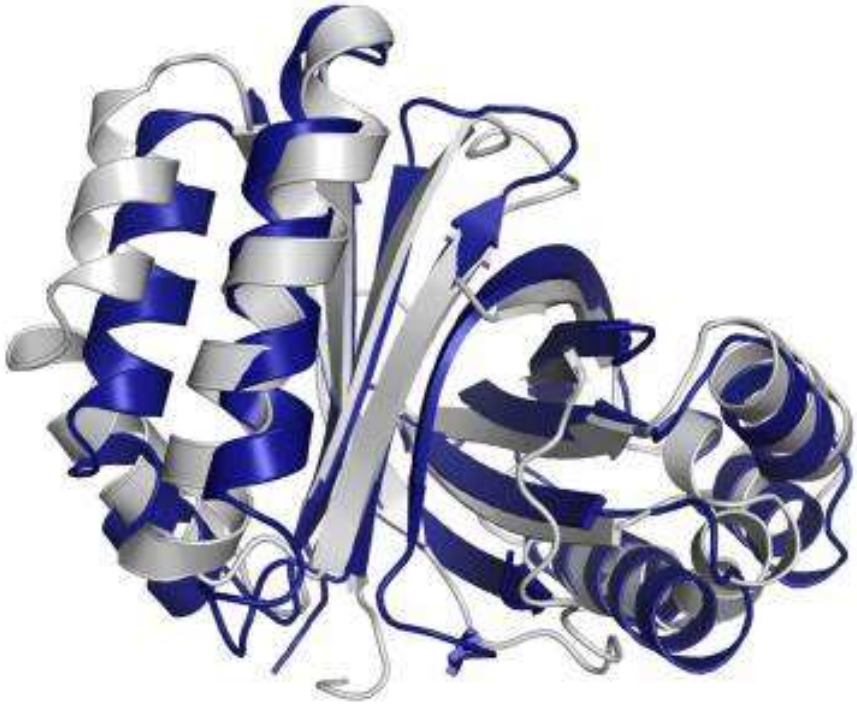
1. Identification of homologue for target sequence
2. Alignment of target sequence to template sequence and structure
3. Side-chain modeling, copy the backbone of the template and model the new side chains onto this backbone
4. Loop modeling, for insertions and deletions in the alignment
5. Refinement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

David F B, Charlotte M D, Hampapathalu A N, Nuria C, An Iterative Structure-Assisted Approach to Sequence Alignment and Comparative Modeling, PROTEINS: Structure, Function, and Genetics Supplementations, 3, pp. 55-60.

Comparative (homology) modeling



Comparative (homology) modeling



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

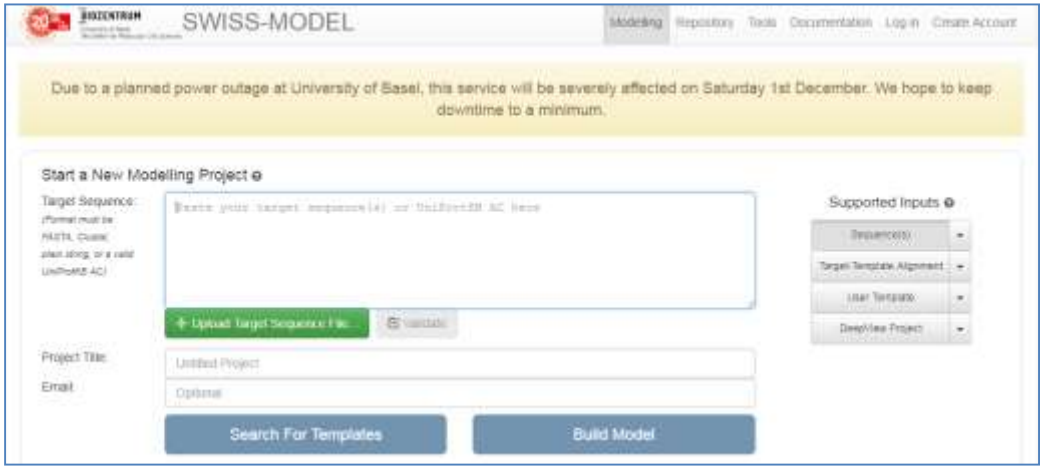
A

B

Comparative (homology) modeling

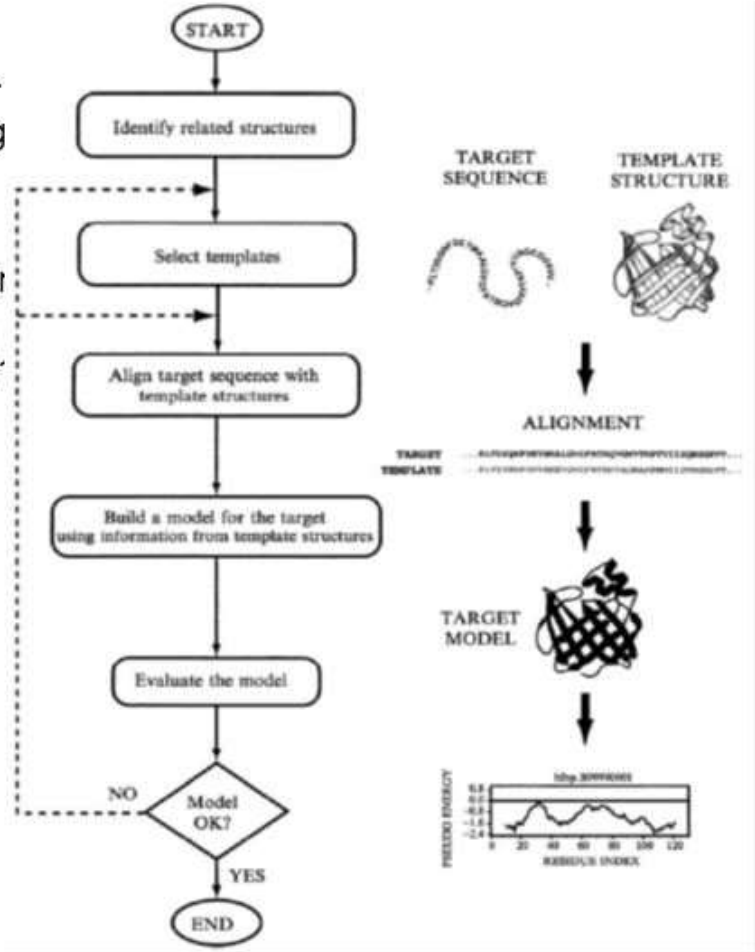
MODELLER
https://salilab.org/modeller/

SWISS-MODEL
https://swissmodel.expasy.org

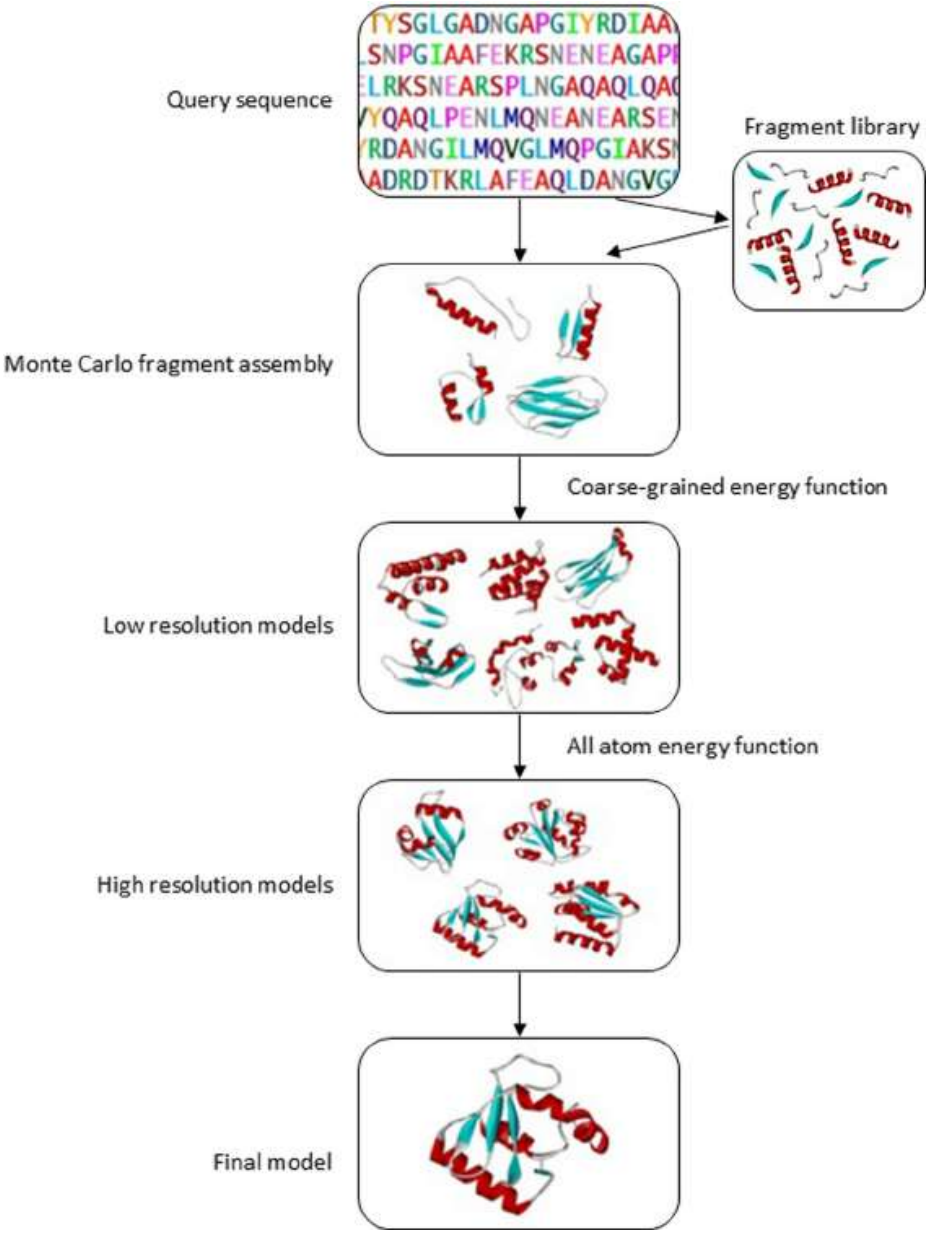


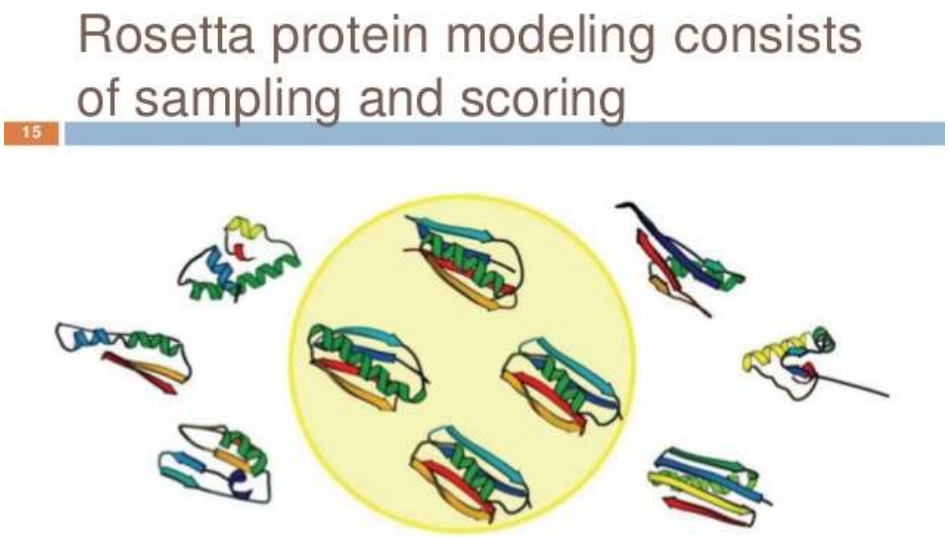
MODELLER (Sali)

- 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. Š ali. Modeling of loops in protein structures. Protein Science 9, 1753-1773, 2000.
  - Fiser A, Sali A. (2003). Modeller: generatio 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 in CASP4

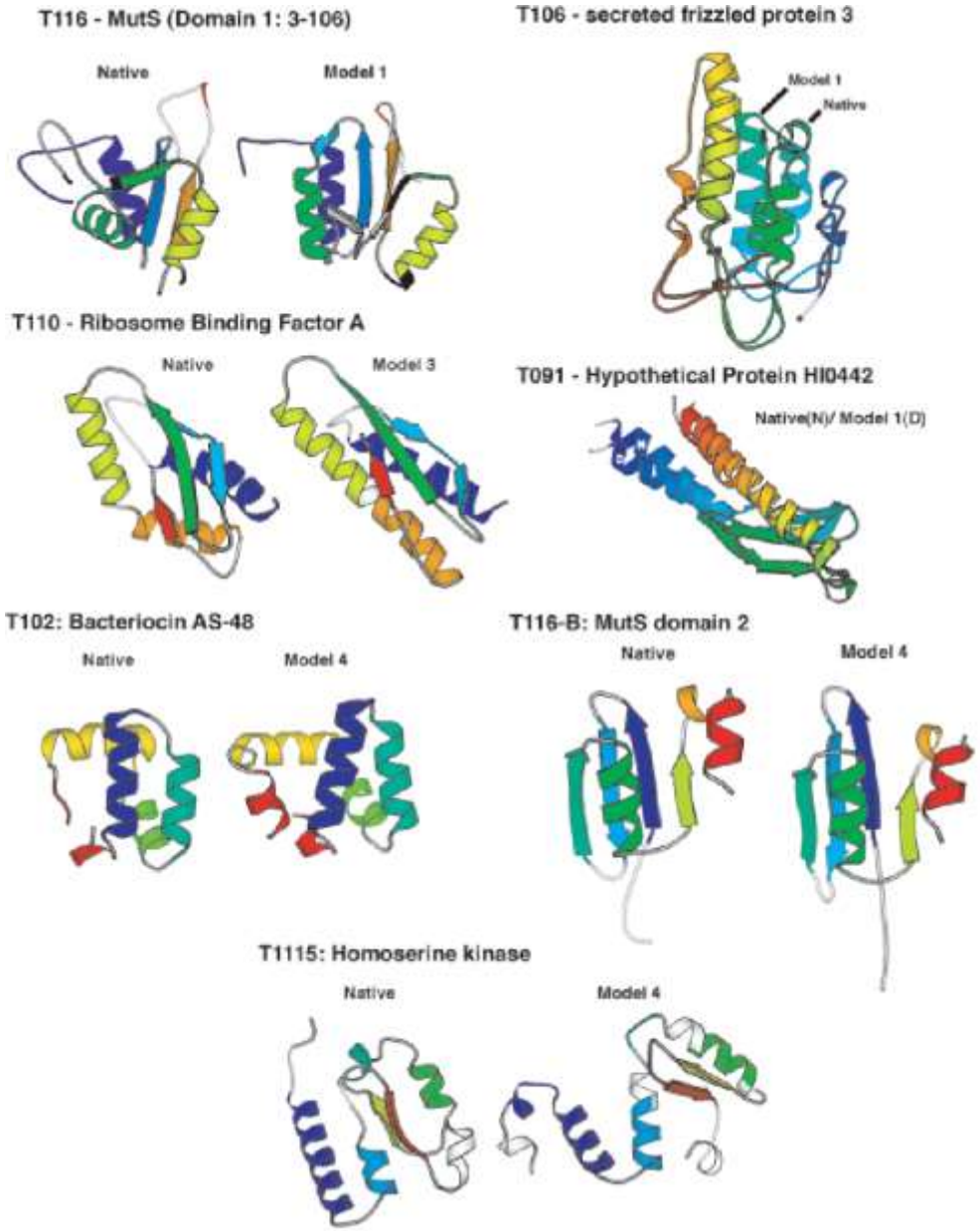


Fig. 3. Comparison of predicted and native structures. Corresponding sequence regions are colored

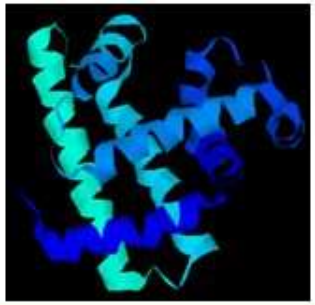
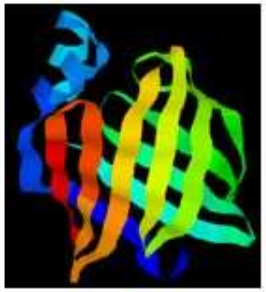
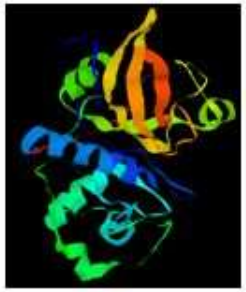
Concept of Threading

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

Query sequence

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

Template set

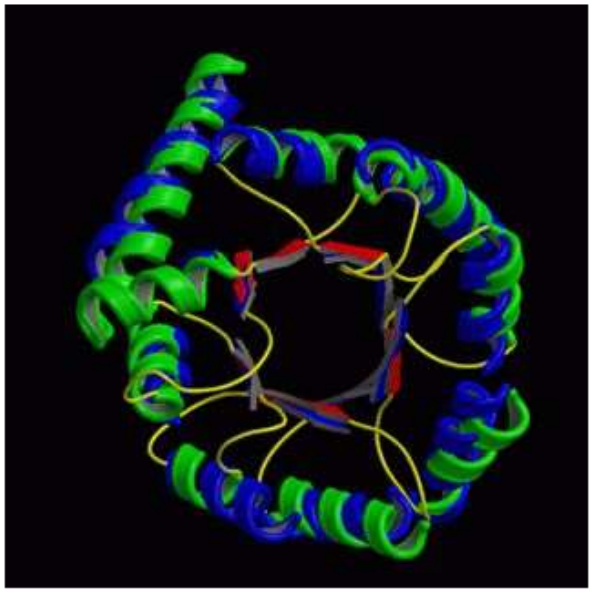

Protein threading

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

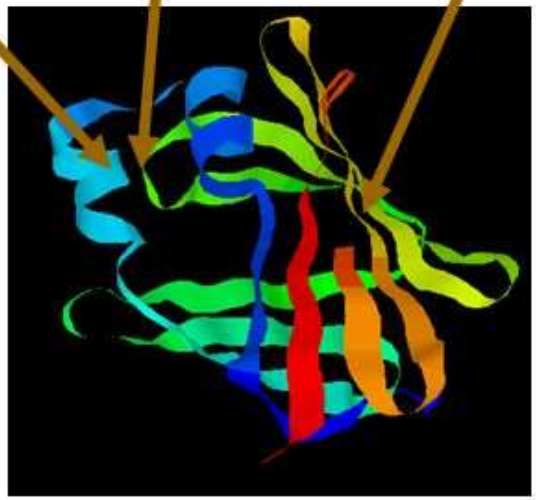


TIM barrel

Protein Threading – energy function

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

how preferable to put two particular residues nearby: E_p



how well a residue fits a structural environment: E_s

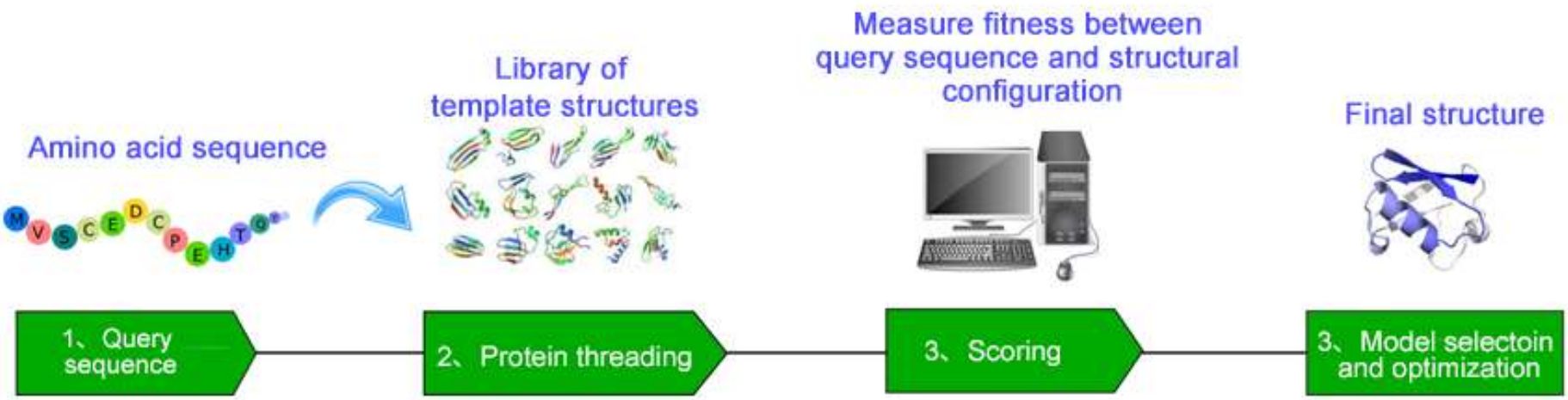
alignment gap penalty: E_g

total energy: E_p + E_s + E_g

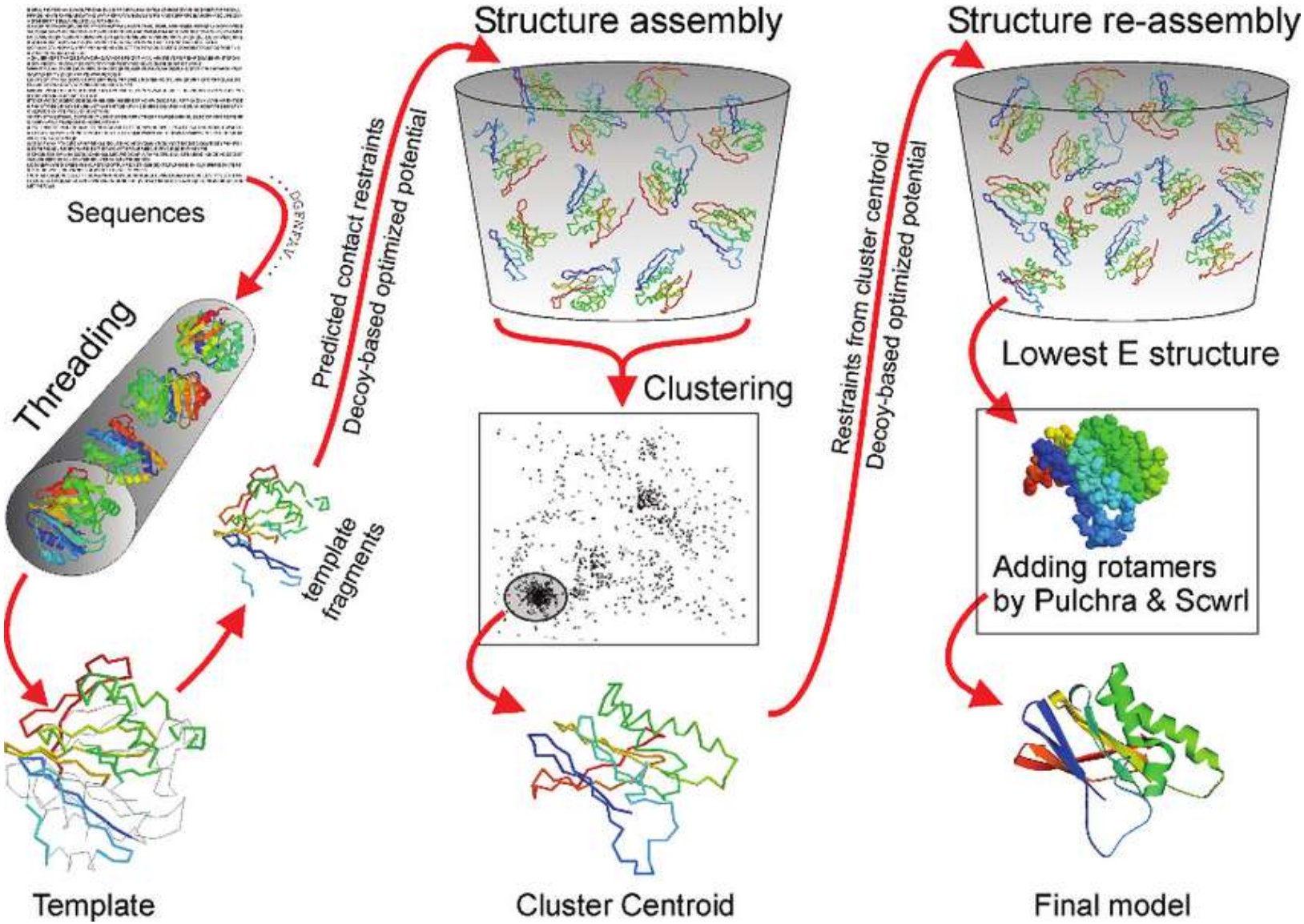
find a sequence-structure alignment to minimize the energy function

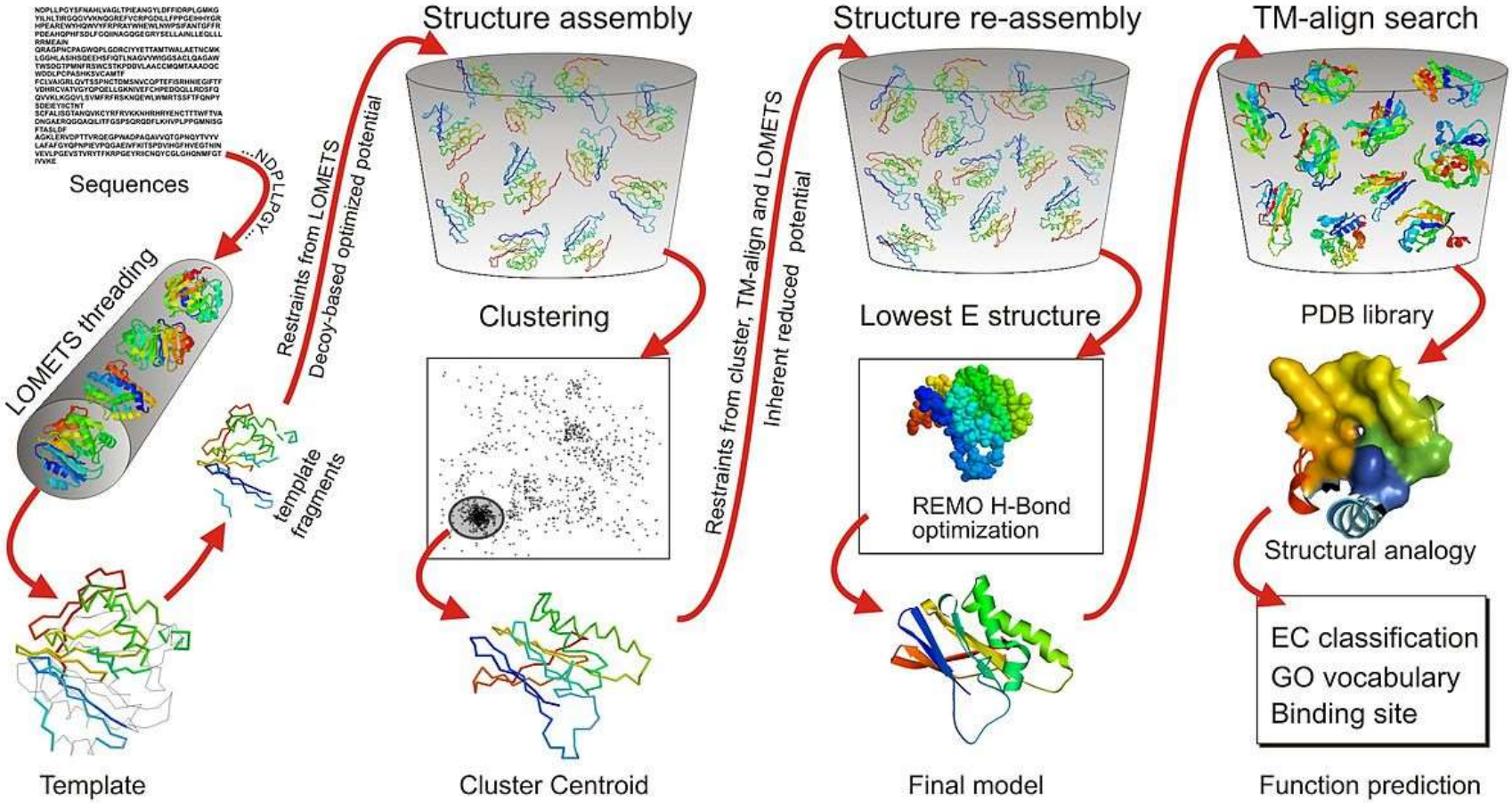
Comparative (homology) modeling

Threading instead of sequence alignment



I-TASSER (Y. Zhang)



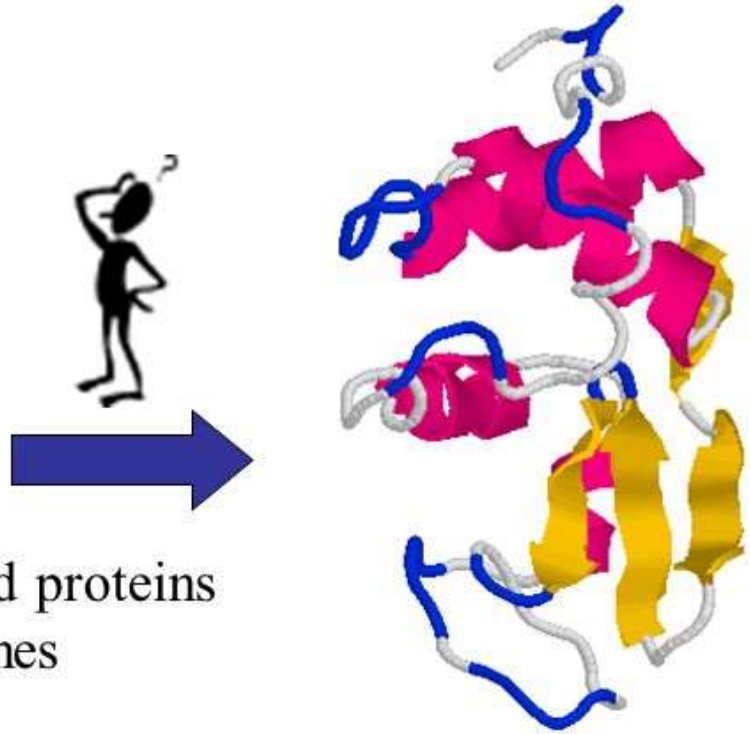

Sequences
LOMETS threading
Template
template fragments
Restraints from LOMETS
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster, TM-align and LOMETS
Inherent reduced potential
Structure re-assembly
Lowest E structure
REMO H-Bond optimization
Final model
TM-align search
PDB library
Structural analogy
EC classification
GO vocabulary
Binding site
Function prediction

Protein Folding Problem

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

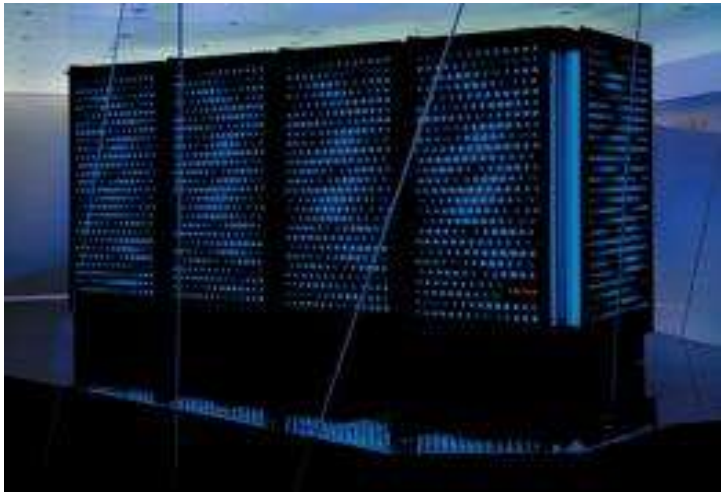
Lysozyme sequence:

```
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

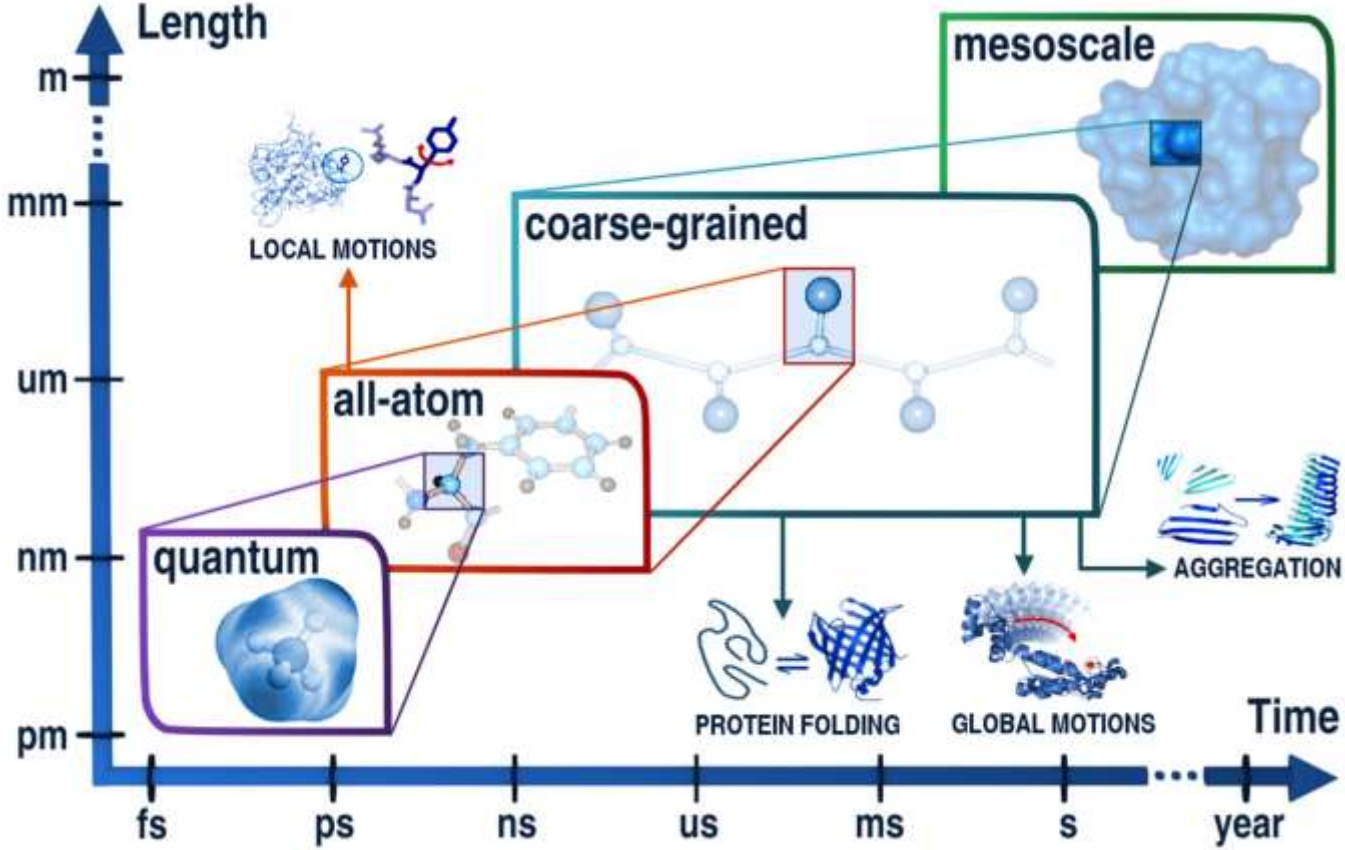


Anton
David E. Shaw Research

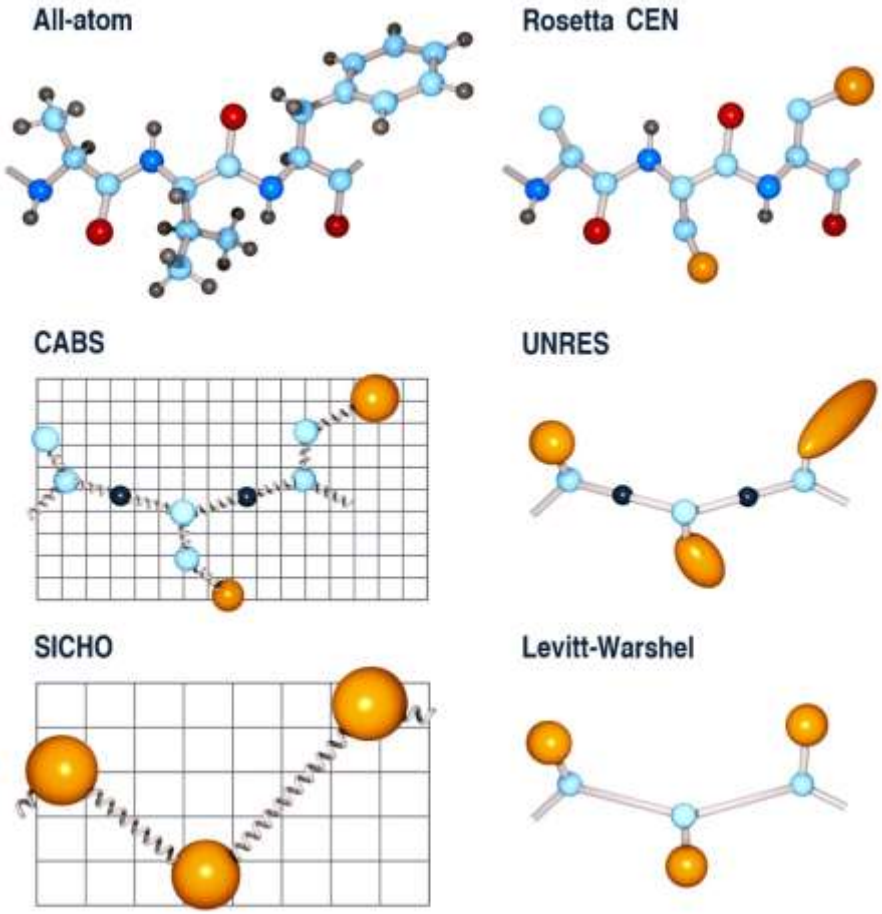
All-atom MD with explicit water
- milliseconds of folding process of a small protein.

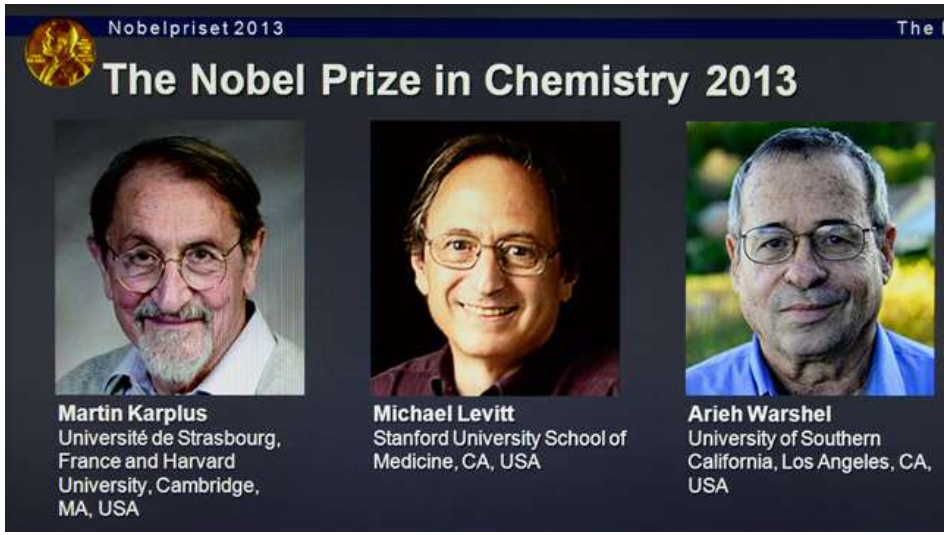
For realistic modeling of larger biomolecular systems, including flexible protein-protein docking, we need much faster simulations.

How to solve the Holy Grail problem



How to solve the Holy Grail problem – Multiscale Modeling





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

CABS model

Cα-Cβ-Side chain

High-coordination lattice

Statistical force-field

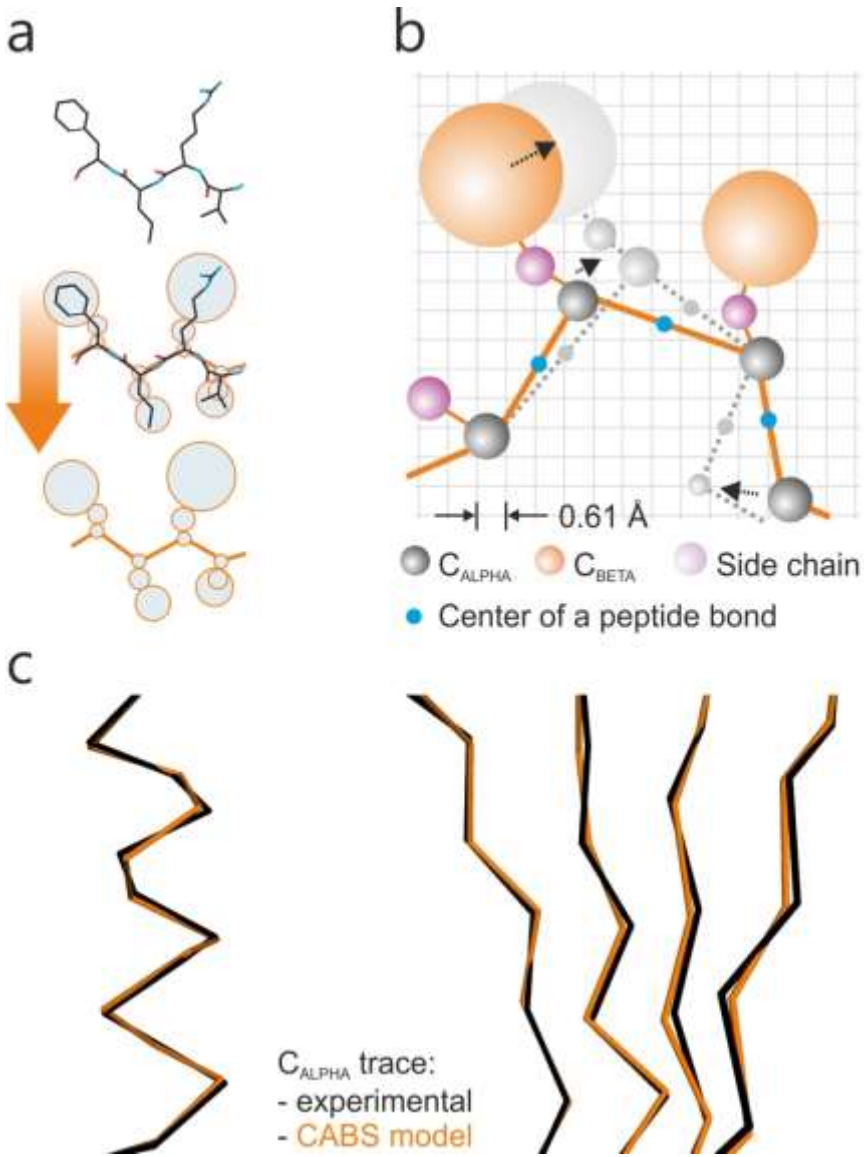
Monte Carlo dynamics

Figures:

a) Building reduced model

b) MC moves on the high-coordination lattice

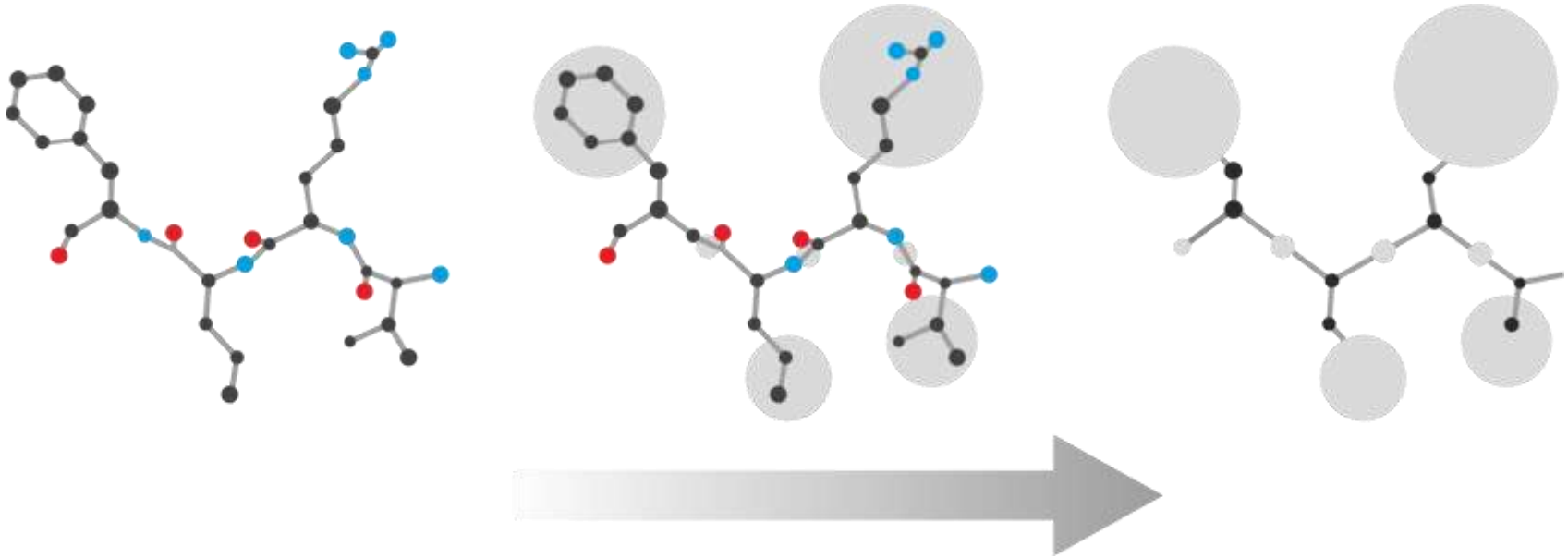
c) Accuracy (Cα-traces)



Time scales MD vs. CABS

All-atom molecular dynamics (MD)

CABS Monte Carlo dynamics

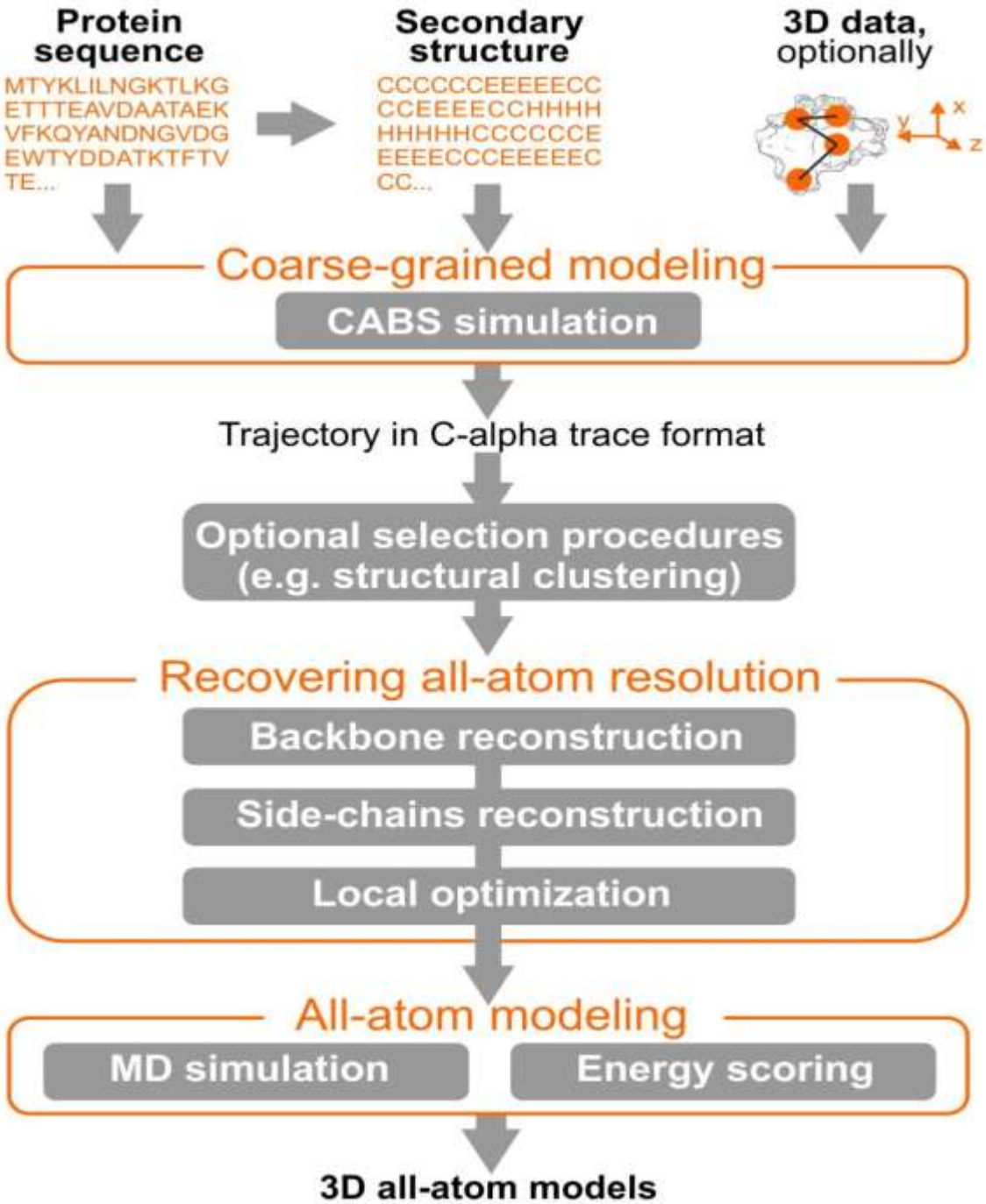


Max ~ 1 millisecond

~ 10^3 / 10^4 faster

Multiscale

modeling with CABS

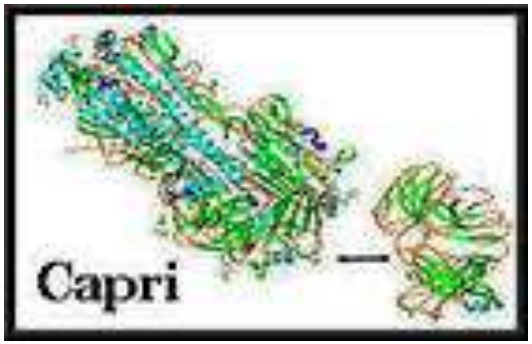


APPLICATINS

Structure prediction

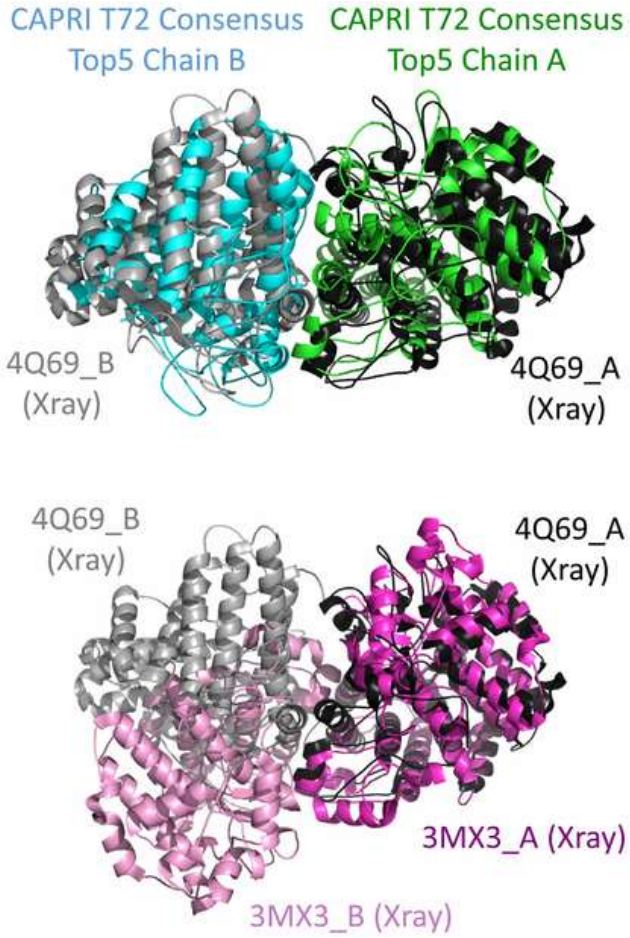
Protein dynamics

Protein docking

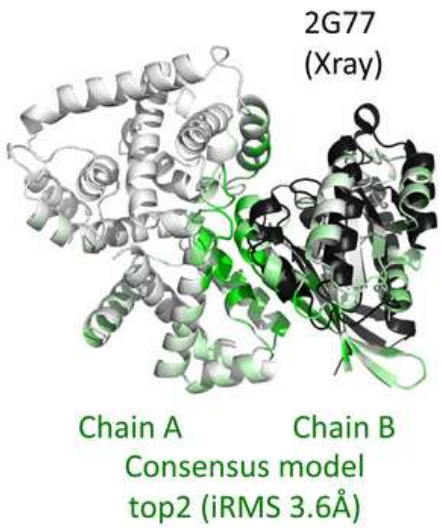


CAPRI: Critical Assessment of PRediction of Interactions

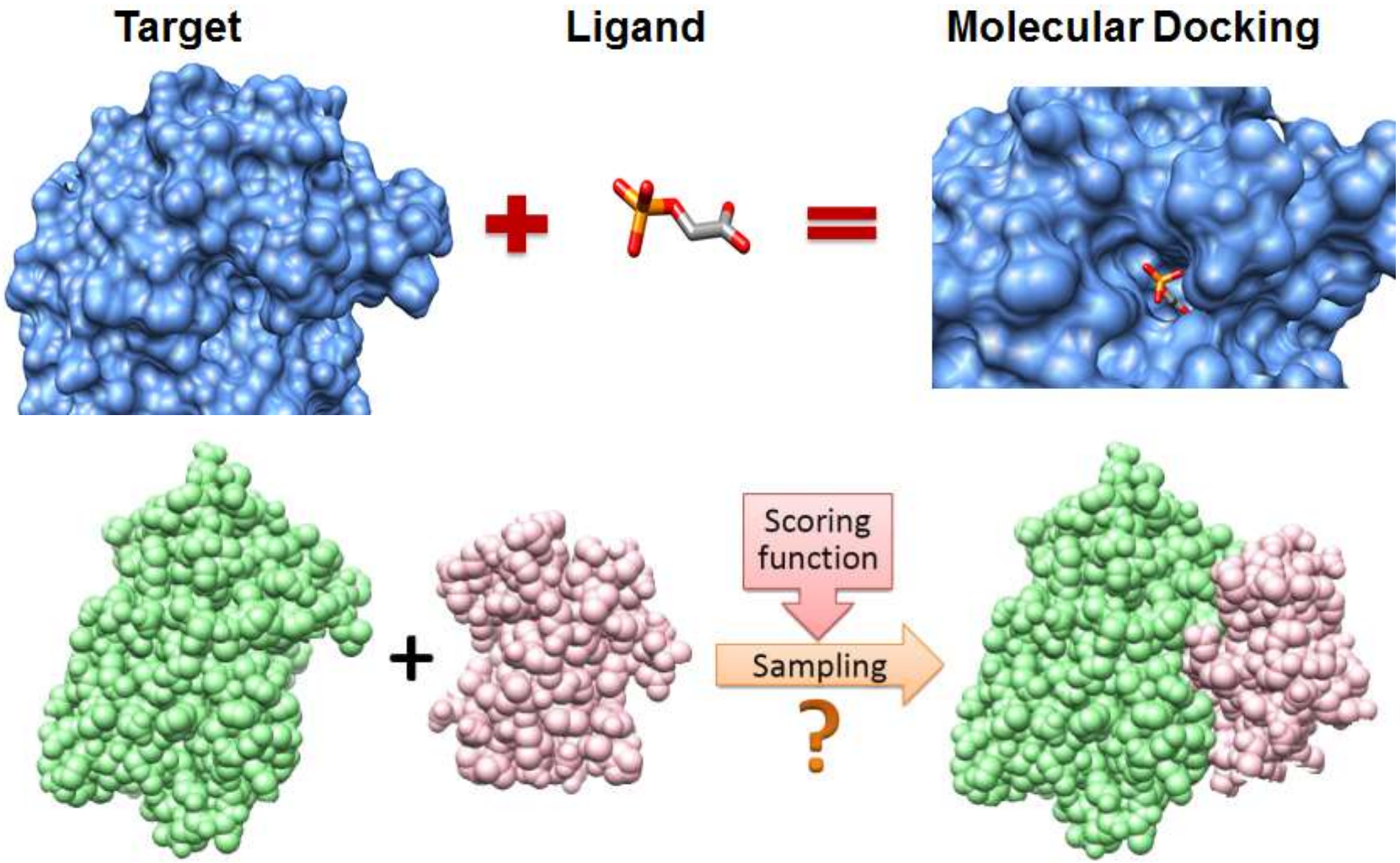
A



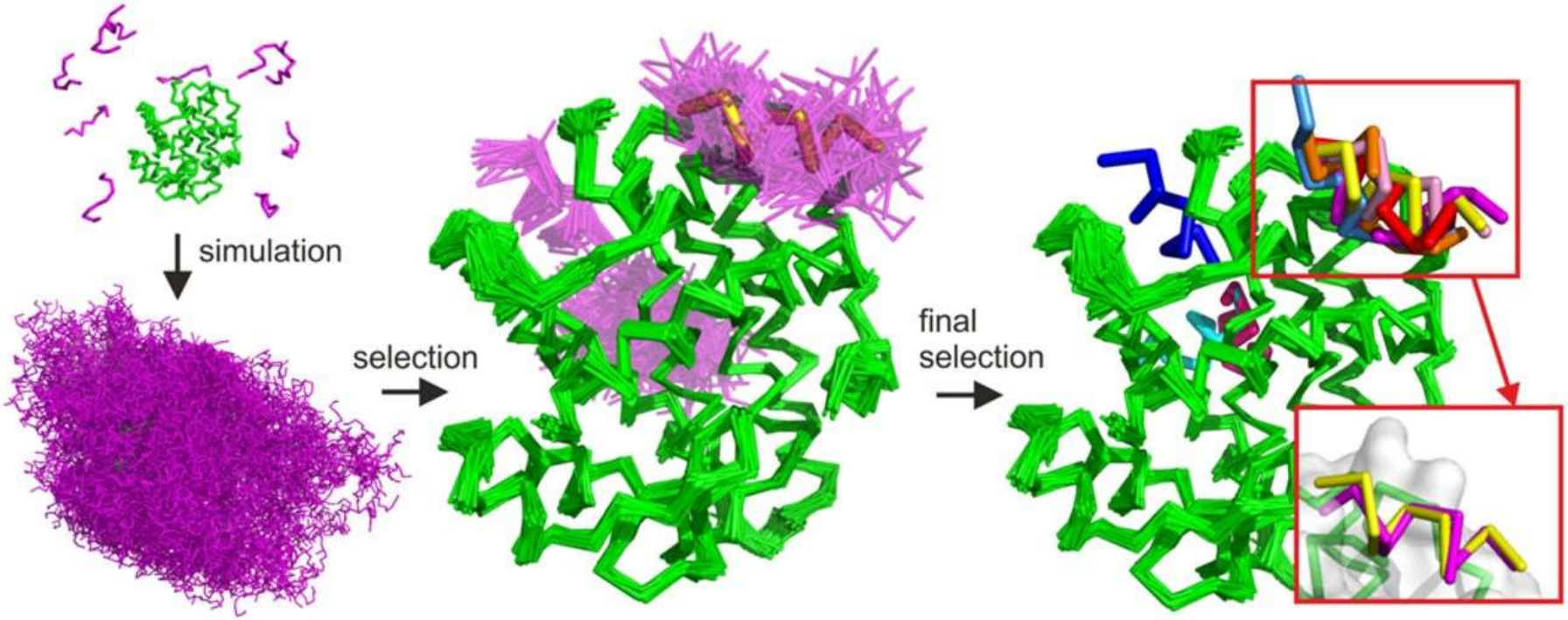
B



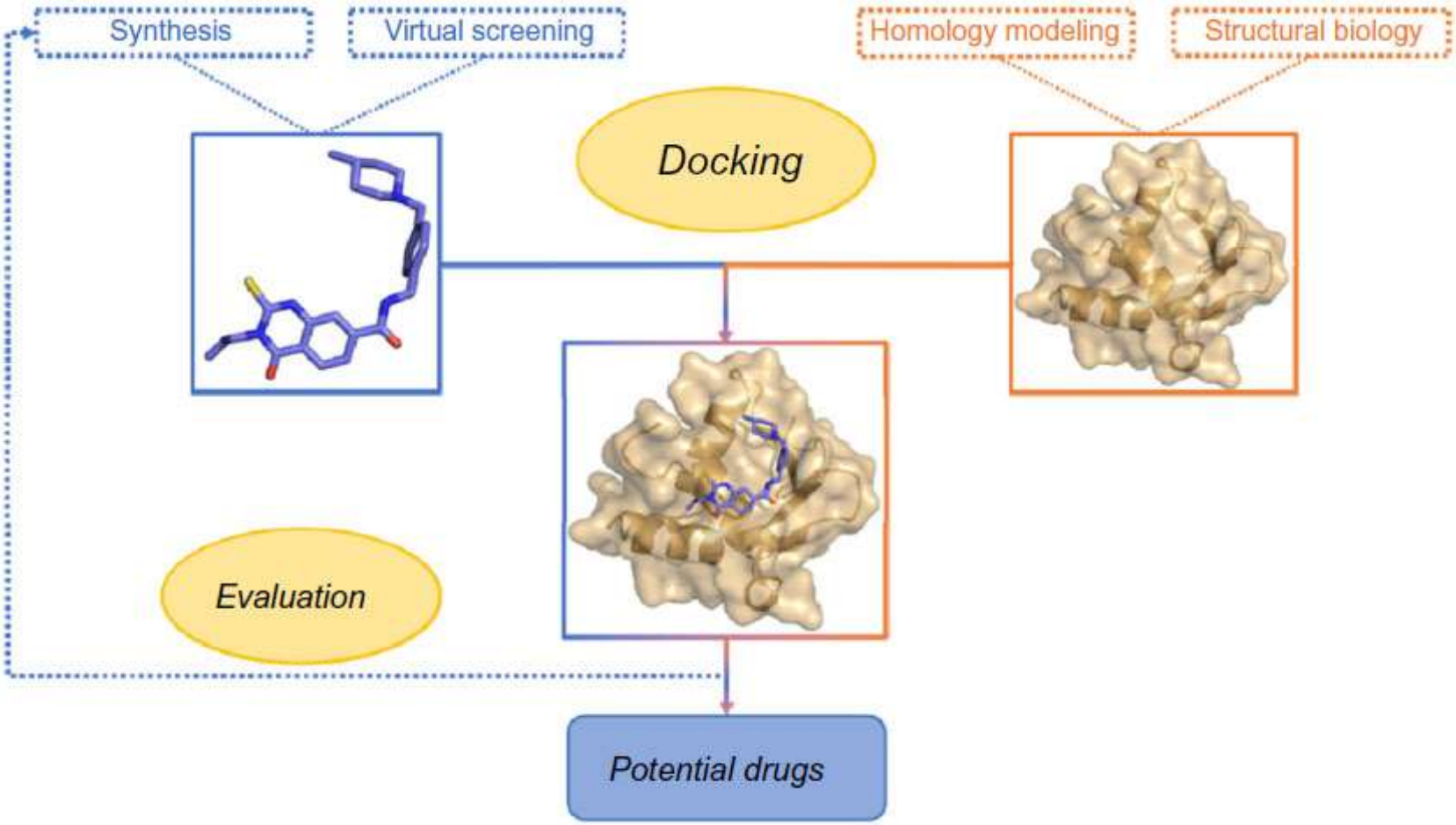
Molecular docking

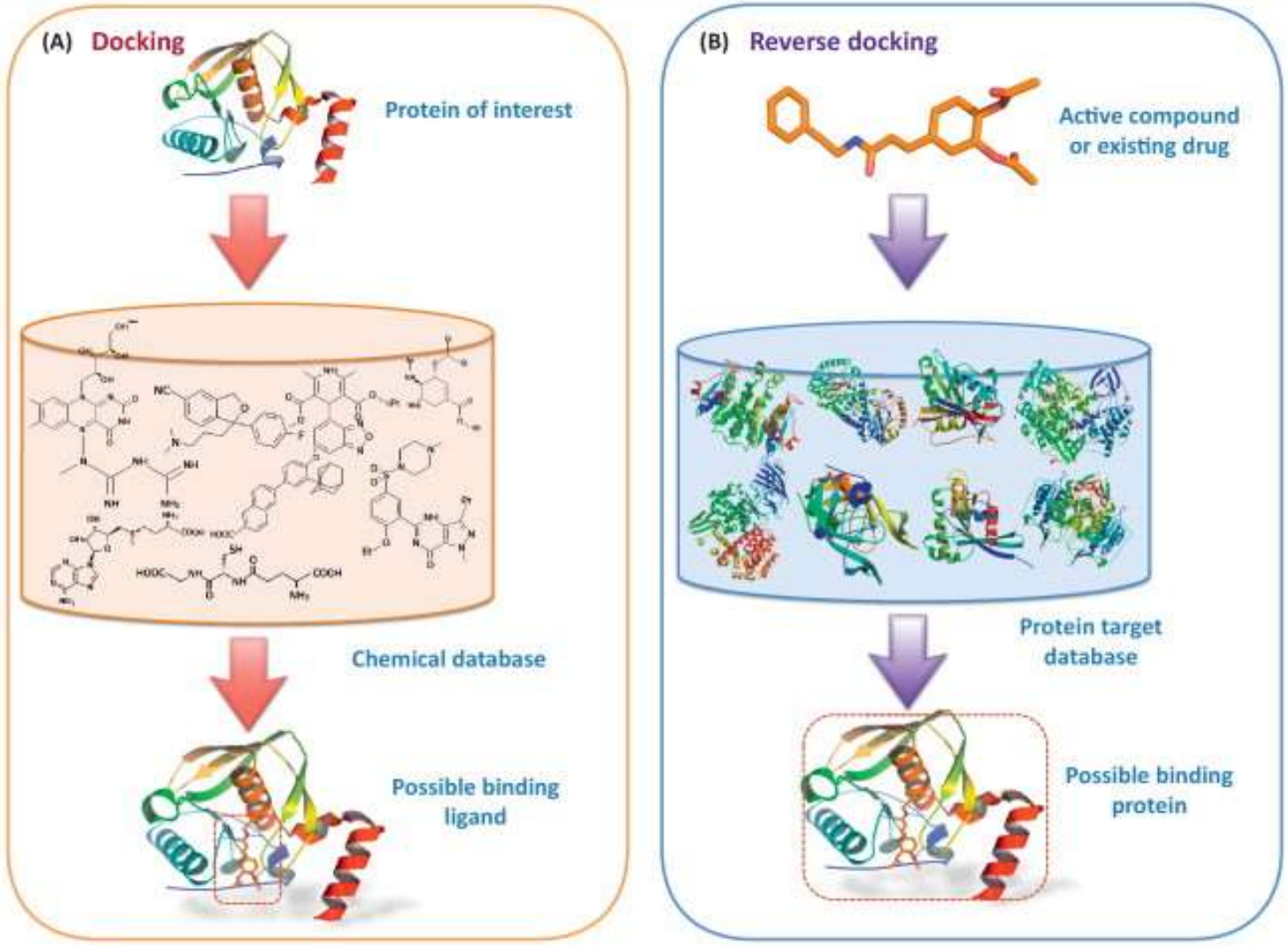


Peptide docking with CABS model



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


Synthesis
Virtual screening
Homology modeling
Structural biology
Docking
Evaluation
Potential drugs



TRENDS in Pharmacological Sciences