LABORATORY of THEORY of BIOPOLYMERS

Proteins - structural bioinformatics (2)

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)

Tertiary Structure and the "Hydrophobic Effect"

What would this protein look like when properly folded?

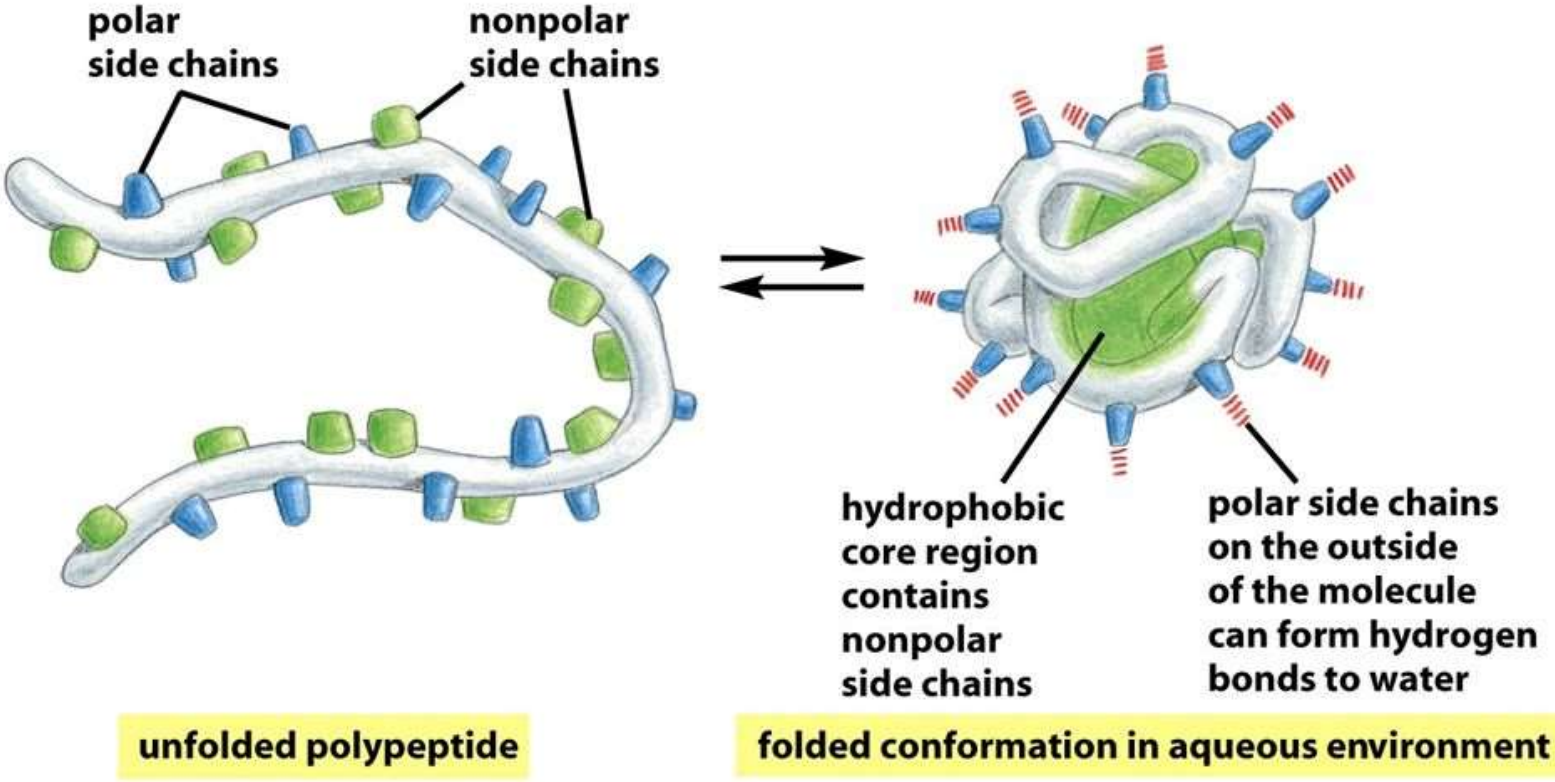
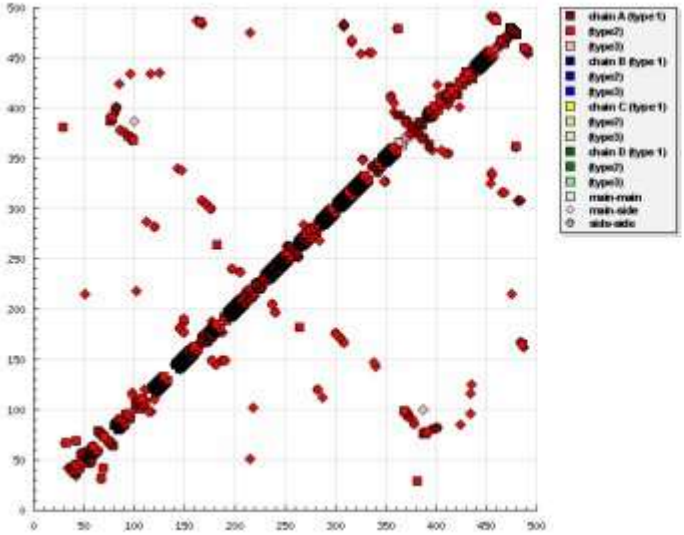
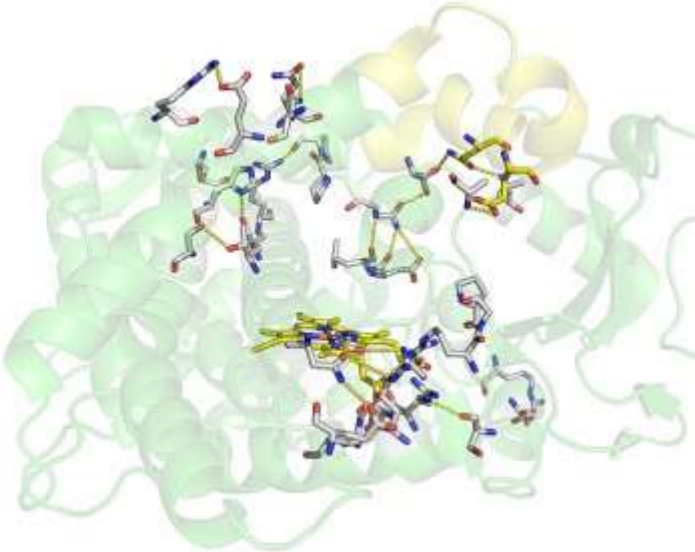
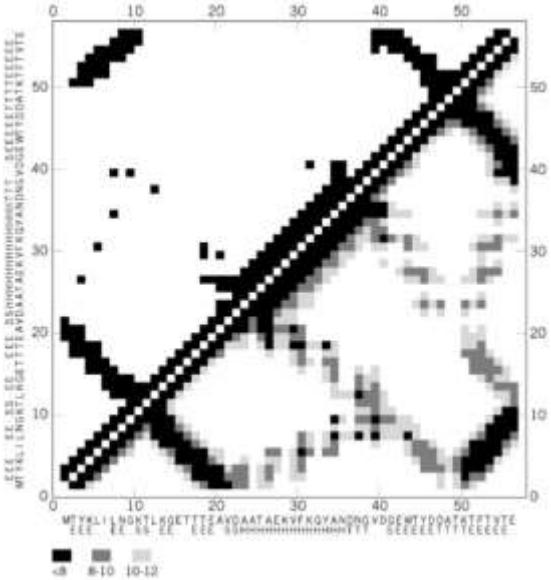


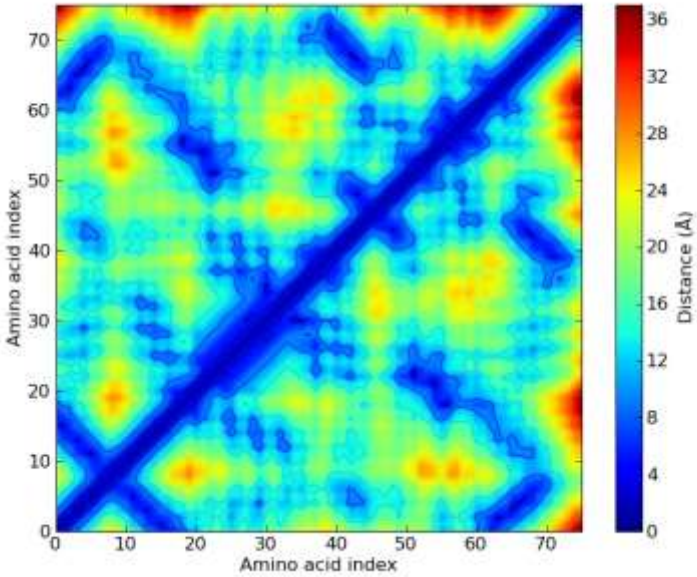
Figure 3-5 *Molecular Biology of the Cell* (© Garland Science 2008)

Side chain packing

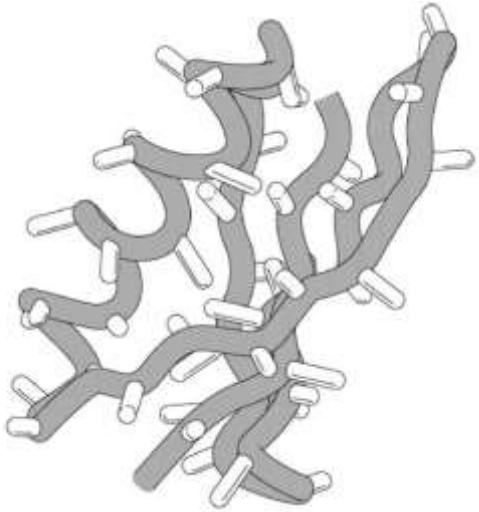








Side chain packing

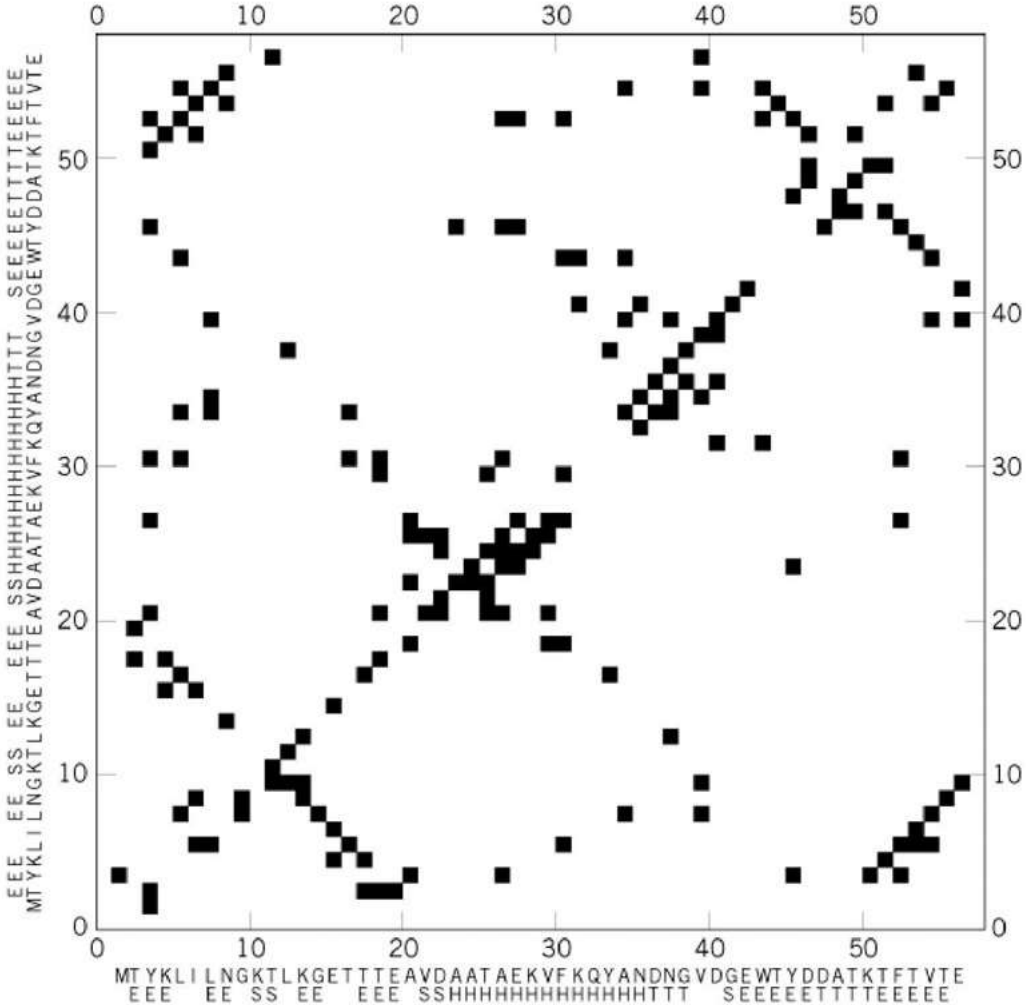


coiled coil

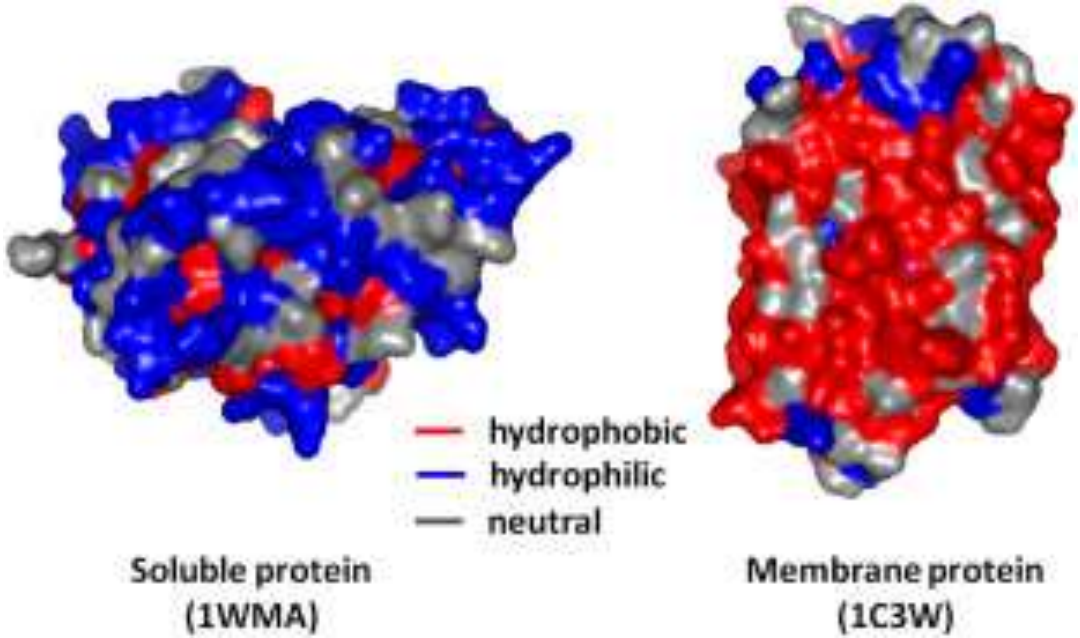
helix-helix

parallel beta-beta

anti-parallel beta-beta



Hydrophobic effects



Tertiary Structure and the "Hydrophobic Effect"

What would this protein look like when properly folded?

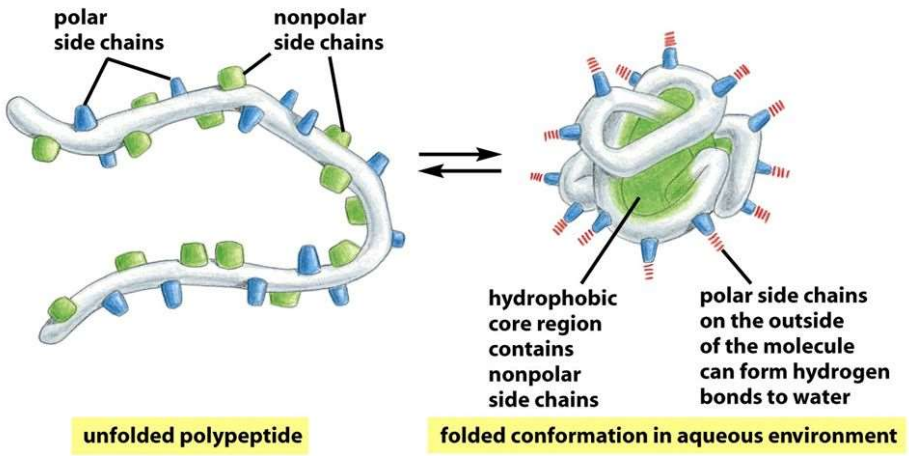
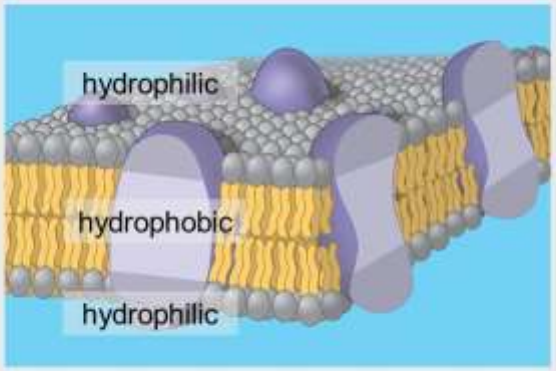


Figure 3-5 *Molecular Biology of the Cell* (© Garland Science 2008)

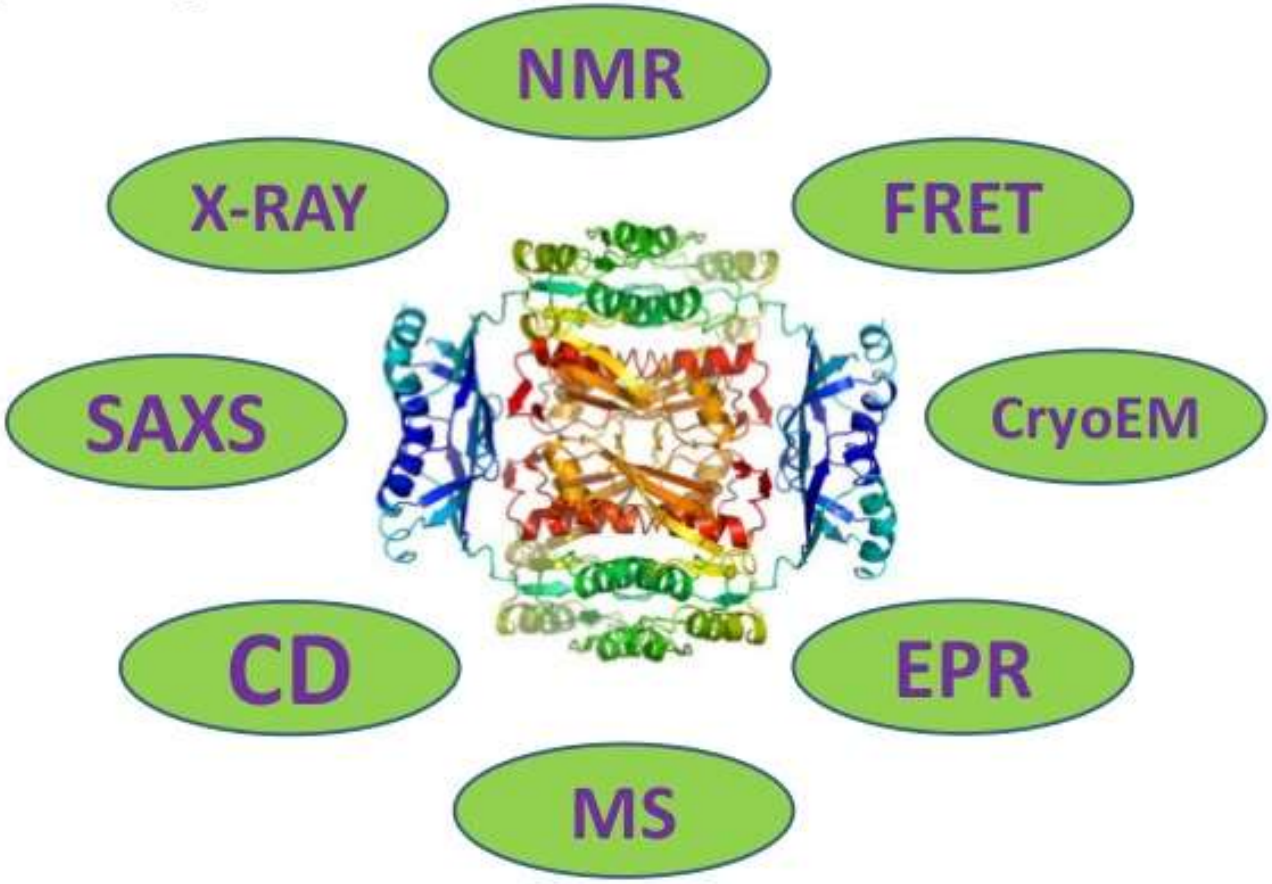
Membrane structure

- cell membrane – amphipathic - hydrophilic & hydrophobic



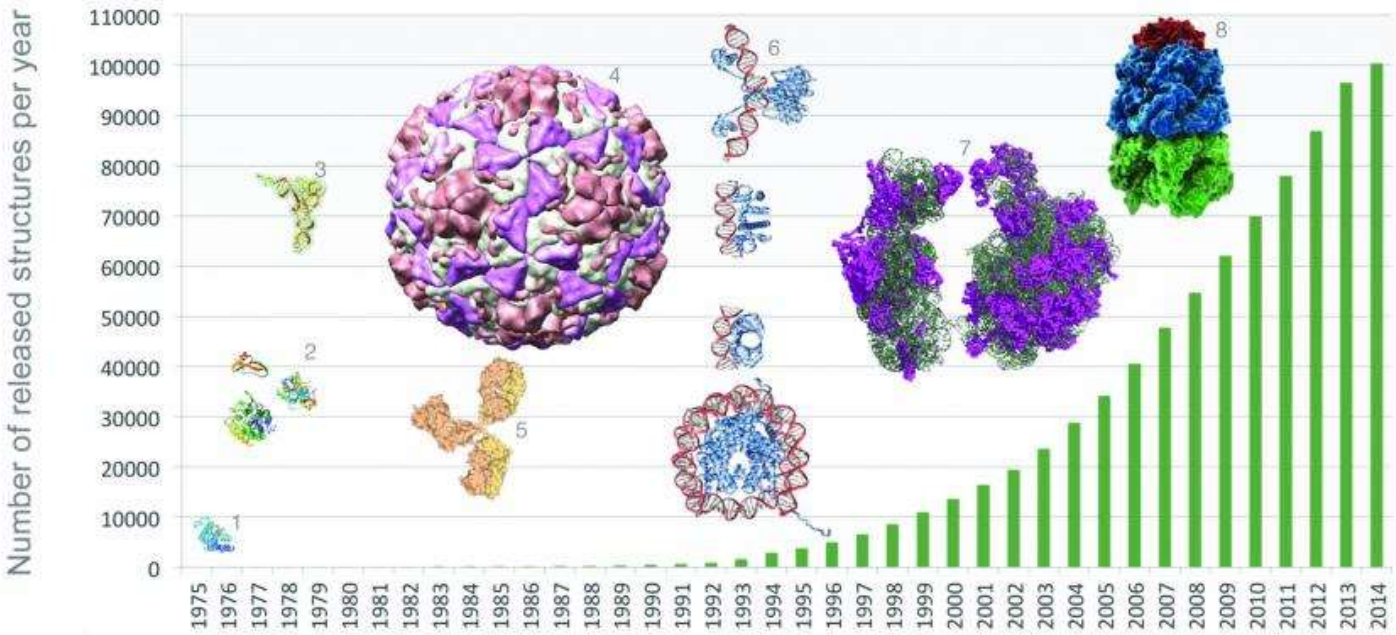
- membrane proteins that are inserted, also amphipathic

Hybrid protein structure determination

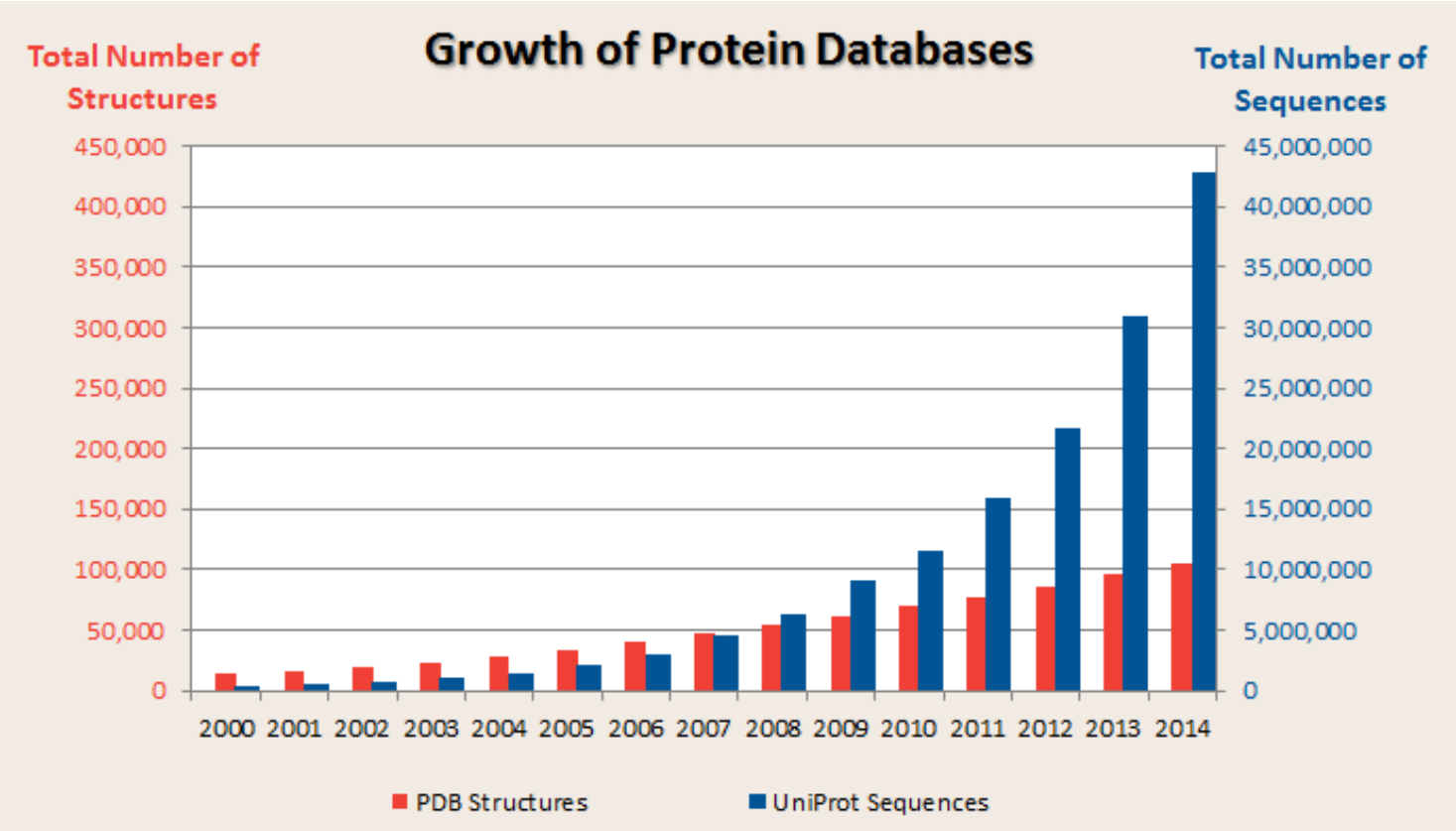


Mark Berjanskii, Edmonton, July 2015

Research Collaboratory for Structural Bioinformatics:
Rutgers and UCSD/SDSC



Sequence - structure



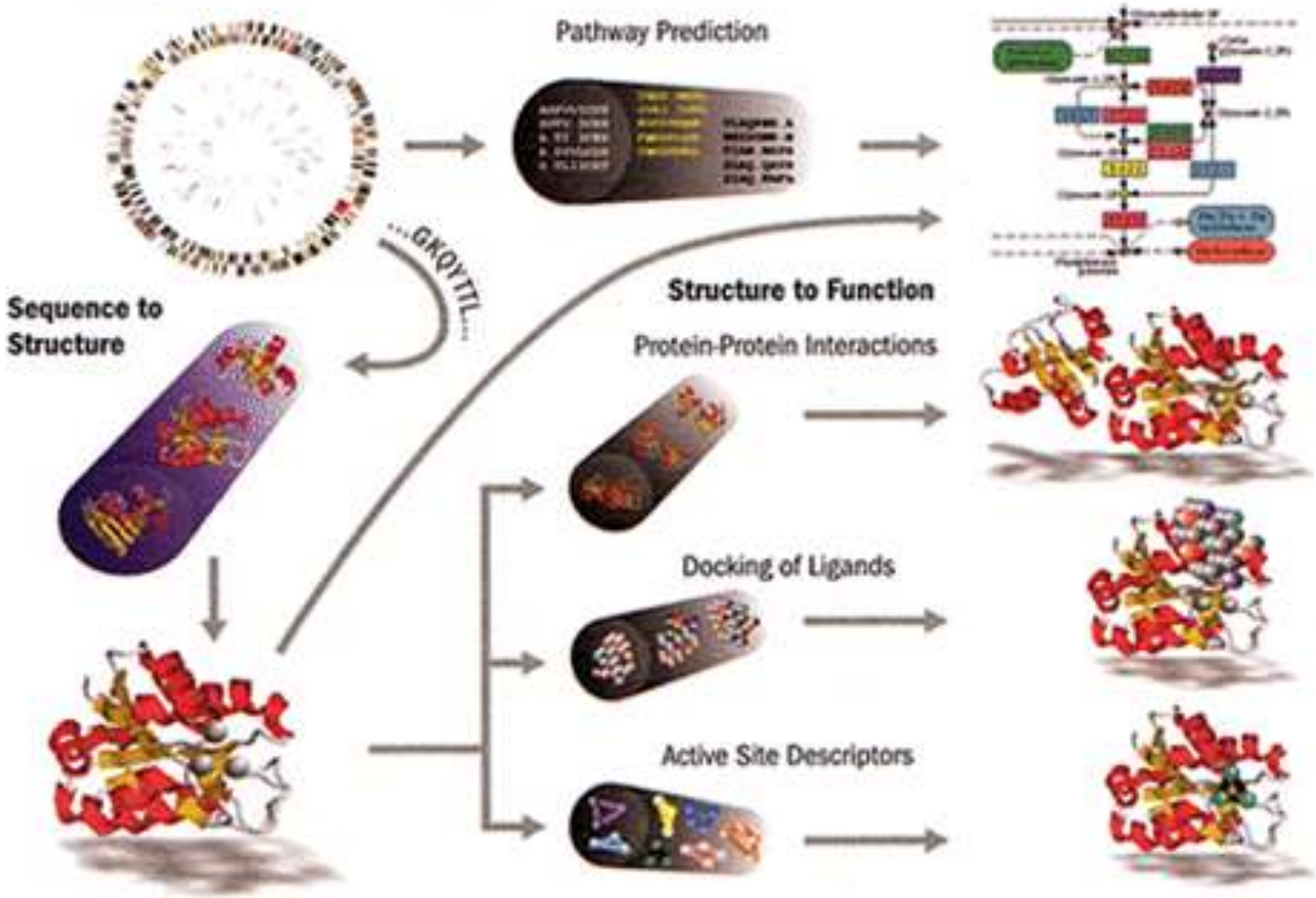
Protein Data Bank (PDB) - 140 000 protein structures

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)

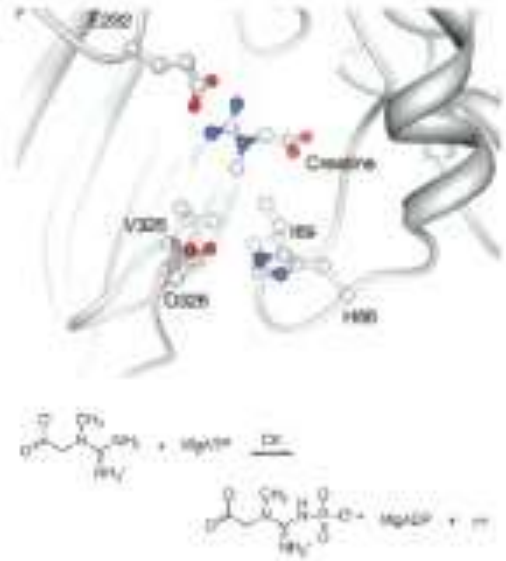
The Sequence-to-Structure-to-Function Paradigm

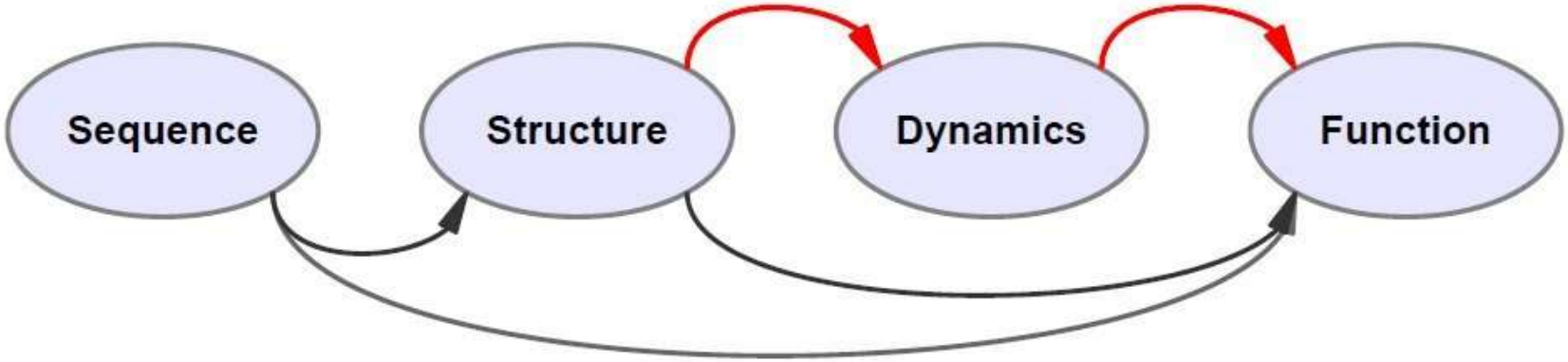


All the potential open reading frames (ORFs) in a protein sequence are threaded through a library of previously solved template protein structures. If a template is found, the structure is scanned for a match to a known active site. Alternatively, ligands can be virtually docked to identify the active site. Threading can also be used to identify potential interacting partners in the genome, or assist ORF pathway assignment.

Sequence → Structure → Function

MPFGNTHNKFKL
NYKPEEEYPDLSK
HNNHMAKVLTLE
LYKKLRDKETPSGF
TVDDVIQTGVDNP
GHPFIMTVGCVAG
DEESYEVFKELFDPI
ISDRHGGYKPTD...





Local Alignment

Pairwise Sequence Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

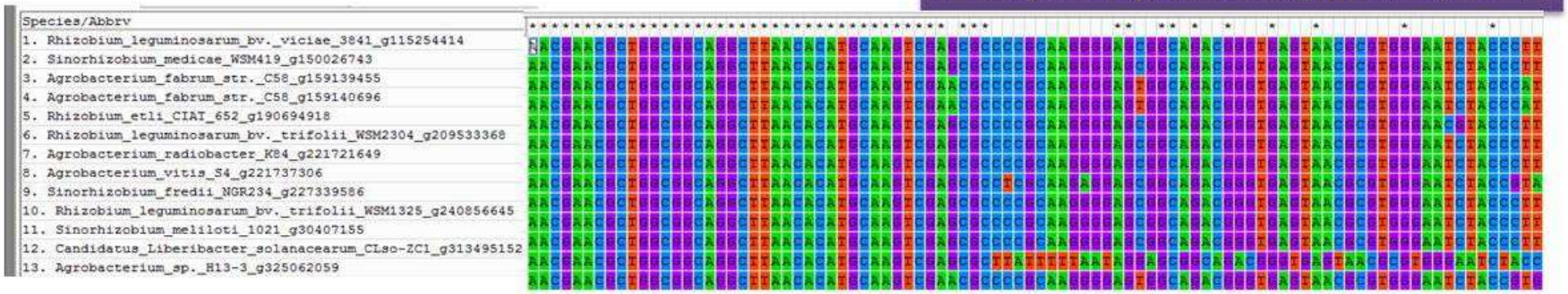
Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

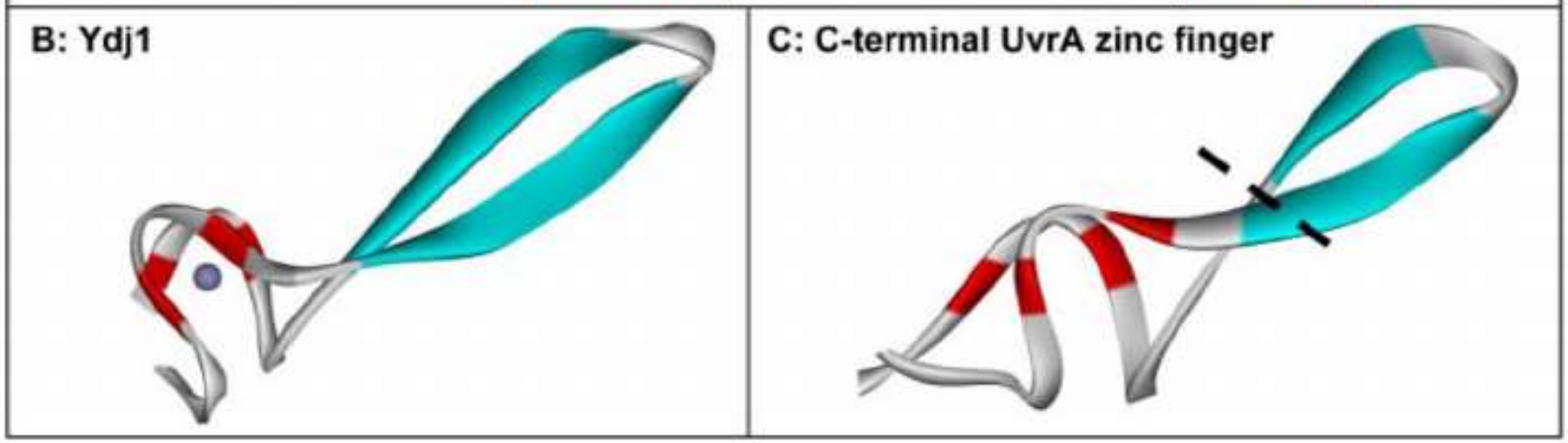
Query Sequence

Multiple Sequence Alignment (MSA)



Structure – Comparative modeling

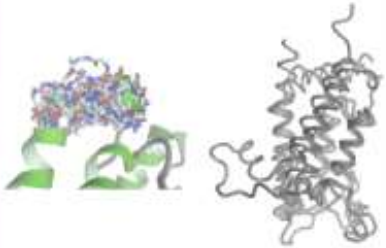
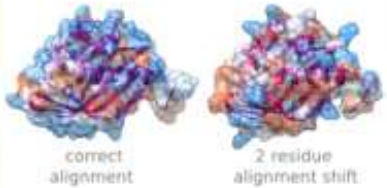
```
A: ECOLI   GRFSFNVRGGRCEACQGDGVIKVEMHFLPDIYVP---CDQCKGKRYNRETLE
   RHIME   GRFSFNVKGGRCEACQGDGVIKIEMHFLPDVYVT---CDVCHGKRYNRETLD
   TREPA   GRFSFNVPGGRCEHCKGDGVITIEMNFLPDVYIT---CDVCHGTRFNRETLA
   HELPJ   SRFSFNVKGGRCEKCQGDGDIKIEMHFLPDVLVQ---CDSCKGAKYNPQTLE
   BCACA   GRFSFNVKGGRCEACHGDGIIKIEMHFLPDVYVP---CEVCHGKRYNRETLE
   ZnG A   GRFSFNVKGGRCEACHGDGII-----G-----VP---CEVCHGKRYNRETLE
   Ydj1    GRGGKKGAVKKCTSCNGQGIKFVTRQMGPMIQRFQTECDVCHGTGDIIDPKD
```



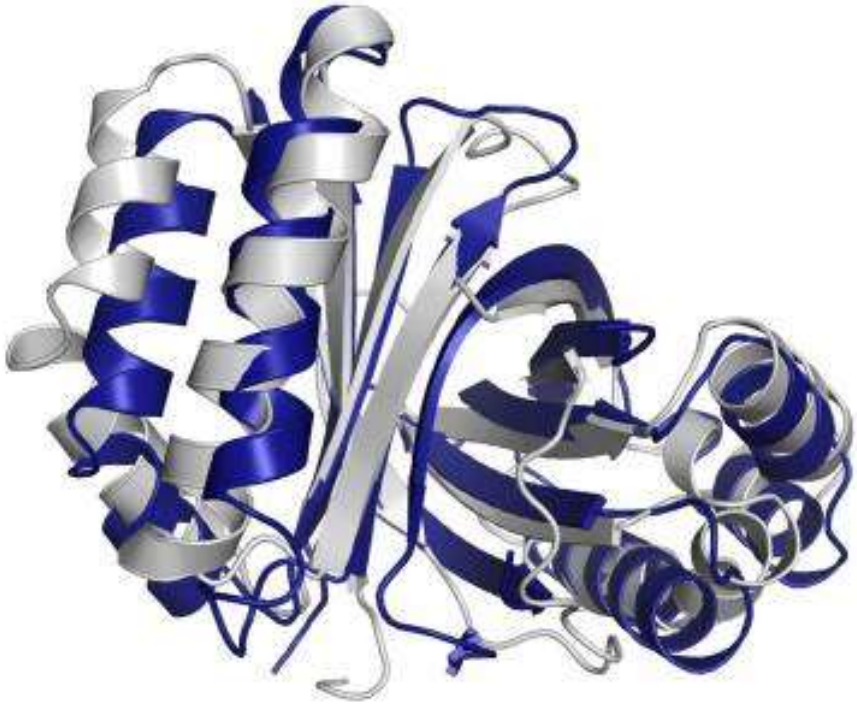
Comparative Modeling--Basic Protocol

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.

Sources of errors			Applications
- experimental errors and uncertainties in X-ray, NMR	1Å 100%		- 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) - distorsions of aligned regions	60%		- structural support for mutagenesis studies - molecular replacement
- alignment errors	3Å 40%		- integrative modeling - modeling into low-resolution density maps
- sub-optimal template selection - model may even have the wrong fold	>3Å <30%		- domain boundaries - identification of structural motives

Comparative (homology) modeling



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

A

B

Protein folding problem

PRIMARY STRUCTURE (amino acid sequence)

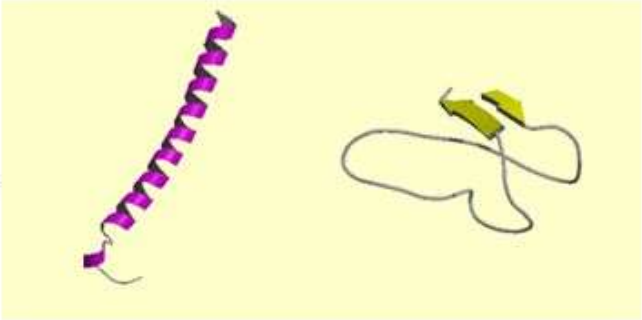
VHLTPEEKSAVTALWGKVNVDE
VGGEALGRLLVVYPWTQRFFE
SFGDLSTPDAVMGNPKVKAHG
KKVLGAFSDGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGN
VLVCVLAHHFGKEFTPPVQAAY
QKVVAGVANALAHKYH

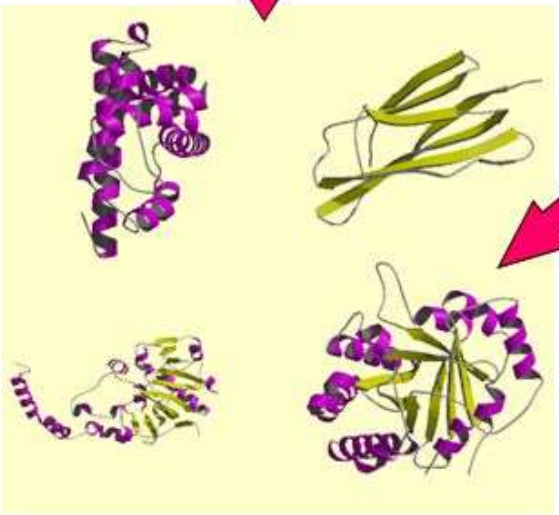
Each protein sequence "knows" how to fold into its tertiary structure. We still do not understand how and why

1-step process

2-step process

SECONDARY STRUCTURE (helices, strands)

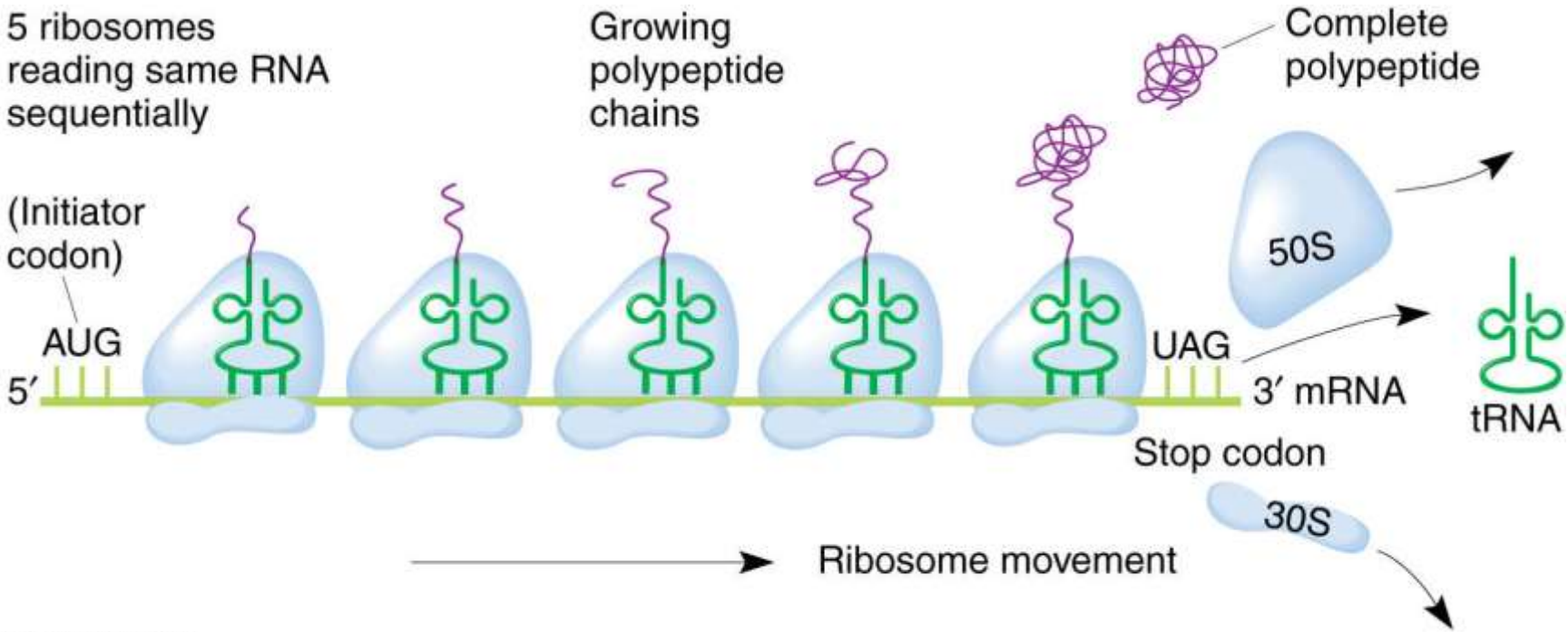




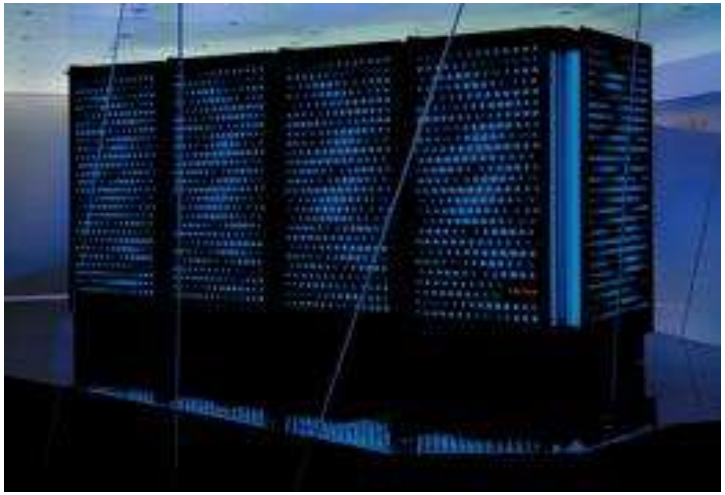
TERTIARY STRUCTURE (fold)

The 1-step process is based on a hydrophobic collapse; the 2-step process, more common in forming larger proteins, is called the framework model of folding

Post-translational or co-translational folding



Protein folding problem - the Holy Grail of the structural biology

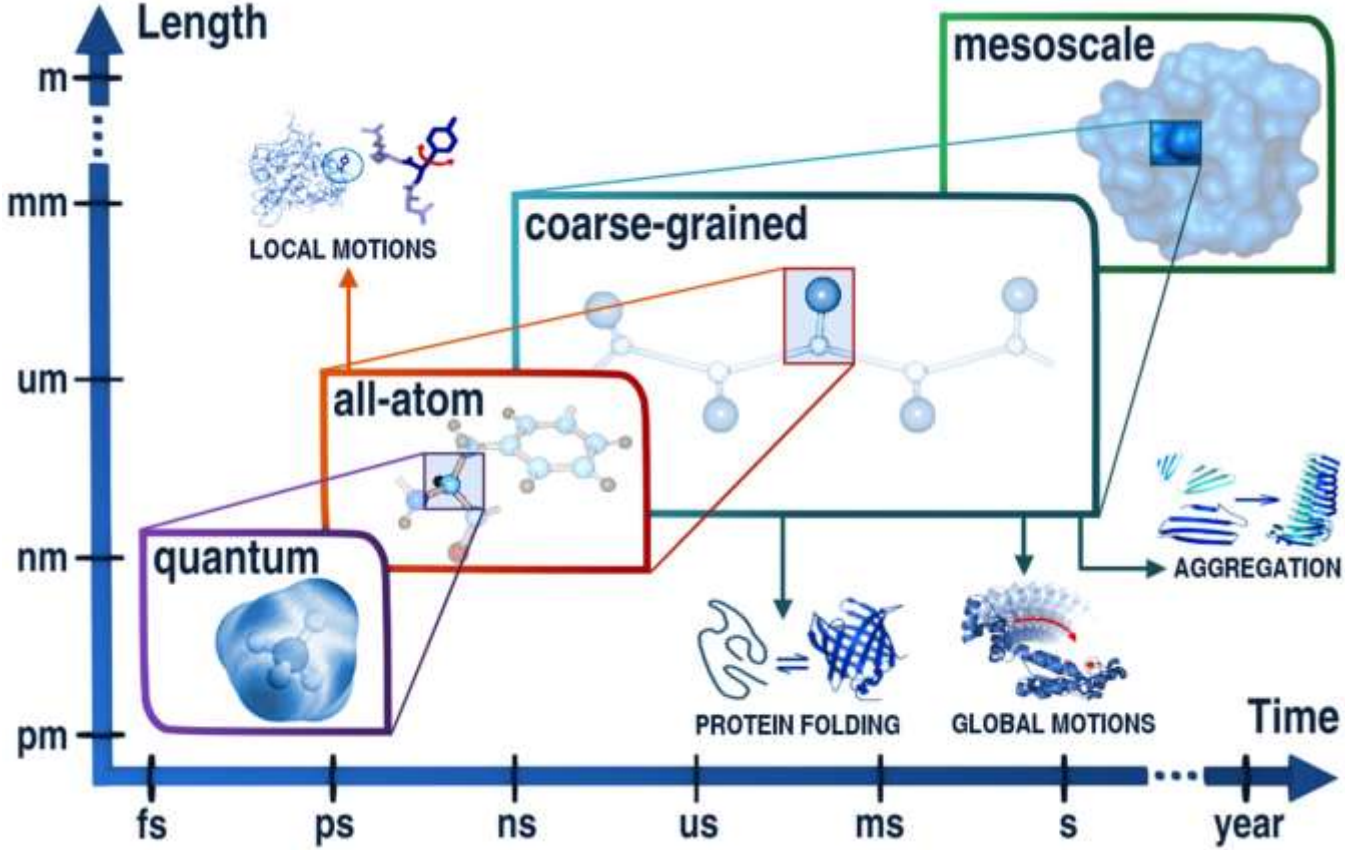


Anton
David E. Shaw Research

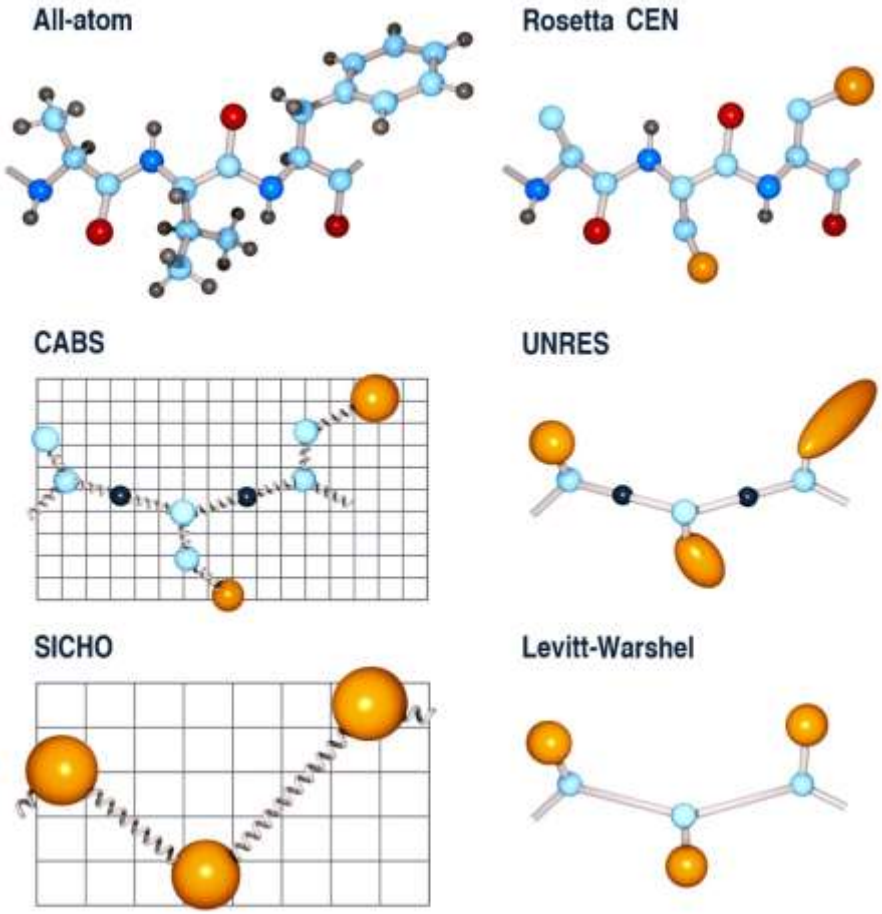
All-atom MD with explicit water
- milliseconds of folding process of a small protein.

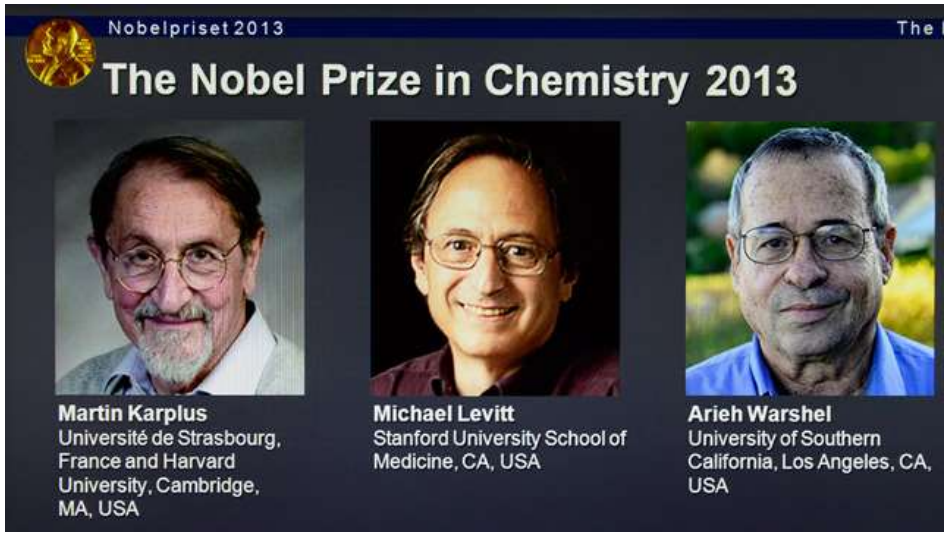
For realistic modeling of larger biomolecular systems, including flexible protein-protein docking, we need much faster simulations.

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

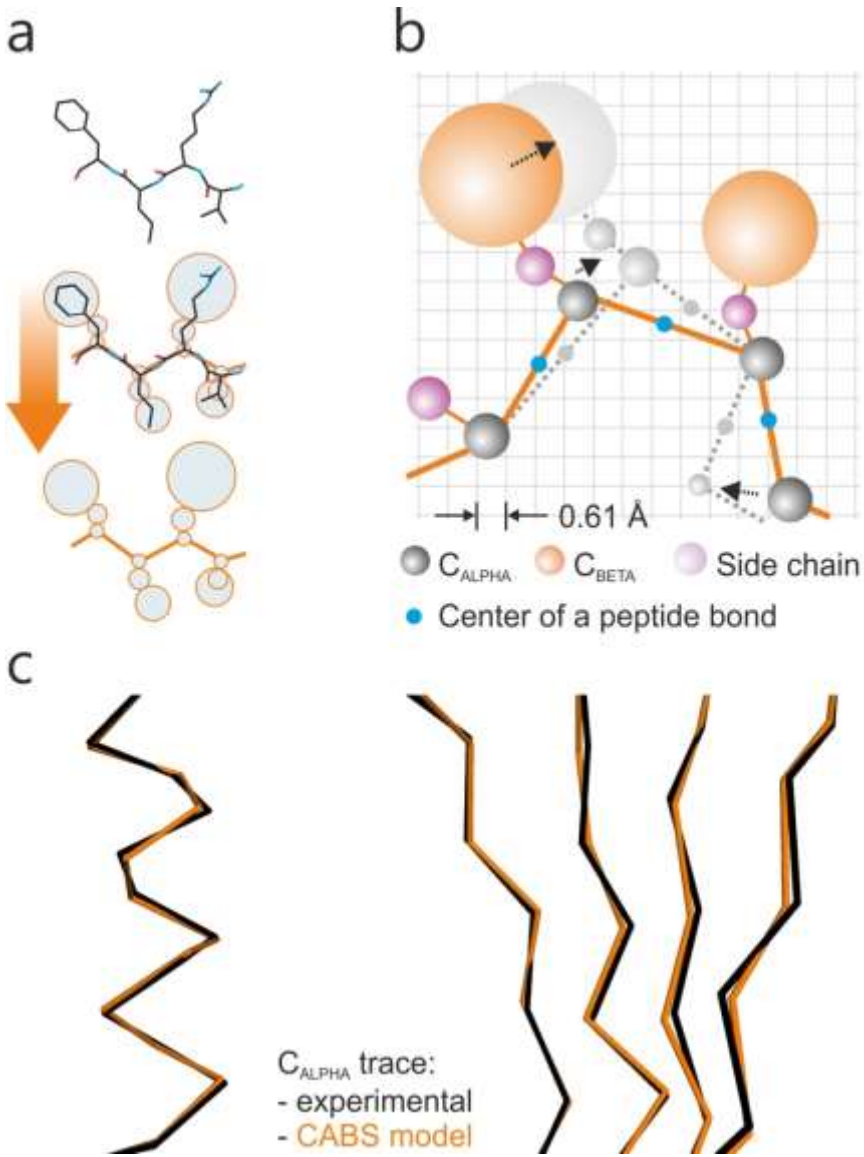
Monte Carlo dynamics

Figures:

a) Building reduced model

b) MC moves on the high-coordination lattice

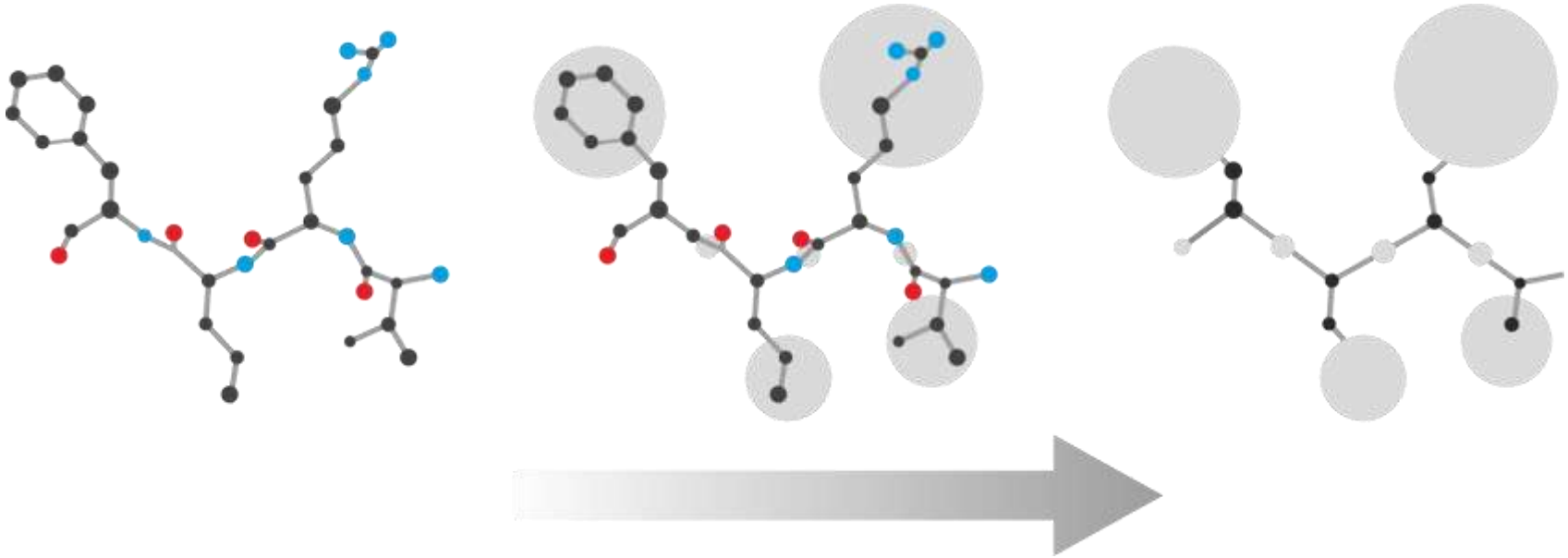
c) Accuracy (Cα-traces)



Time scales MD vs. CABS

All-atom molecular dynamics (MD)

CABS Monte Carlo dynamics

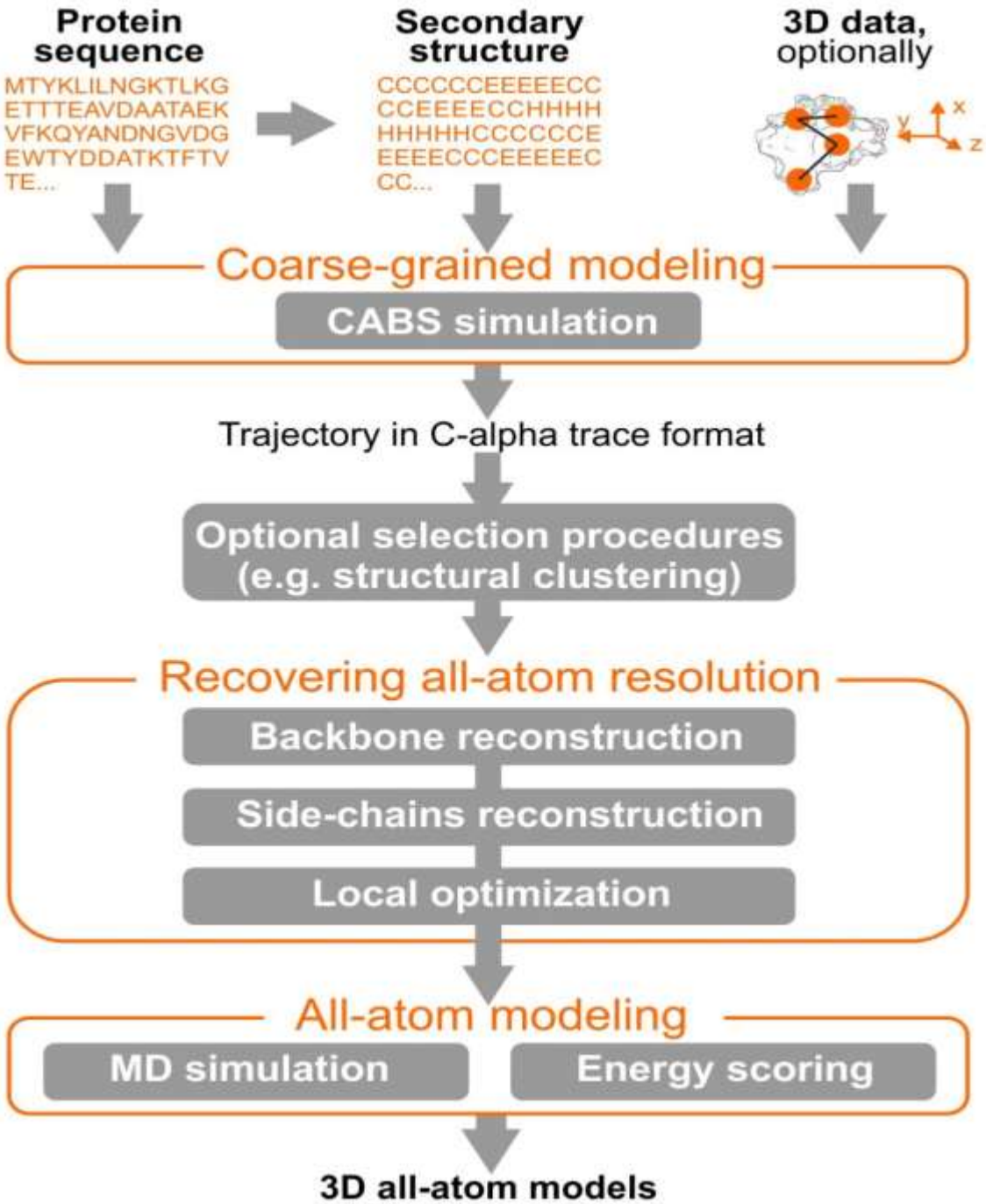


Max ~ 1 millisecond

~ 10^3 / 10^4 faster

Multiscale

modeling with CABS



APPLICATINS

Structure prediction

Protein dynamics

Protein docking

CASP and CAPRI

CASP Competition

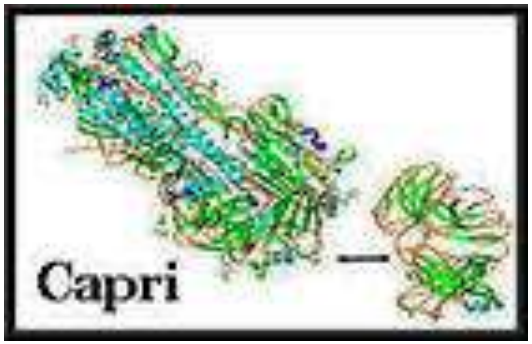
- CASP competition (Critical Assessment of Techniques for Protein Structure Prediction) http://predictioncenter.llnl.gov/
- Their goal is to help advance the methods of identifying protein structure from sequence.

29

CASP Experiment

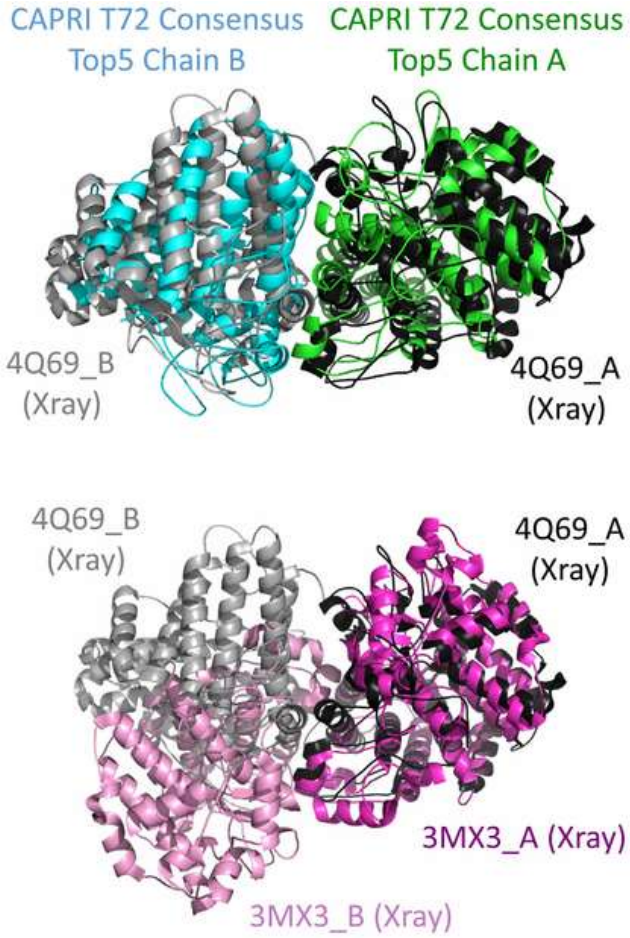
- Experimentalists are solicited to provide information about structures expected to be soon solved
- Predictors retrieve the sequence from prediction center (predictioncenter.llnl.gov)
- Deposit predictions throughout the season
- Meeting held to assess results

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

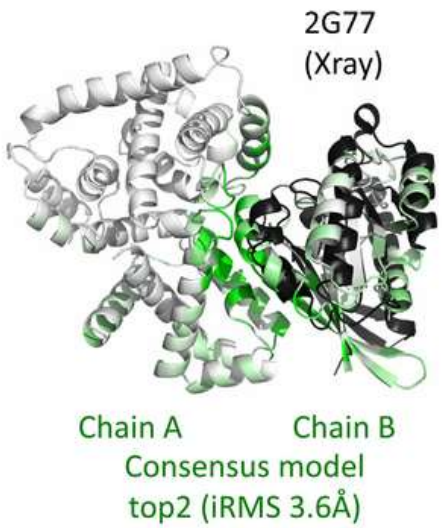


CAPRI: Critical Assessment of PRediction of Interactions

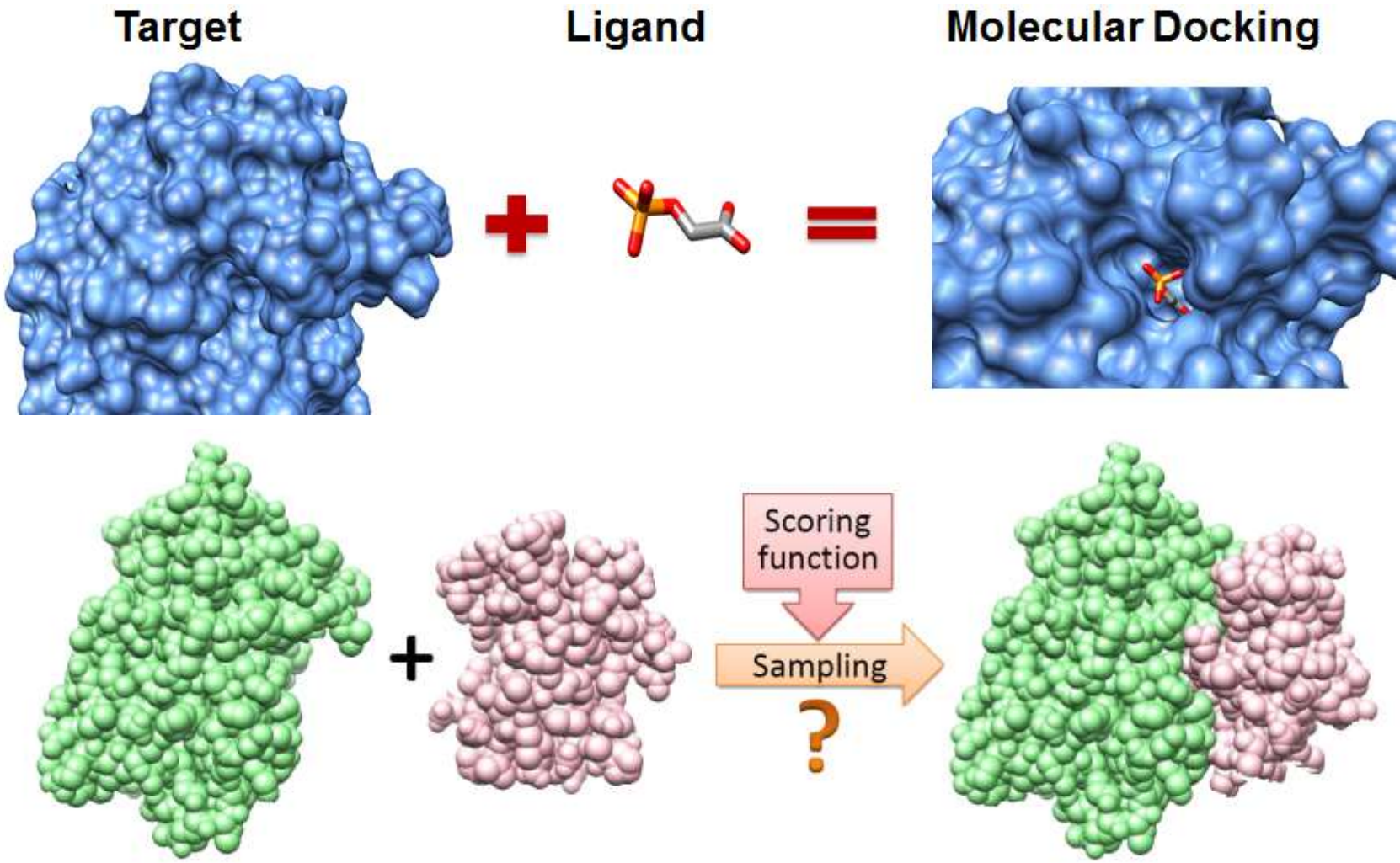
A



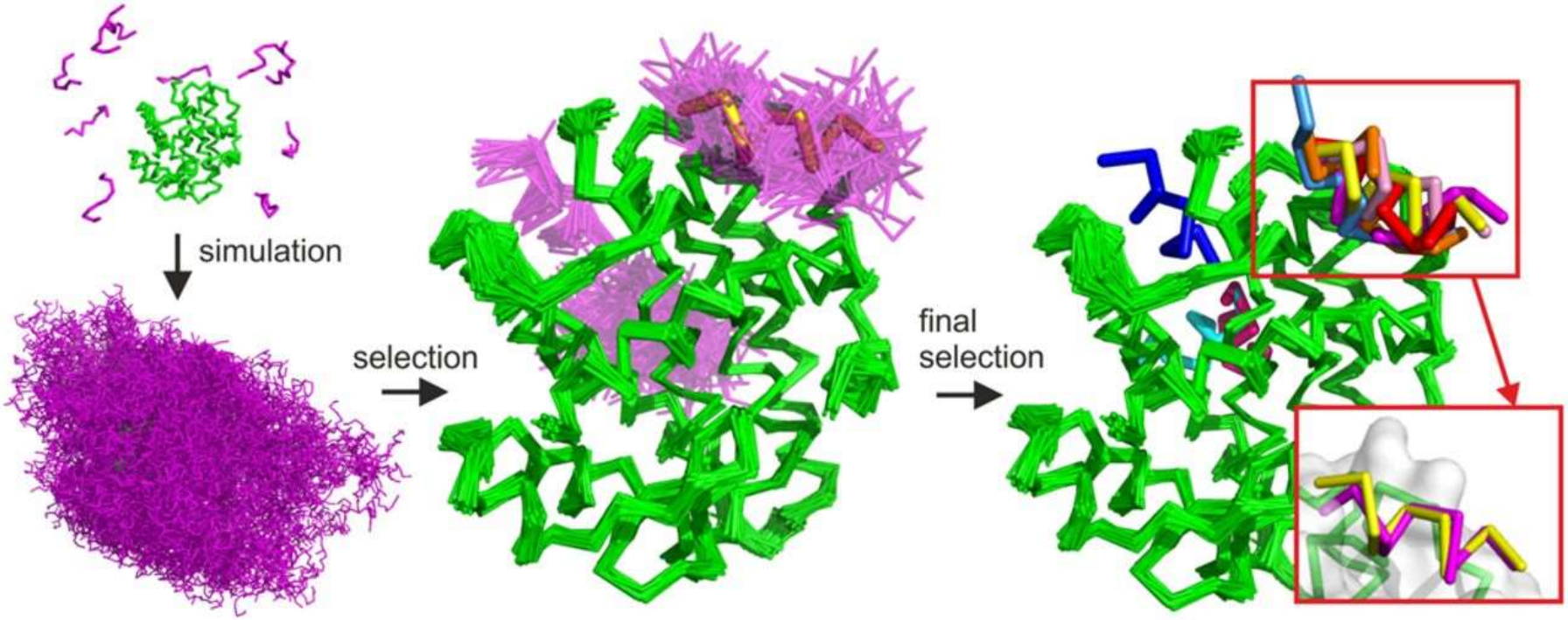
B



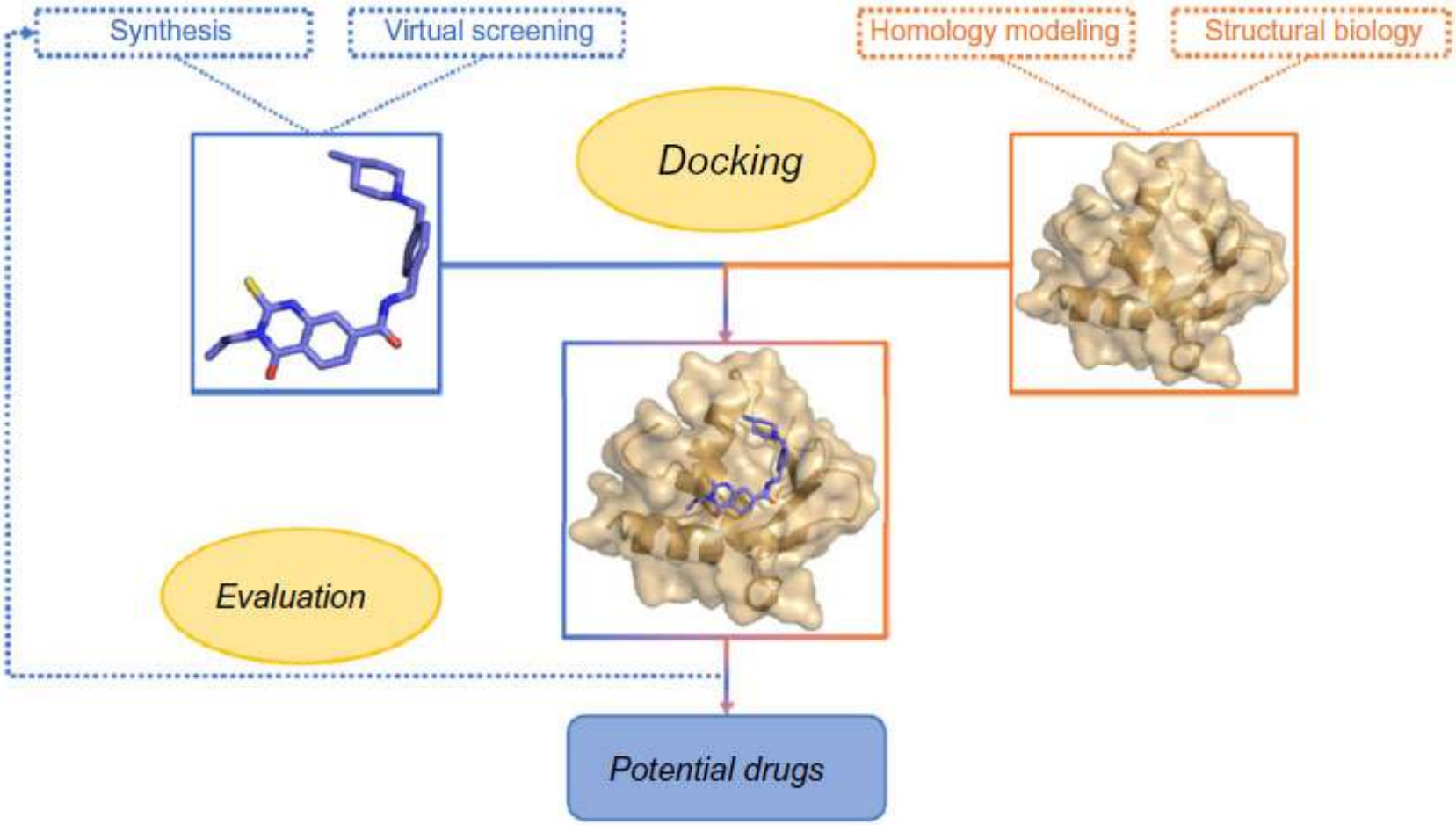
Molecular docking

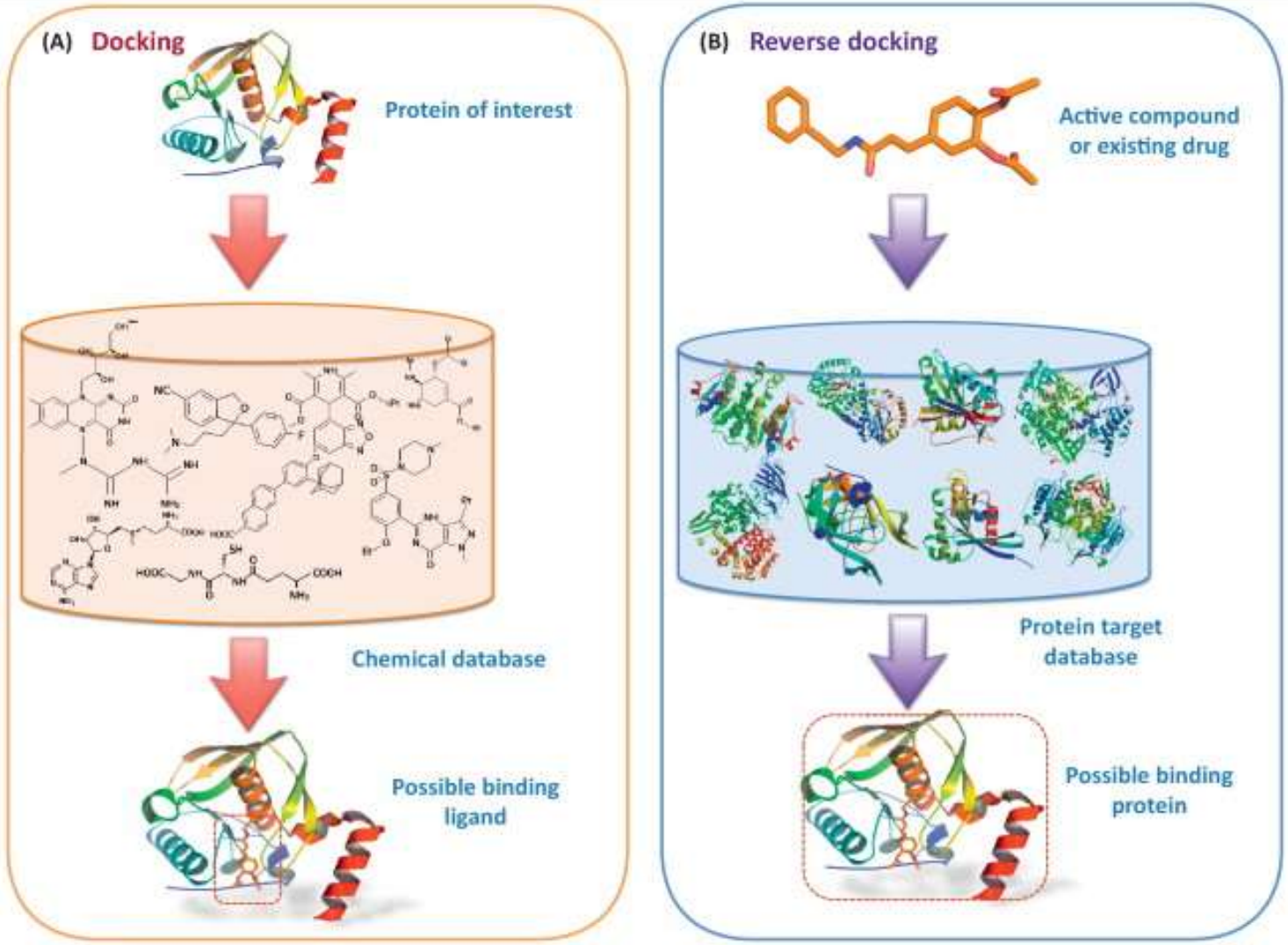


Peptide docking with CABS model



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

Check our tools at:

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

LABORATORY of THEORY of BIOPOLYMERS

Andrzej Kolinski Research Gr

RESEARCH PEOPLE PUBLICATIONS TOOLS CONTA

TOOLS

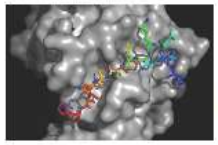
- CABS-DOCK
- CABS-FLEX
- CABS-FOLD
- AGGRESCAN3D
- CABS
- SURPASS
- CABS-NMR
- PYCABS
- CLUSCO
- BBQ
- BIOSHELL
- CCOMP
- MSITE
- BIODESIGNER
- IMOL

METHODS: The used methodologies were reviewed in our review paper on Coarse-Grained Protein Models and Their Applications (pdf) in Chemical Reviews journal.

FUNDING: CABS-dock, CABS-flex, CABS-fold and Aggrescan3D tools were funded by the Foundation for Polish Science TEAM project (TEAM/2011-7/6) co-financed by the European Regional Development Fund operated within the Innovative Economy Operational Program.

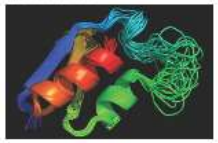
CABS-dock

server for protein-peptide docking and prediction of binding sites



CABS-flex

server for fast simulations of flexibility of protein structures

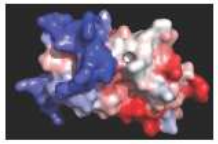


CABS-fold

server for de novo and consensus-based prediction of protein structure

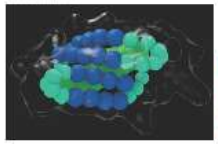
AGGRESCAN3D

server for prediction of aggregation properties of protein structures



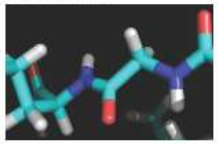
SURPASS

SURPASS coarse-grained protein model of low-resolution



BBQ

program for protein backbone reconstruction from C-alpha coordinates



pyCABS

package for simulations of long-time protein dynamics using CABS reduced model

ClusCo

a software for GPU/CPU clustering and comparison of protein models