School on Molecular Modeling for Life Science

Pula, Sardinia, Italy
June 6th – 10th 2016

Role, relevance and limitations of homology modeling in biomolecular sciences

Allegra Via
Sapienza Università di Roma, Italy
IBBE, National Research Council, Italy
ELIXIR Italy

IBBE
Istituto di Biomembrane e Bioenergetica

http://bioinformaticstraining.pythonanywhere.com/

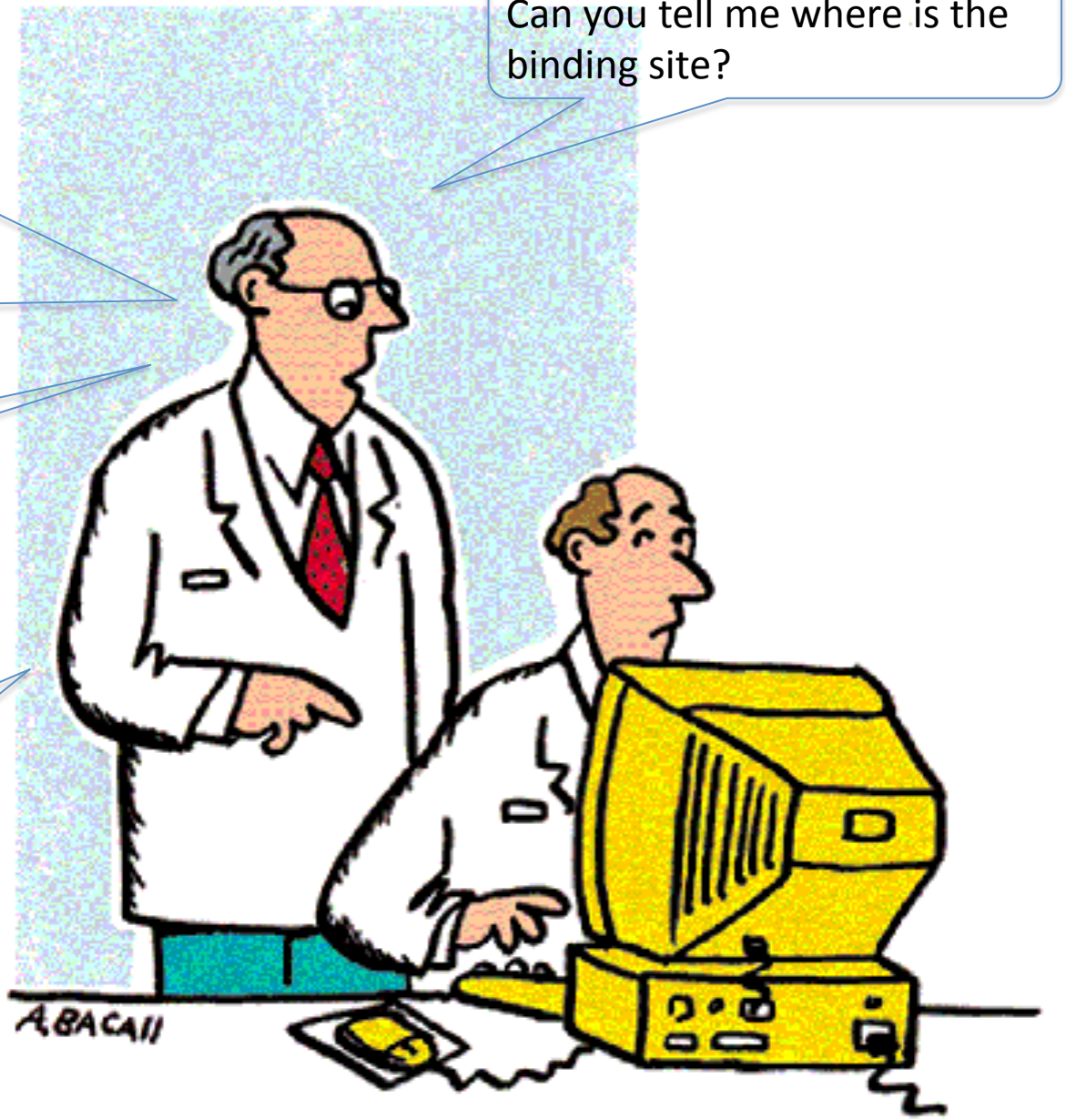
http://www.123rf.com

I want to run site-directed experiments in order to identify functional residue(s). Do I have to mutate all the residues of my protein?

Can you tell me where is the binding site?

Which serine residues in my protein are likely to be phosphorylated?

Which residues are likely to be hotspots?



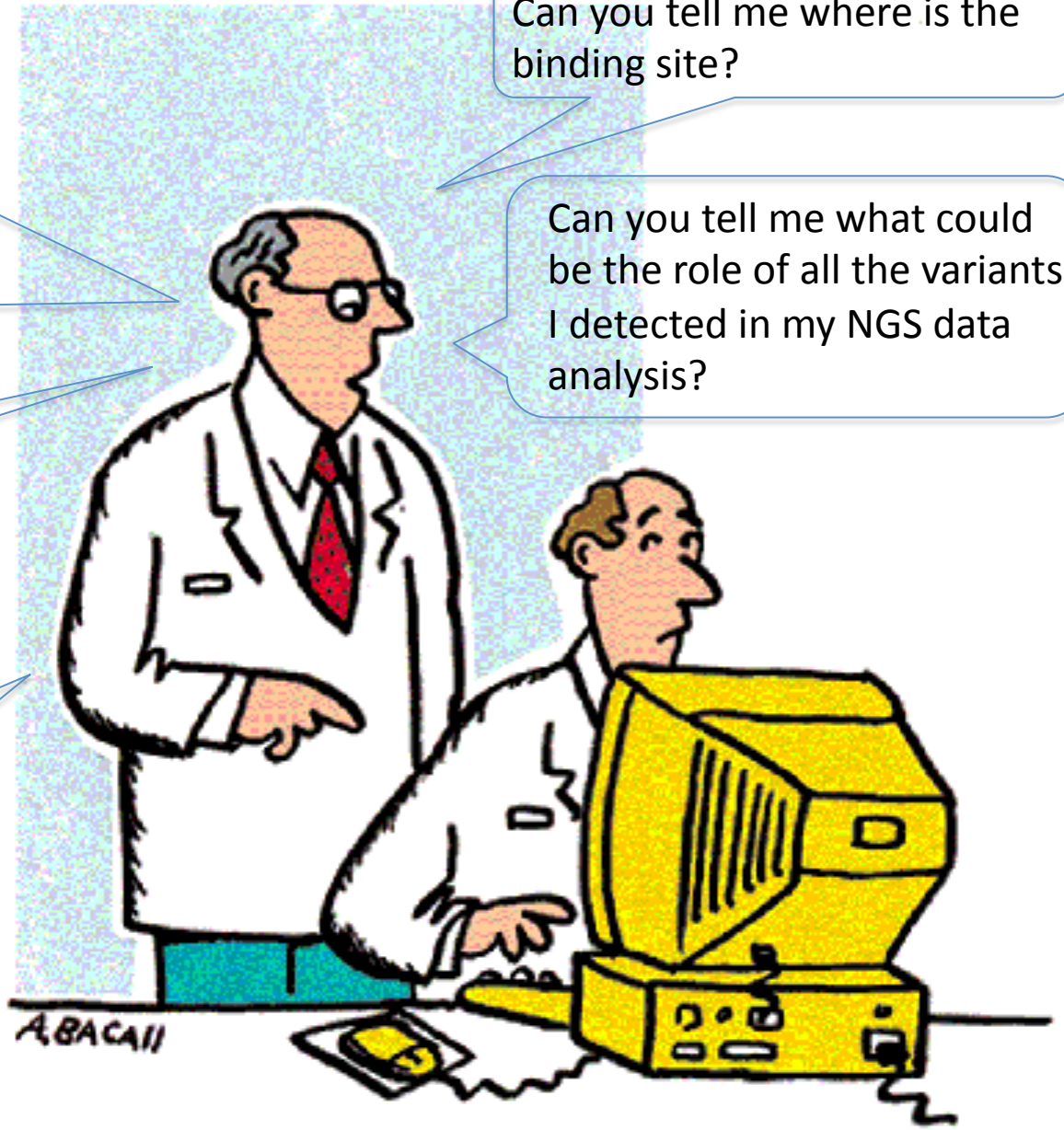
I want to run site-directed experiments in order to identify functional residue(s). Do I have to mutate all the residues of my protein?

Can you tell me where is the binding site?

Can you tell me what could be the role of all the variants I detected in my NGS data analysis?

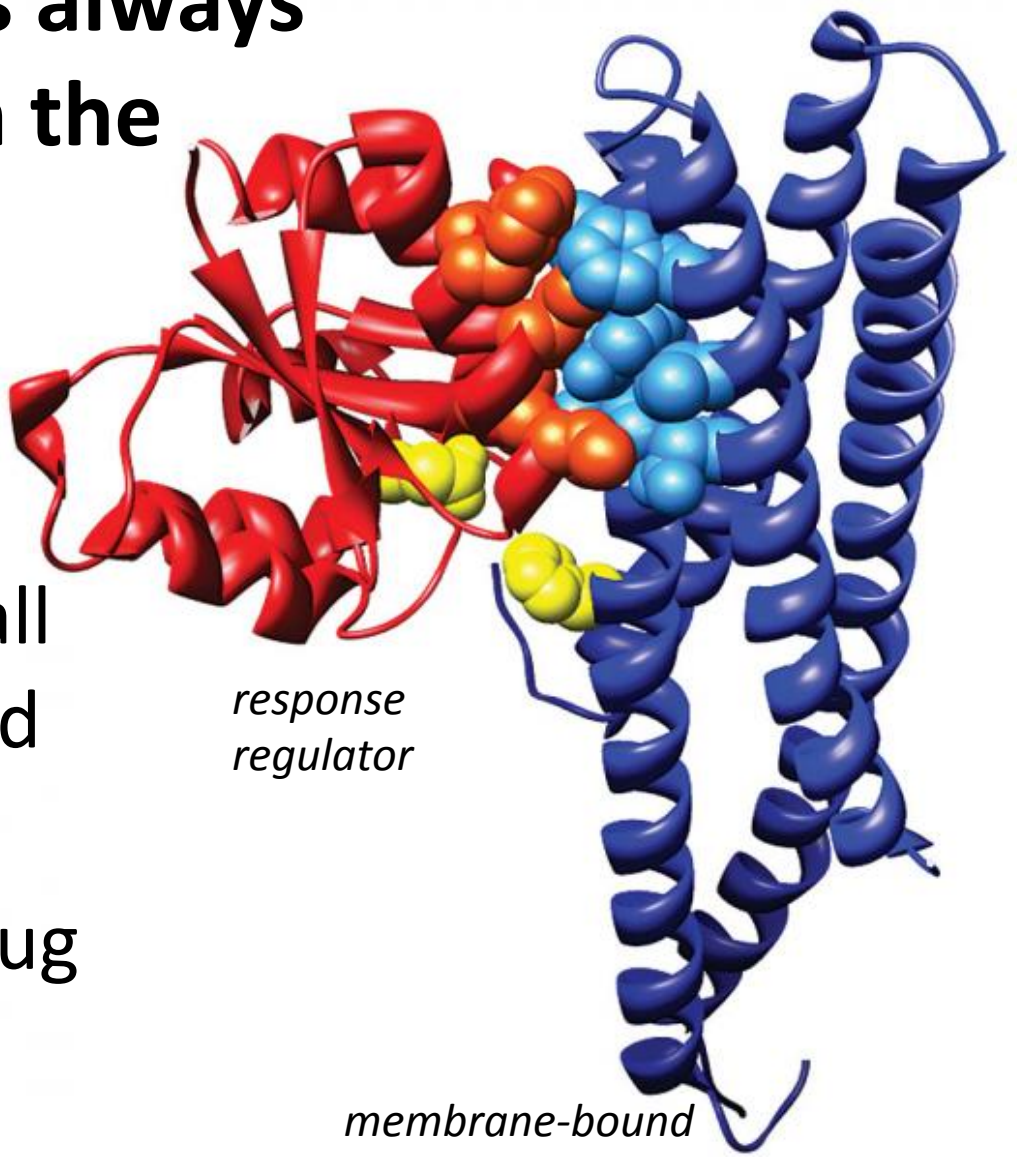
Which serine residues in my protein are likely to be phosphorylated?

Which residues are likely to be hotspots?



A protein structure is always of great assistance in the study of:

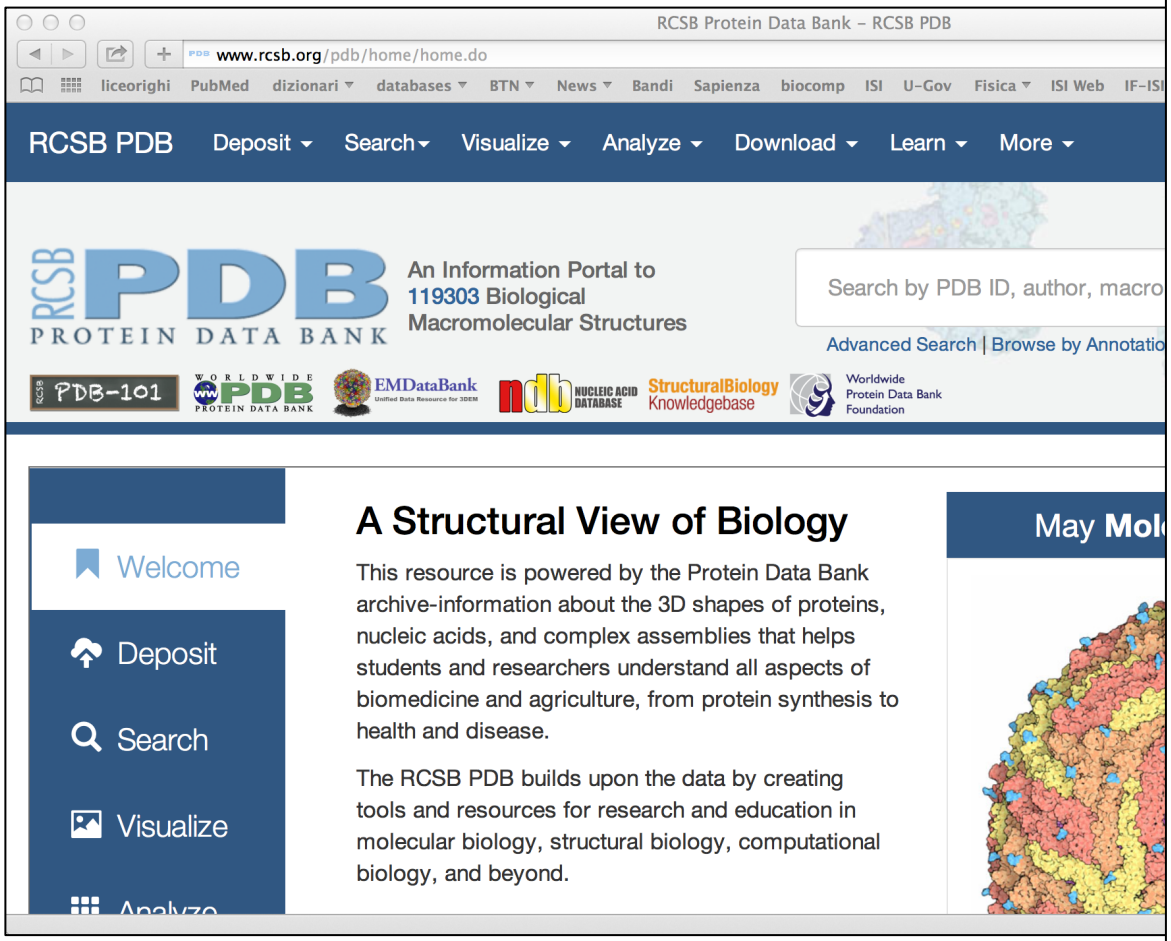
- protein function
- protein dynamics
- interactions with small ligands, DNA/RNA and other proteins
- in structure-based drug discovery and drug design

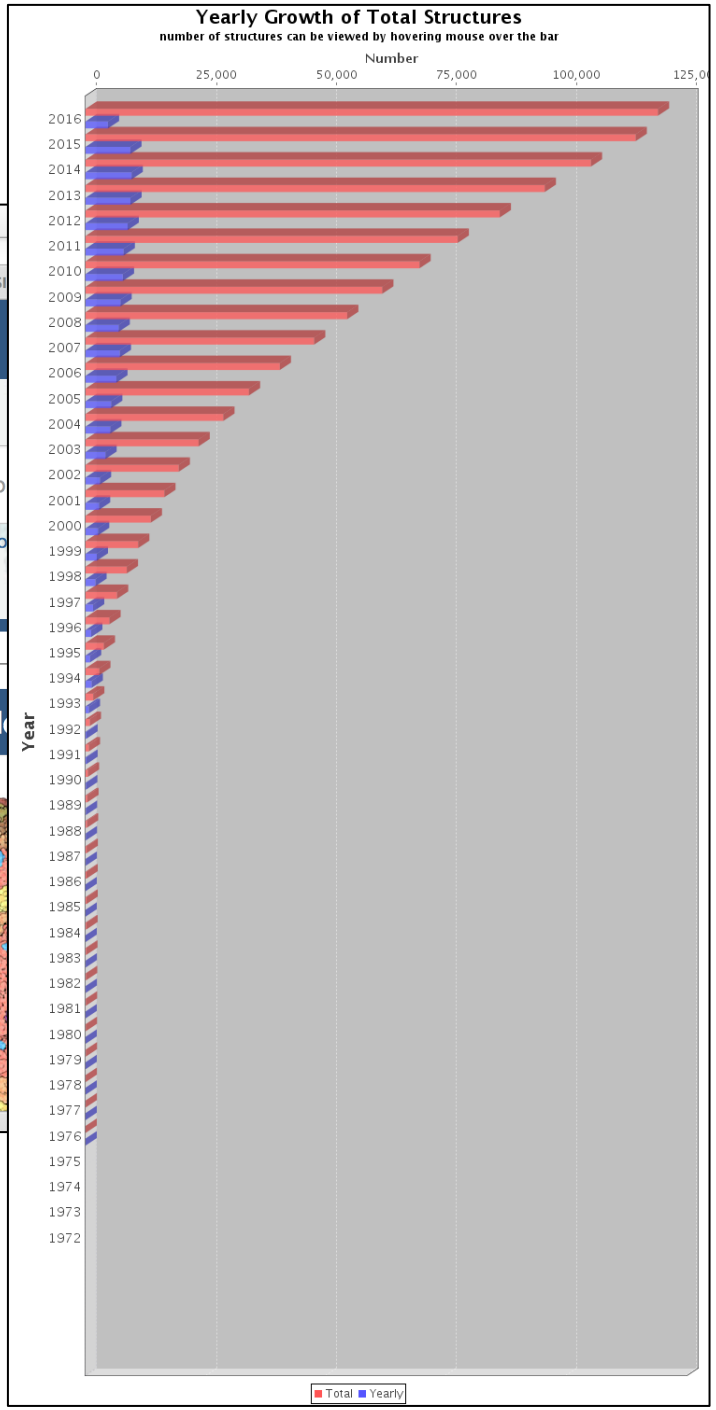


response regulator

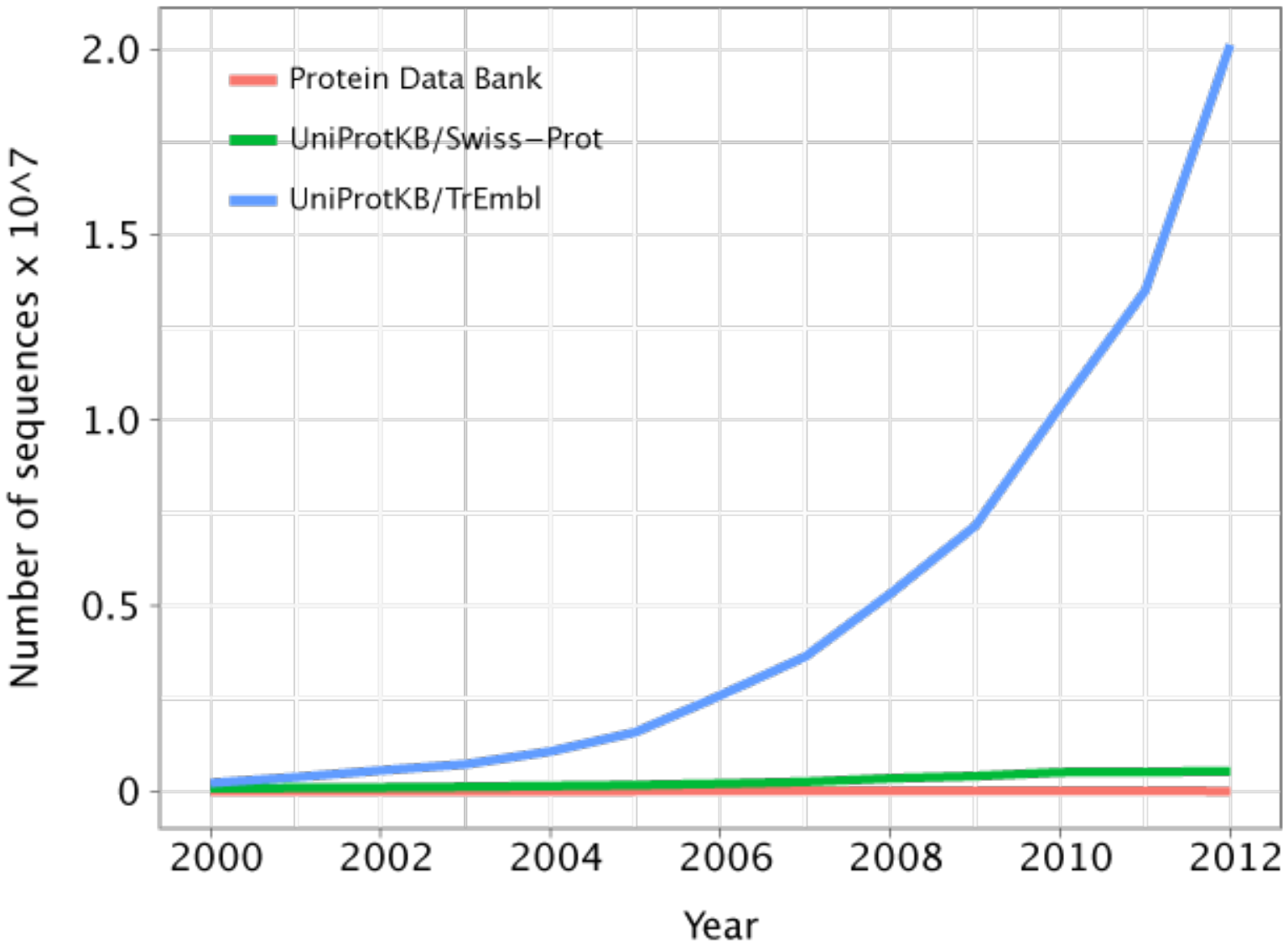
membrane-bound sensor histidine kinase

The PDB is continuously growing



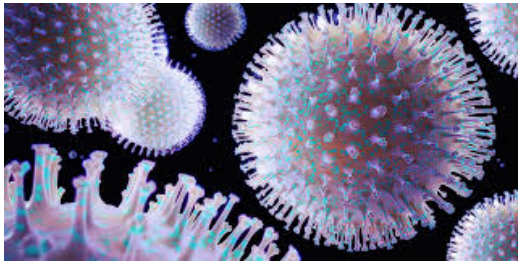


Last update: *Tuesday May 31, 2016 at 5 PM PDT*



Given the drop in the cost of the genome sequencing techniques, this gap is only destined to grow.

Entire genomes



> 75,000 viruses

> 350 fungi

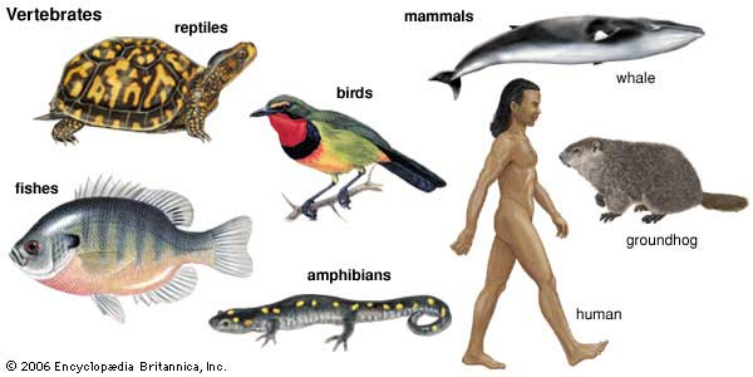
> 100 invertebrate

> 360 archea

> 150 plants

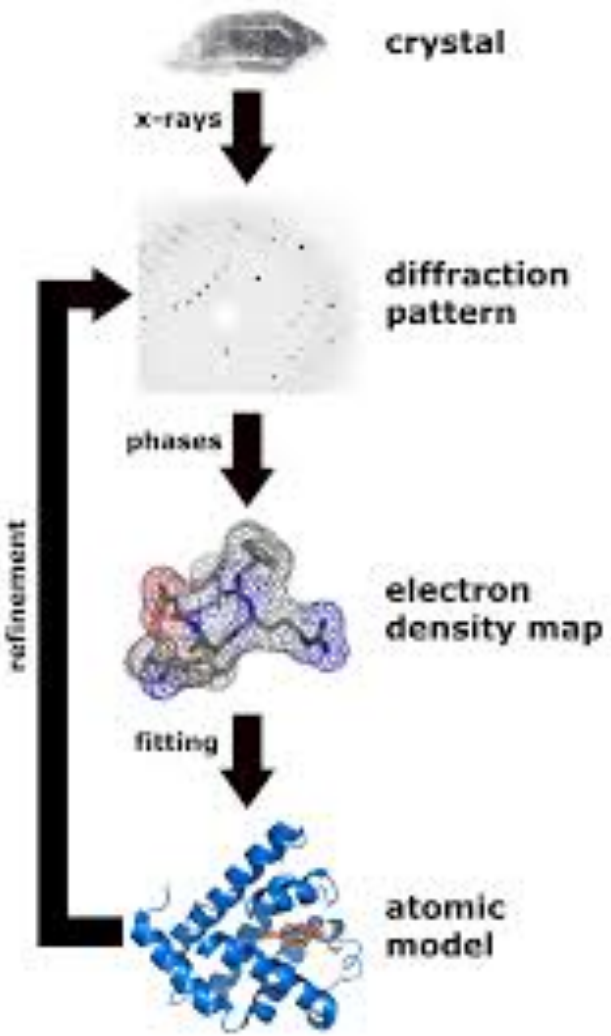
> 17,000 bacteria

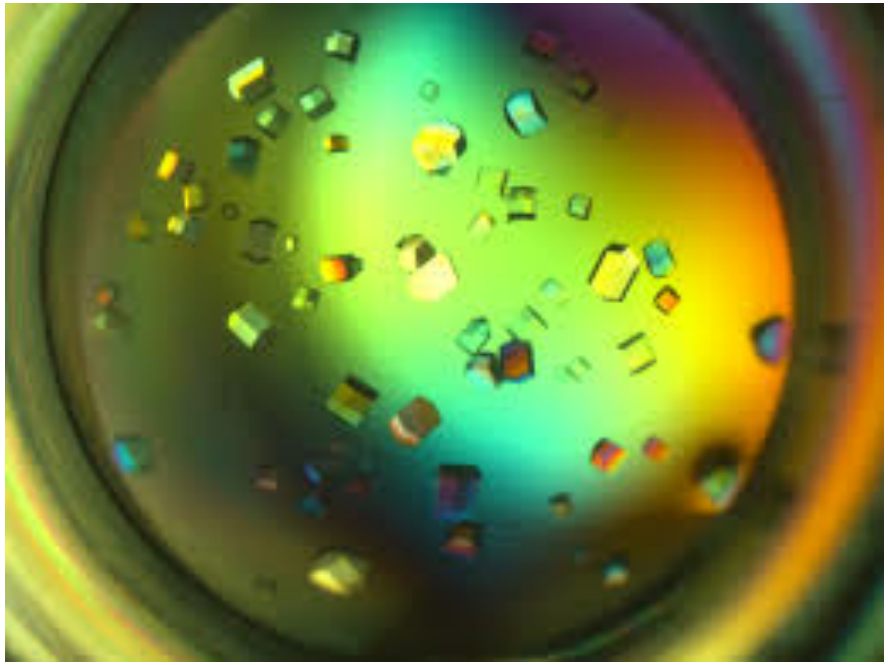
> 100 insects



> 230 vertebrates
(80 mammalian)

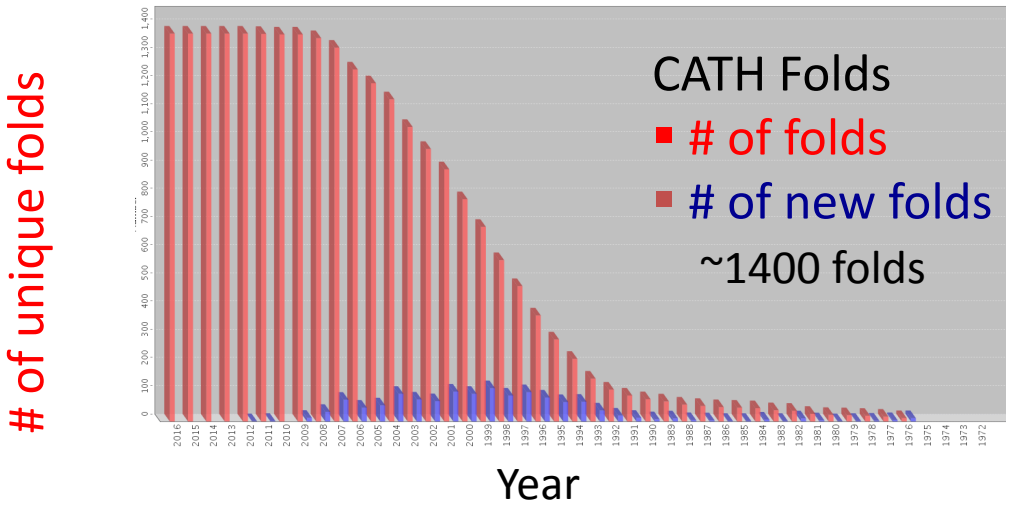
X-ray crystallography….



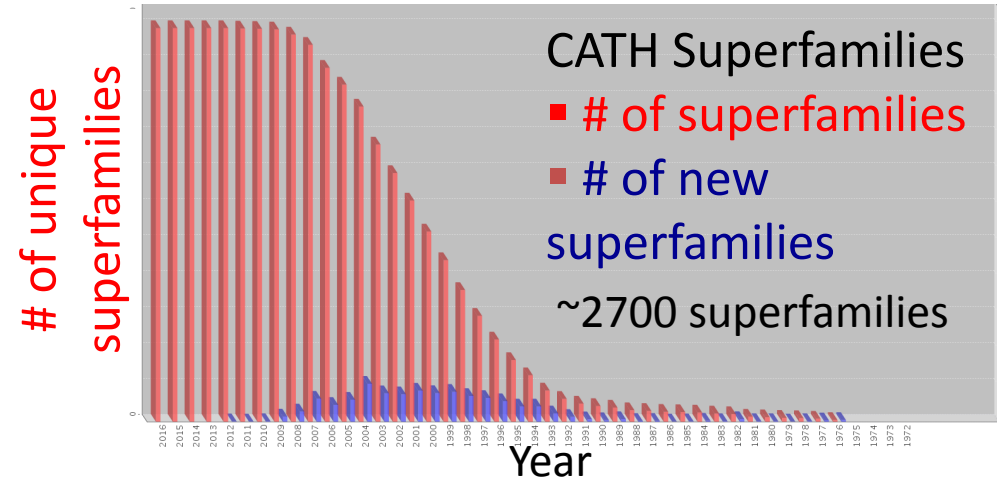


- The protein sample must be pure
- The protein sample must "be able" to crystallise
- Several proteins cannot be crystallised (i.e., transmembrane proteins)
- Determining the structure of a protein is usually rather expensive (~$100K per structure)

Few new folds & superfamilies lately -> template available for nearly everyone



Few new folds in the last years!



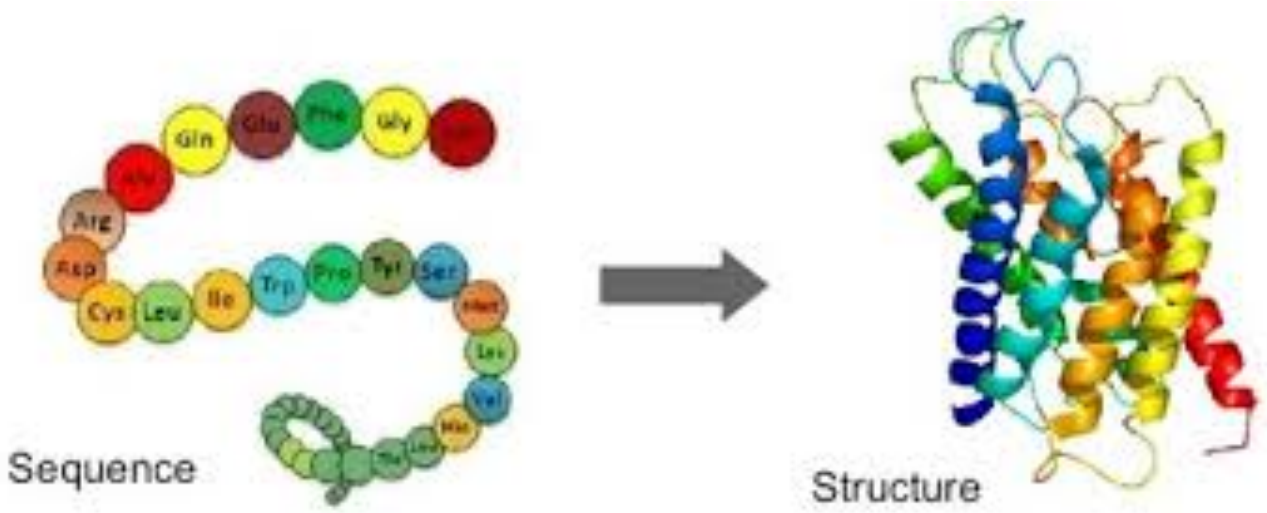
Few new superfamilies in the last years!

- A large number of polypeptides takes a small number of different folds
- At least 50% of the available sequences has a homologue in the PDB

Computational methods…

– Fast (minutes or hours)

– Not expensive (PC)

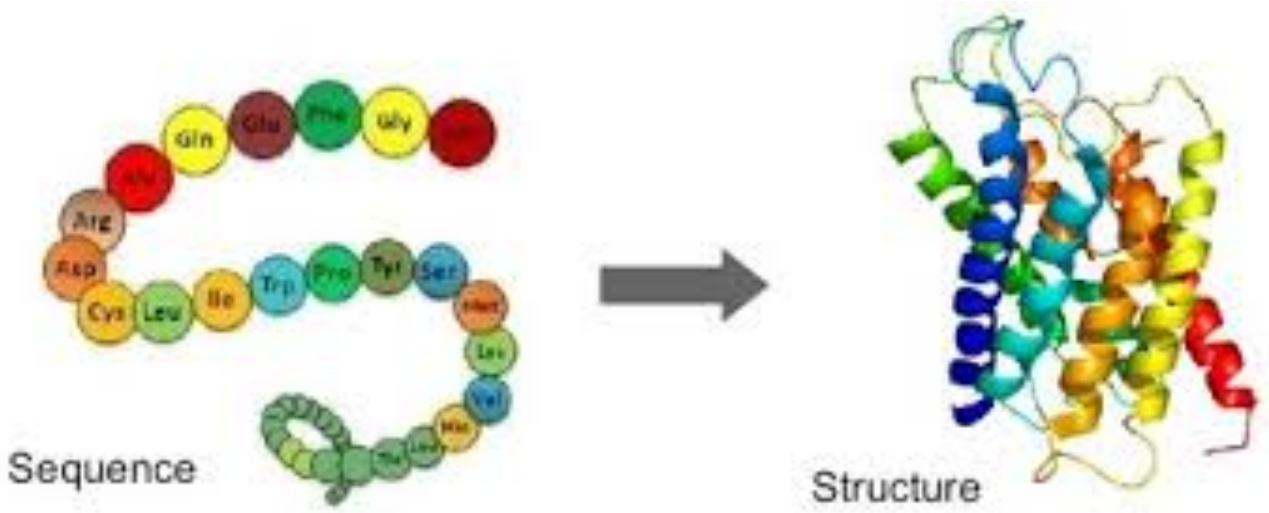


3D structure prediction or modelling

- Comparative or homology modelling
- Threading or fold recognition
- *Ab initio*

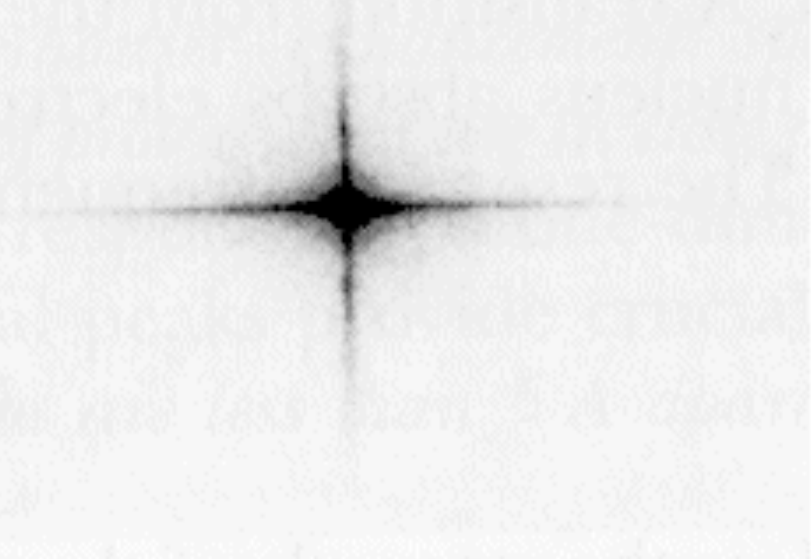
Computational methods…

- Fast (minutes or hours)
- Not expensive (PC)
- In 60% of the cases provide correct solutions
- Solutions are low resolution but often sufficient to study protein function



photograph of Parthenon

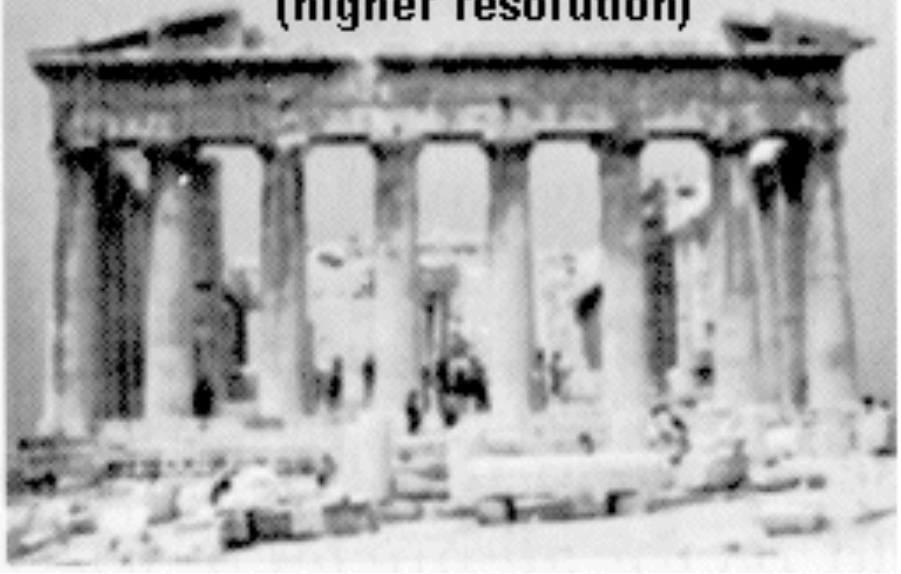
diffraction pattern of Parthenon



http://www.sci.sdsu.edu/TFrey/Bio750/FourierTransforms.html

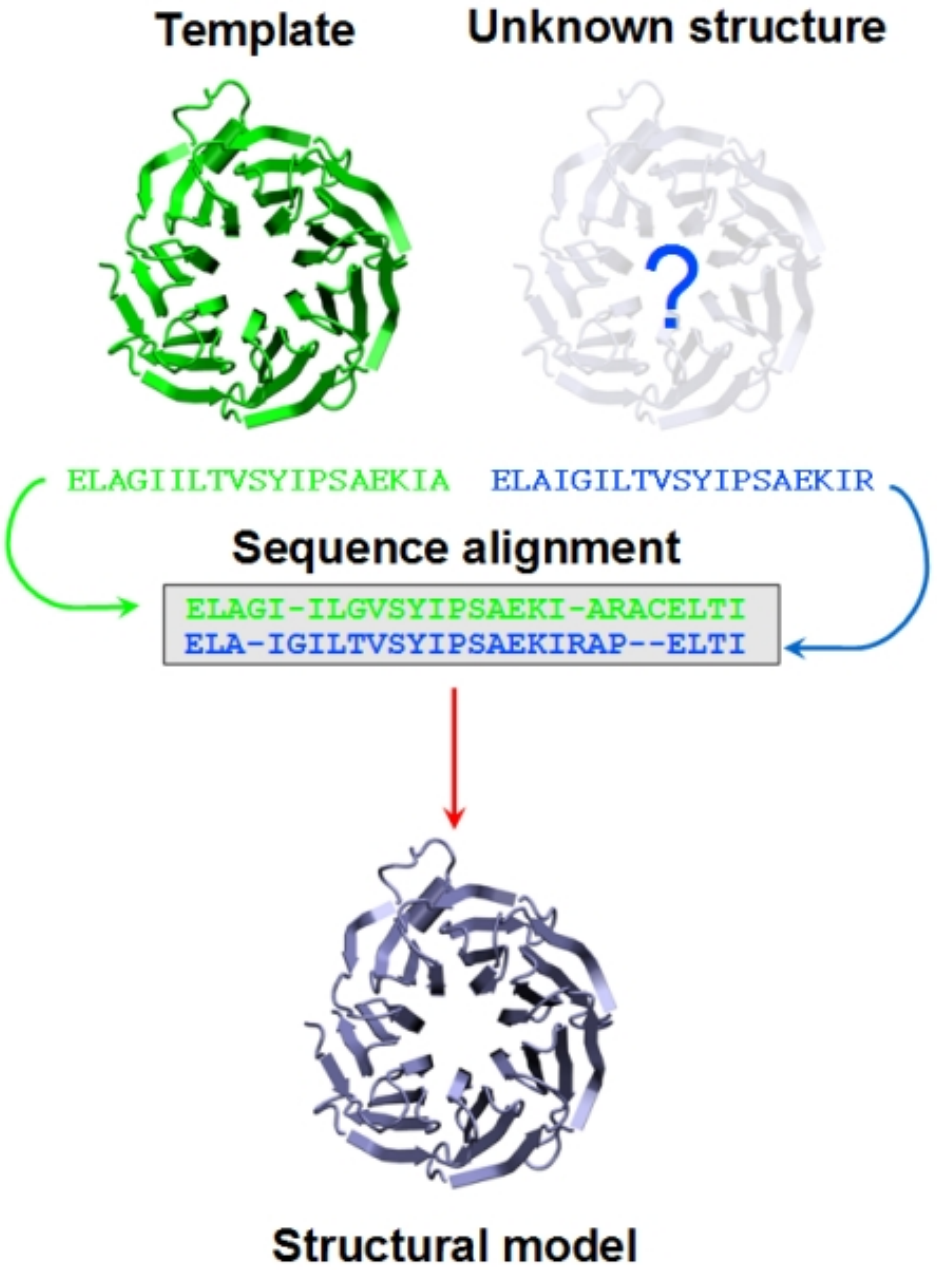
reconstruction from diff. pattern

reconstruction from diff. pattern
(higher resolution)

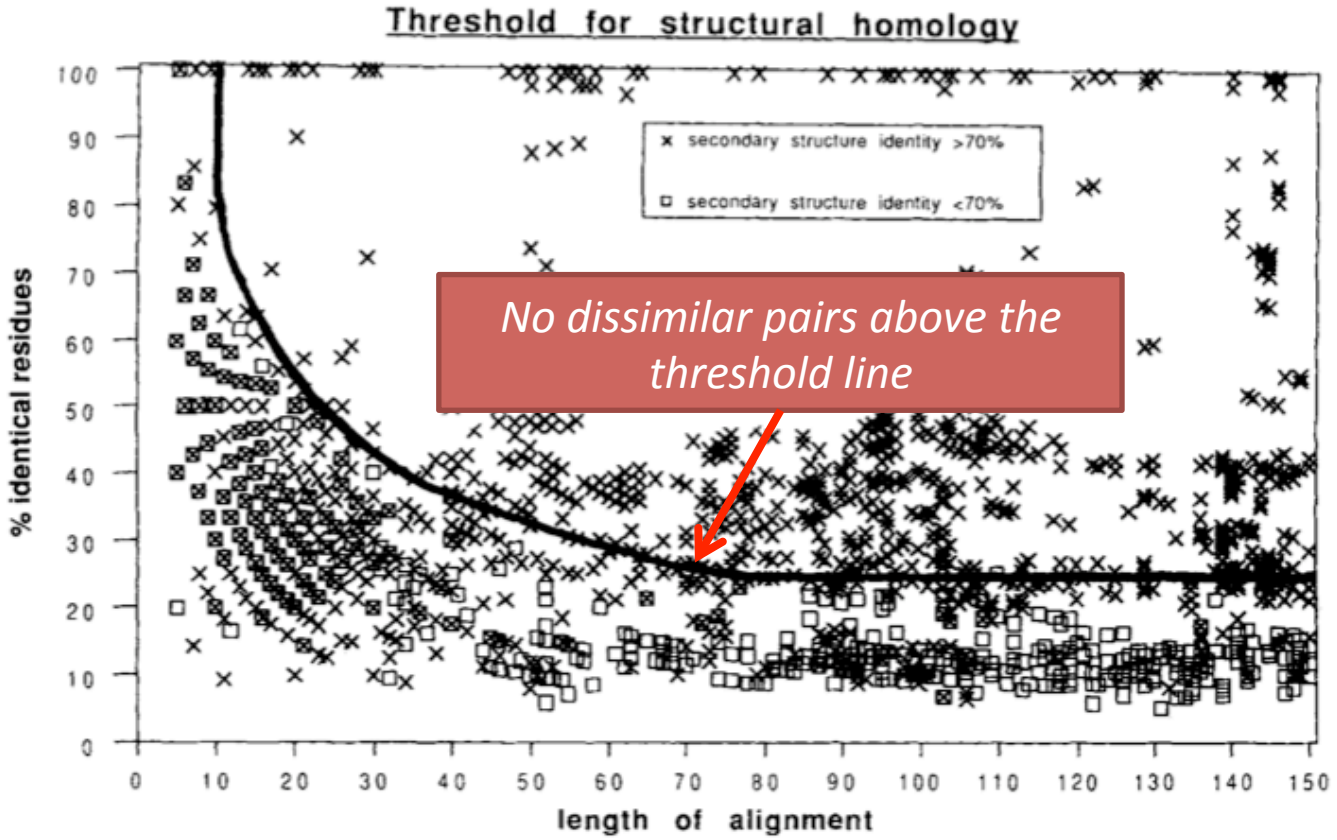


Building by homology

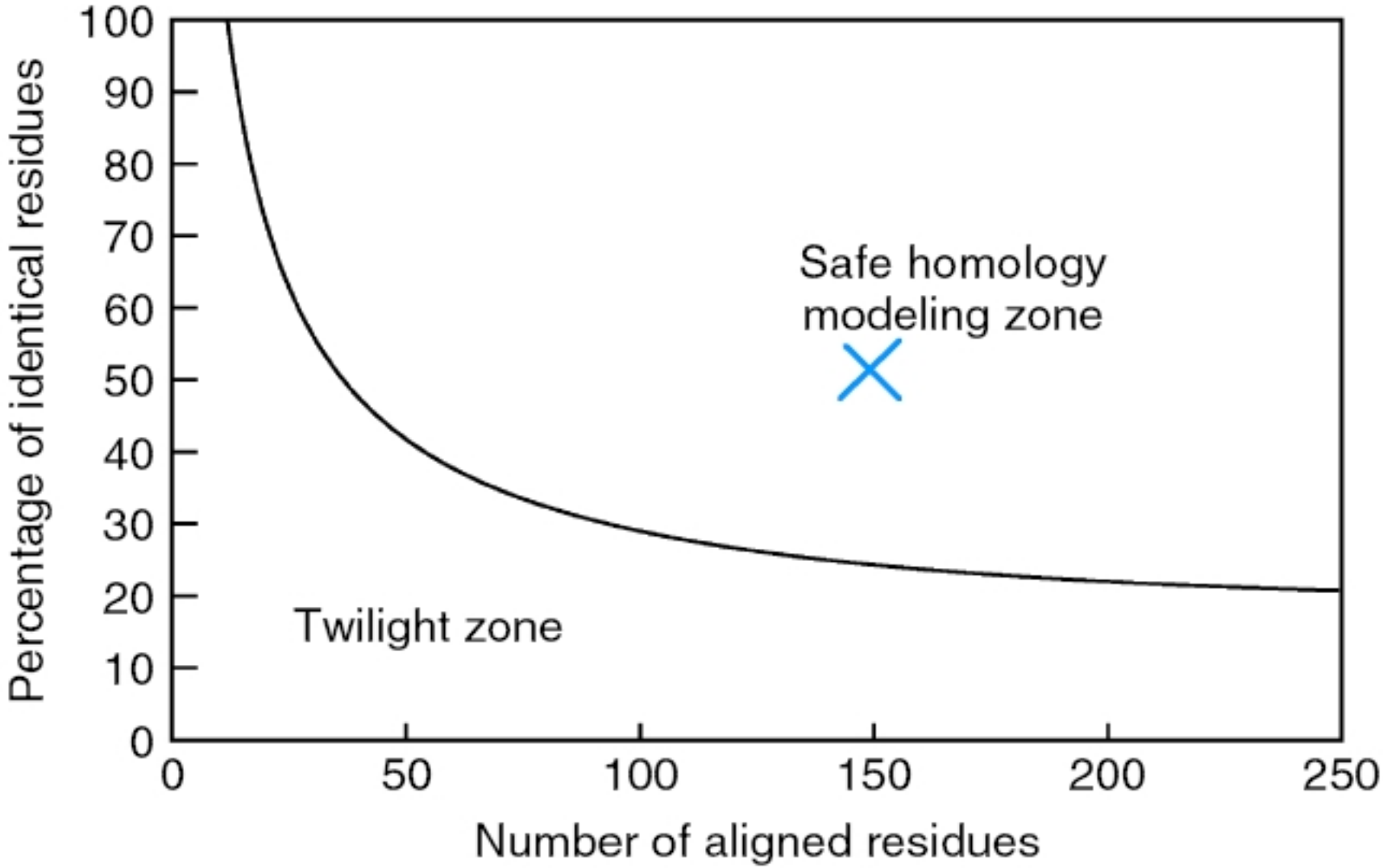
Can provide "low-resolution" structures, which may contain sufficient information about the spatial arrangement of important residues and which may guide the design of new experiments



Sequence-structure identity depends on length of protein



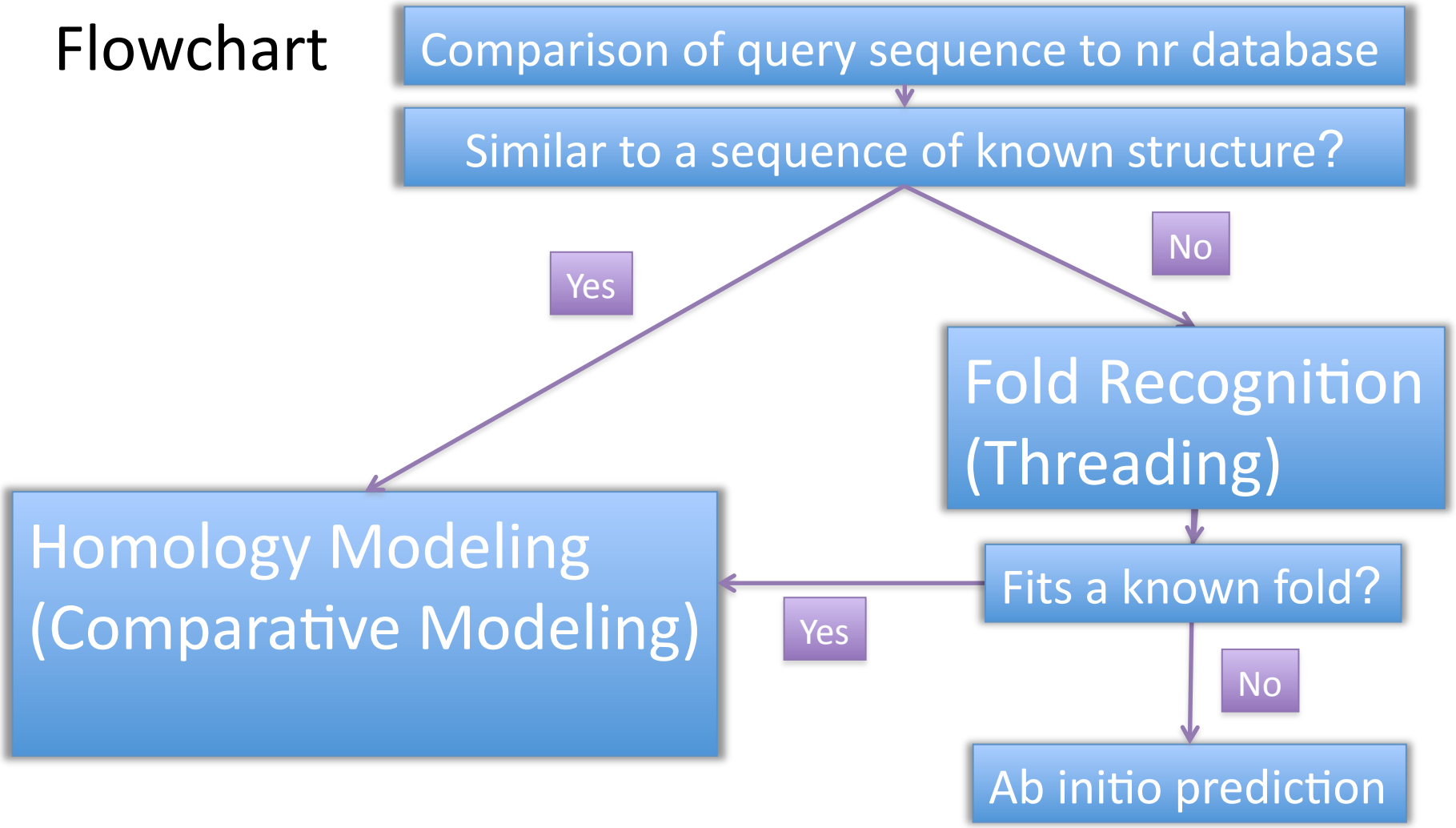
Sander & Schneider, 1991

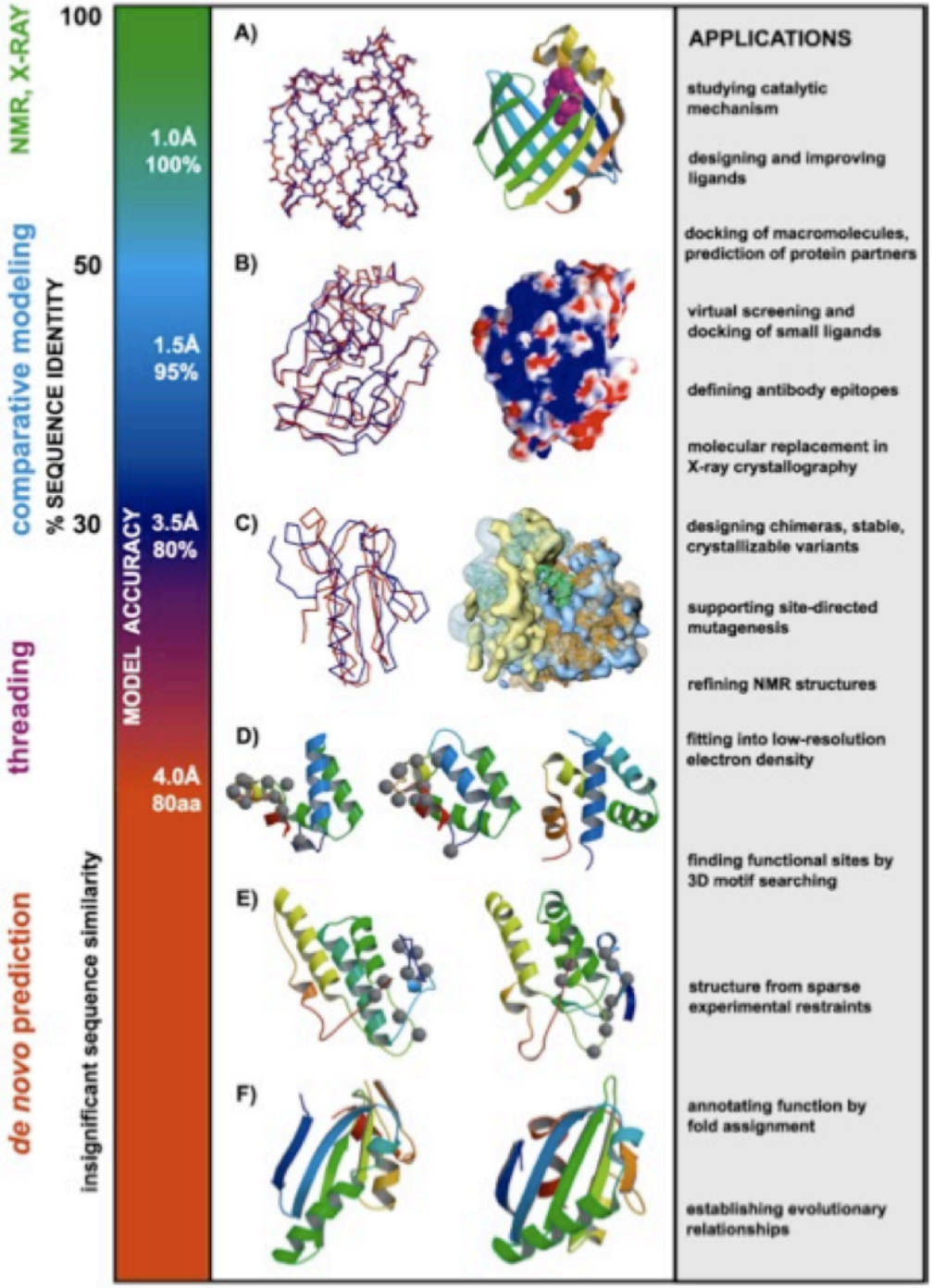

Percentage of identical residues
100
90
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
Number of aligned residues
Safe homology
modeling zone
Twilight zone

Approach	Identity percentage
Homology modelling	> 30%
Threading/fold recognition	0 – 30%
ab initio/de novo	no homologous

Prediction of structure from sequence

Flowchart

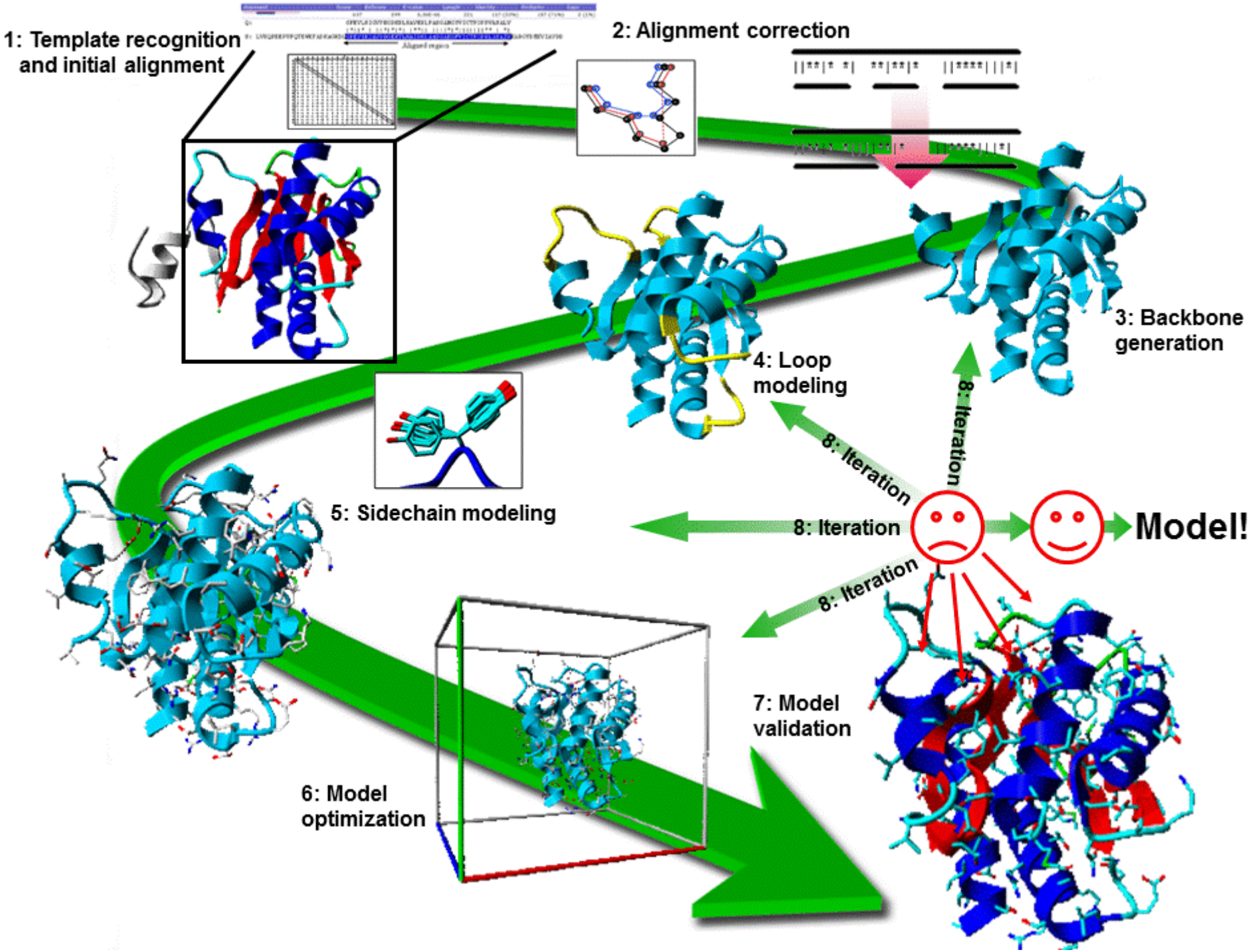


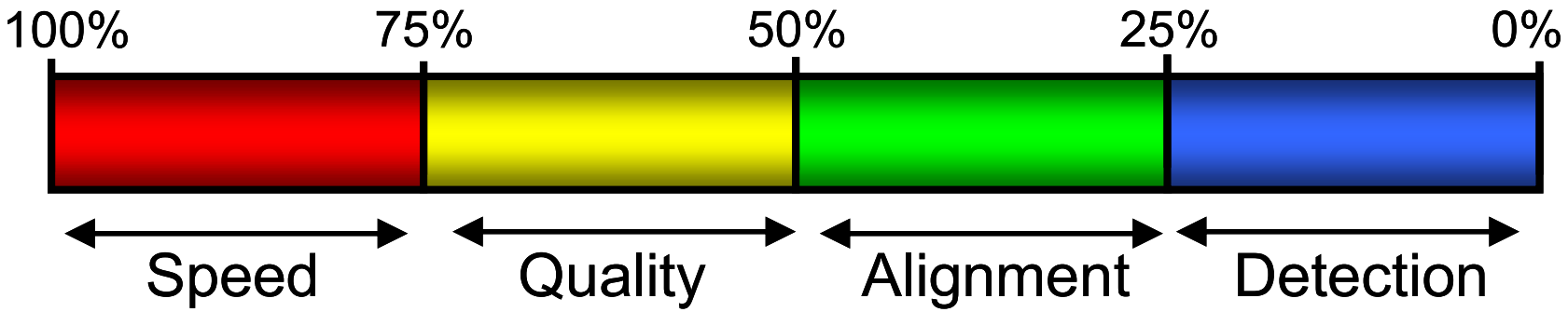

NMR, X-RAY
comparative modeling
threading
de novo prediction
100
50
30
% SEQUENCE IDENTITY
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

- Homology models can be constructed for 25-65% of the amino acids in a genome (→ support structural genomics projects) (Xiang, 2006)
- This value
  - differs between individual genomes
  - increases steadily thanks to the continuous growth of the PDB
- the remaining 75-35% → threading/ab initio/ NMR/X-Ray

Homology modeling is a multi-step process

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side chain modelling
6. Model optimisation
7. Model validation (by hand or using different servers)
8. Iteration to correct mistakes (if any)


1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Model!



The limiting steps in homology modeling as function of percentage sequence identity between the structure and the model. (Figure based on Rodriguez and Vriend, 1997)

Template recognition and initial alignment

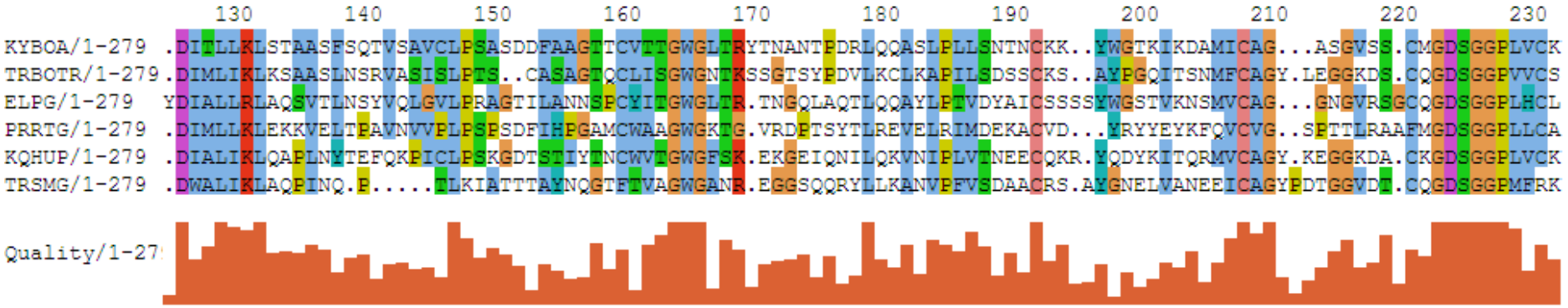
- Given: sequence
- Wanted:
  - structural template
  - sequence-structure alignment
- Easiest approach:
  - Blast / Psiblast against sequences with known structure
  - select template based on sequence identity
    - >70%: straight forward
    - ~40-50%: usually clear
    - lower seqid: alignment is a challenge
- State of the art protocols include
  - more sophisticated searches
  - additional information for improved template selection
  - Profile-profile comparison (HHSEARCH)
  - Seq-structure compatibility (Threading: RAPTOR)

Alignment step is critical

The alignment determines the quality of the model

Search for the highest number of homologues in the PDB

If PSSM or HMM are used in the search, they will also include sequences with no structure



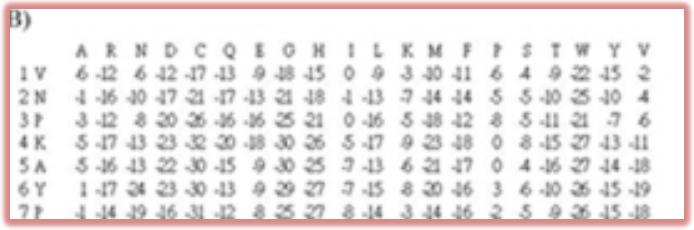
Build an accurate multiple alignment between the target sequence and the sequences of the templates

Sequence-sequence alignment

- Information content:

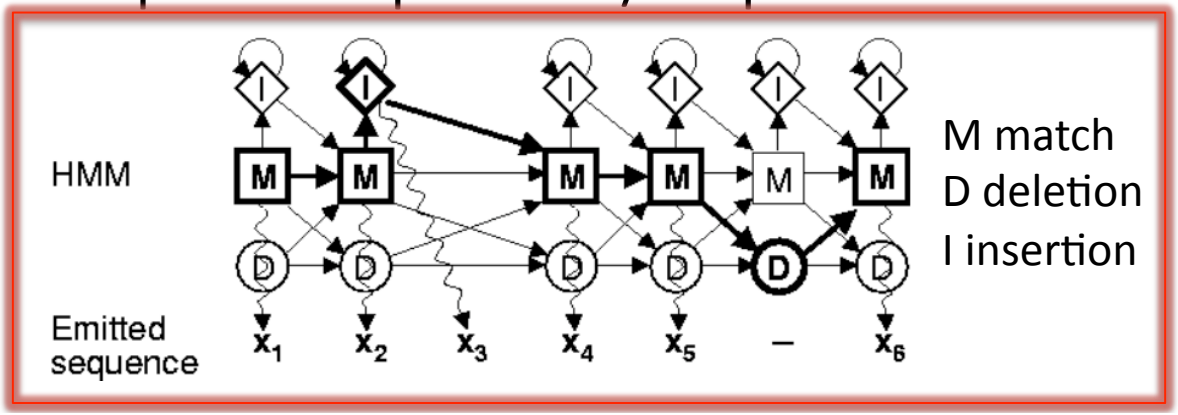
1. Sequence
2. Profile (Position specific scoring matrix -PSSM)

   aa preferences for each position

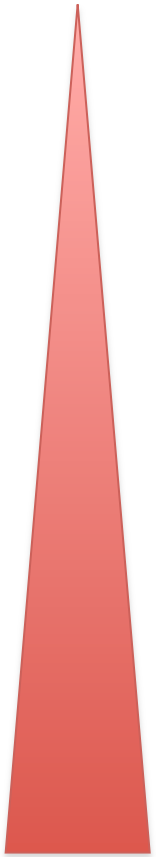


3. Hidden Markov Models (HMM)

   Contains in addition position-specific in/del penalties



Sequence-sequence alignment

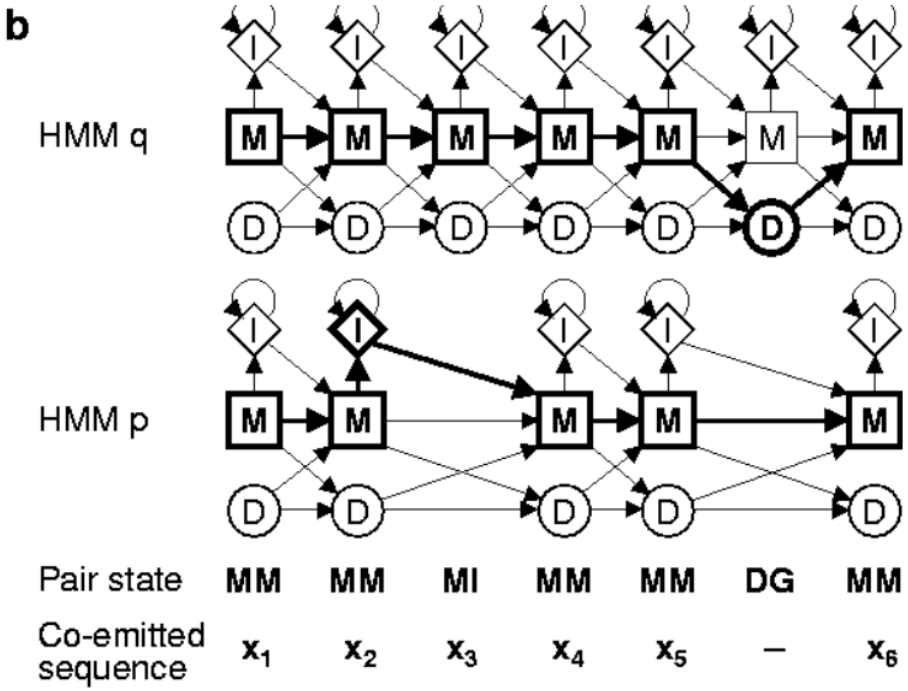


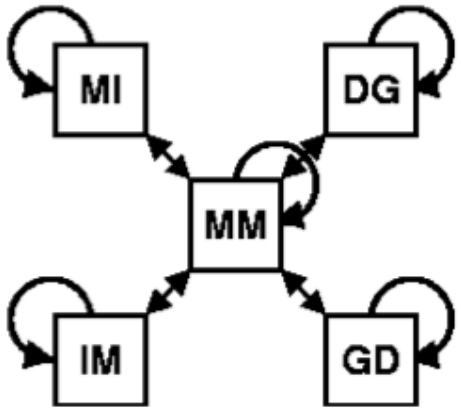
- Information content:
  - Sequence-sequence comparison
    - e.g. BLAST
  - Profile-sequence comparison
    - e.g. PSI-BLAST
  - Profile-profile comparison
    - e.g. LAMA, PROF_SIM, COMPASS
  - HMM-HMM comparison
    - e.g. HHSEARCH

More information – increased sensitivity in detecting template

HHSEARCH: HMM-HMM alignment

- Formalization:





- more sensitive (for hard cases with <20% seqid) than:
  - Profile-profile comparison
  - Profile-sequence comparison
  - Sequence-sequence comparison

** Söding. Protein homology detection by HMM-HMM comparison. Bioinformatics (2005) 21: 951*

HHSEARCH includes structural information about template

Include ***secondary structure preference*** in model:

•Score pairs of aligned secondary structure elements with substitution matrix

- Query sequence:

  Predicted secondary structure

  (PSIPRED: H/E/C) with confidence [0..9]
- Structural template:

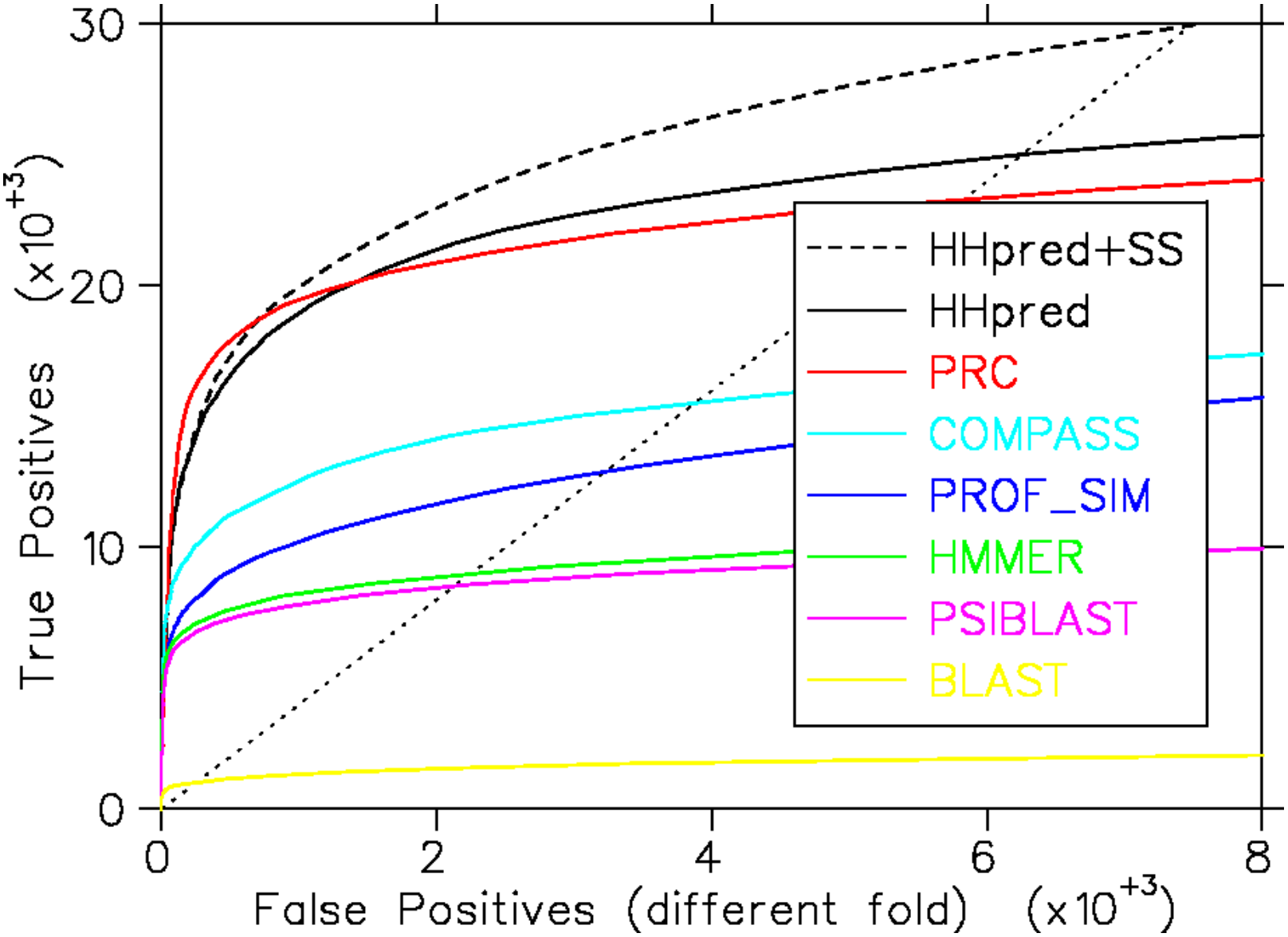
  Secondary structure

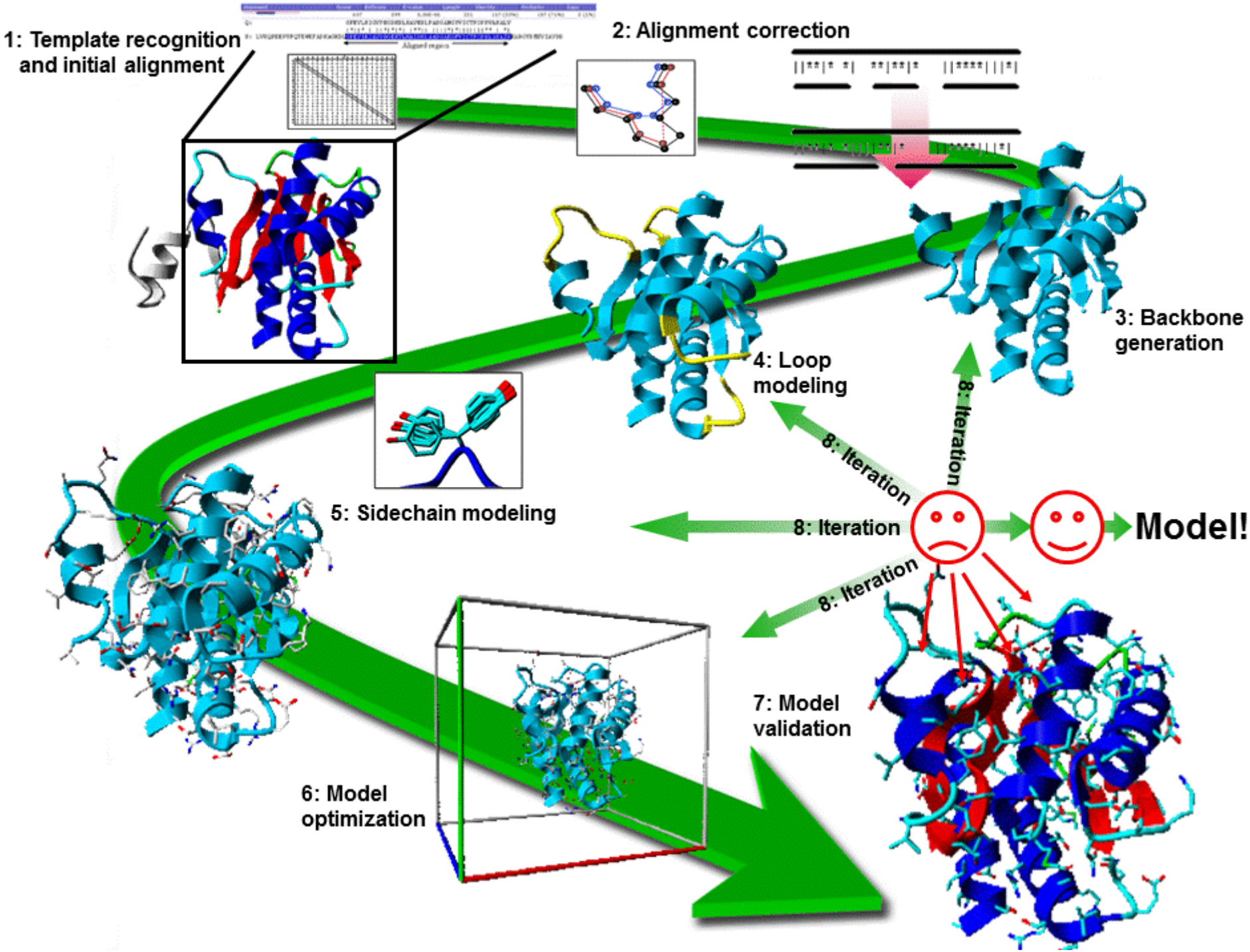
  (DSSP: H/E/B/G/I/T/S)

DSSP:
H = alpha helix
E = extended strand

B = residue in isolated beta-bridge
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
T = hydrogen bonded turn
S = bend

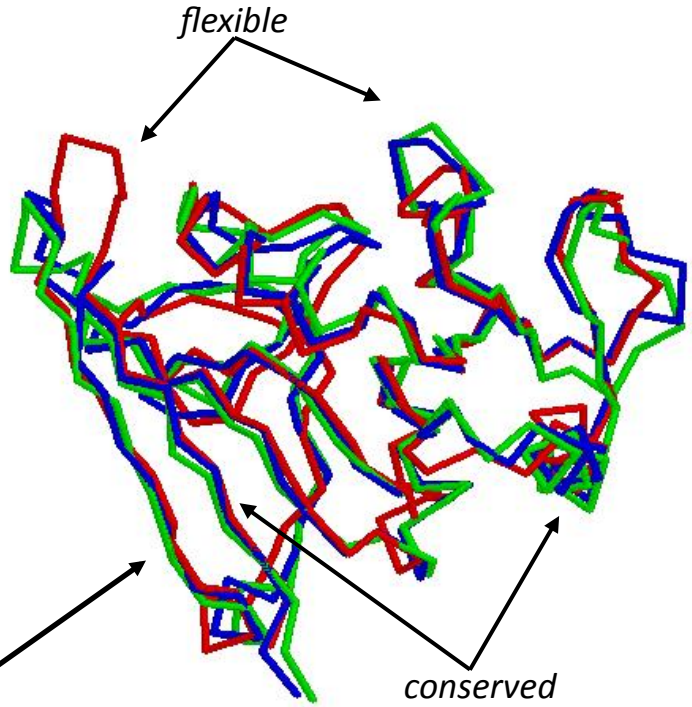
* Söding. Protein homology detection by HMM-HMM comparison. Bioinformatics (2005) 21: 951

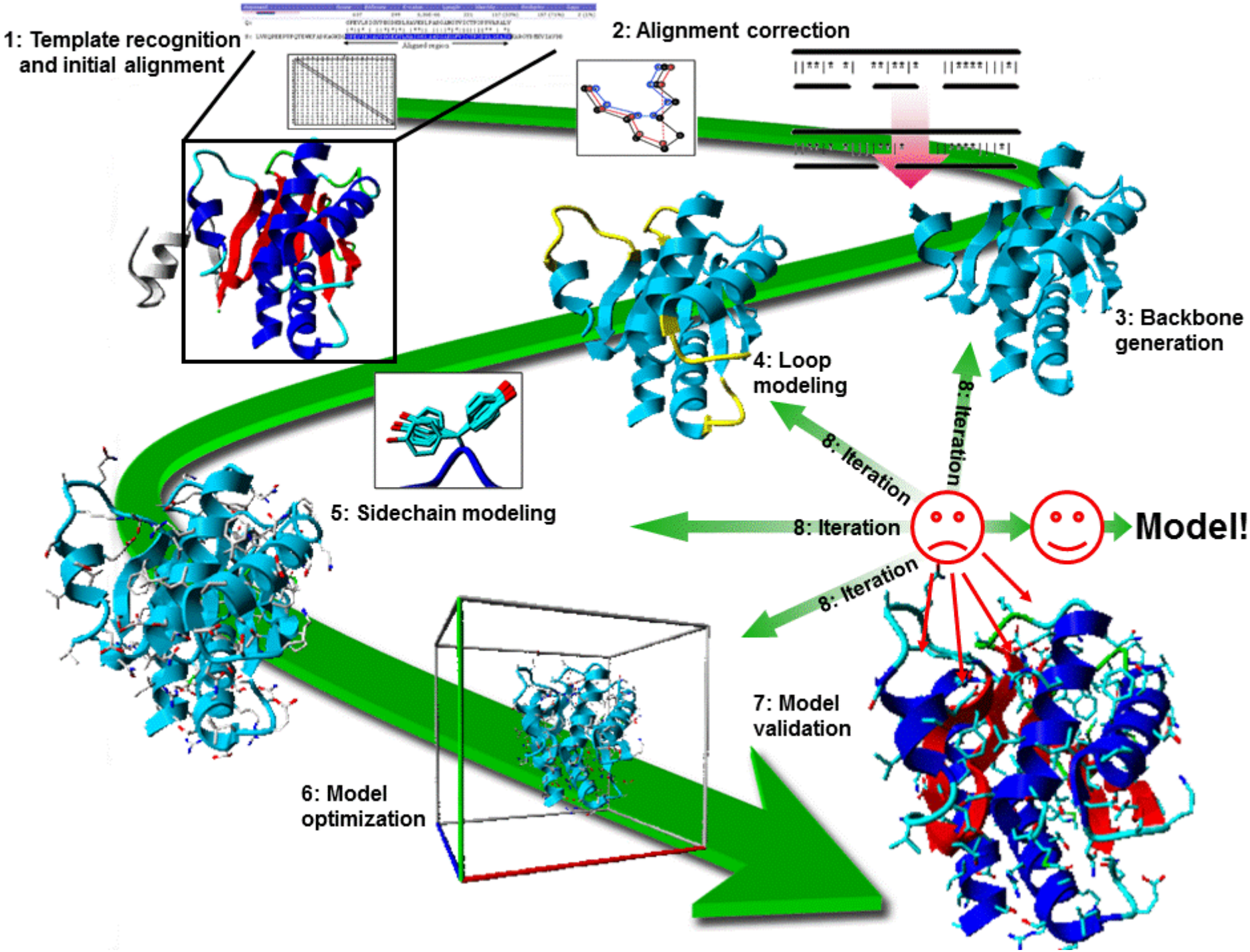

30
20
10
0
True Positives (x10+3)
0
2
4
6
8
False Positives (different fold) (x10+3)
HHpred+SS
HHpred
PRC
COMPASS
PROF_SIM
HMMER
PSIBLAST
BLAST


1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Model!

Building the pre-model (backbone generation)

- The template structure is used as a mould
- Secondary structure regions are identified
- Copy aligned regions from template
- Variable regions (loops) cannot be copied



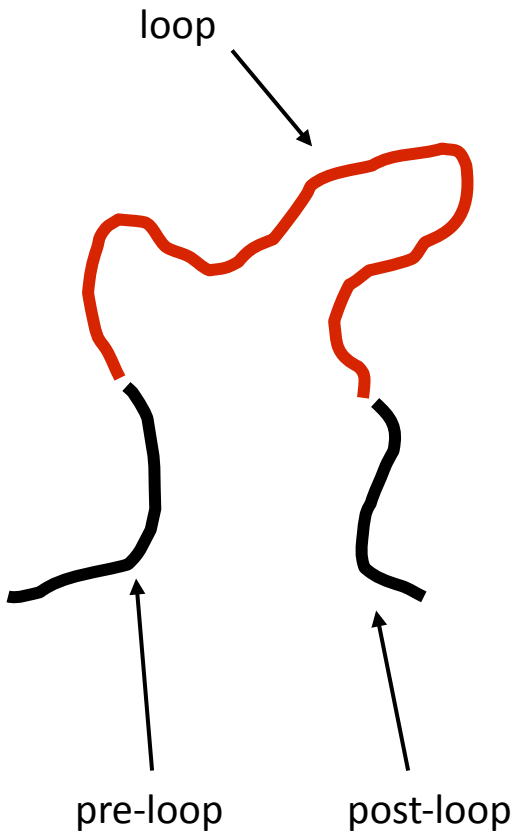

1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Model!

Loop modelling

- The pre-model might lack entire fragments of backbone
  - –not conserved in the protein family
  - –insertions
  - –deletions

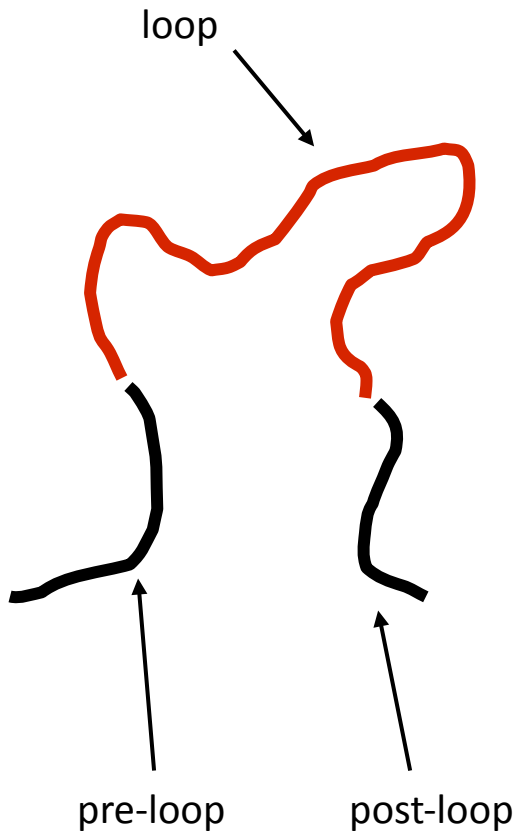
- Input:
  - –2 anchors (pre-and post-loop)
  - –Length k of missing residues

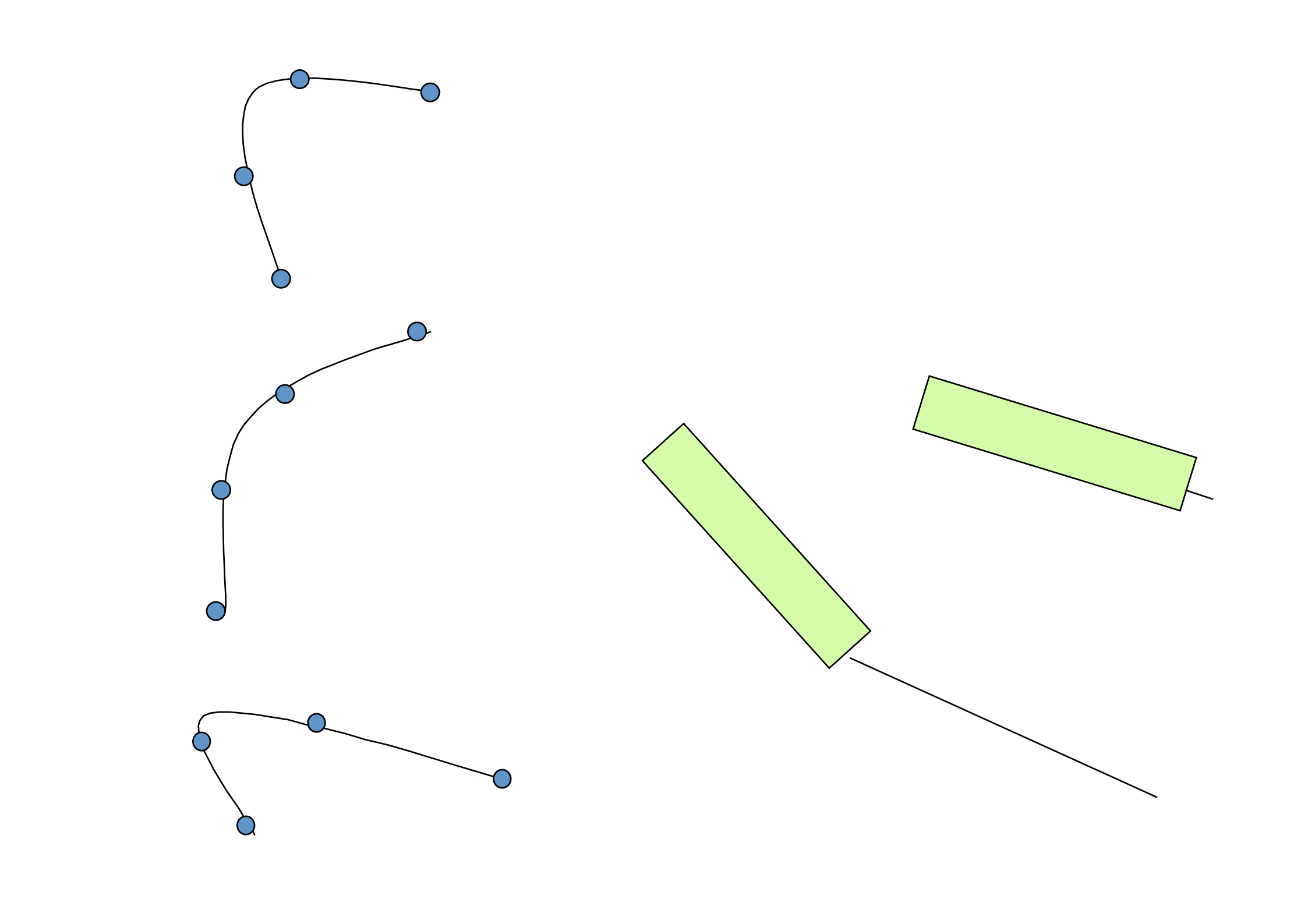
- Problem description:
  - –Search a fold of k residues linking the N- and C-term anchors

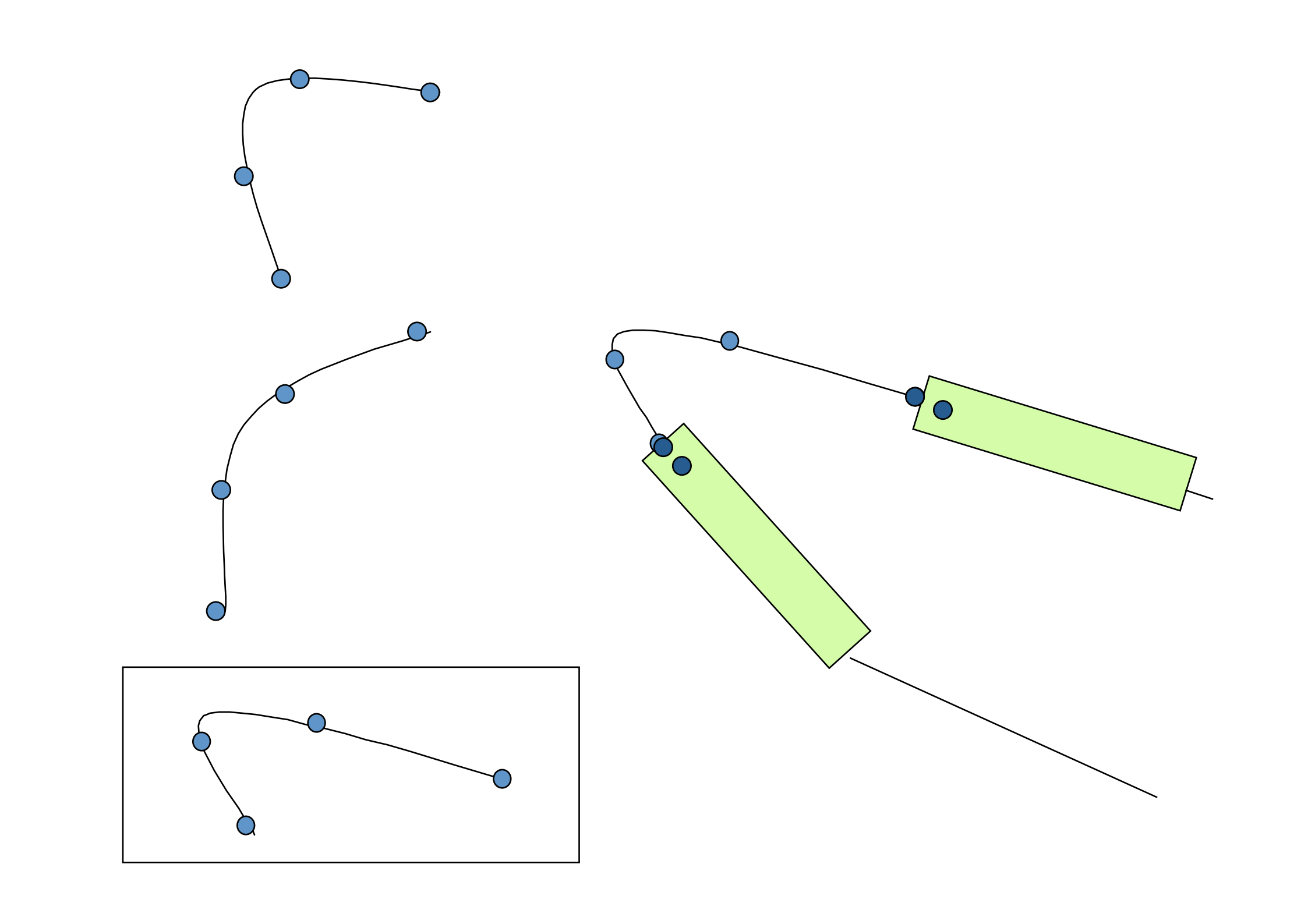


Loop modelling

- Loop libraries
  - –Extract fragments from the PDB
  - –Pick the one that best fullfils geometric restraints
  - –Good for short loops



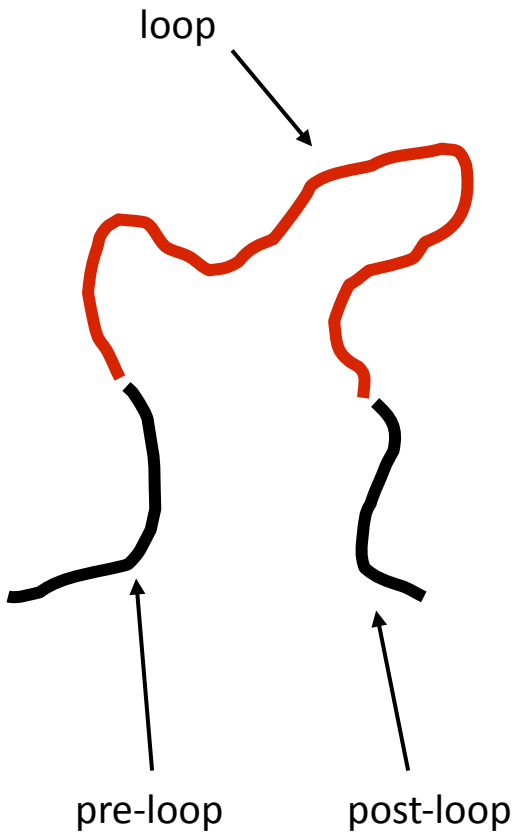


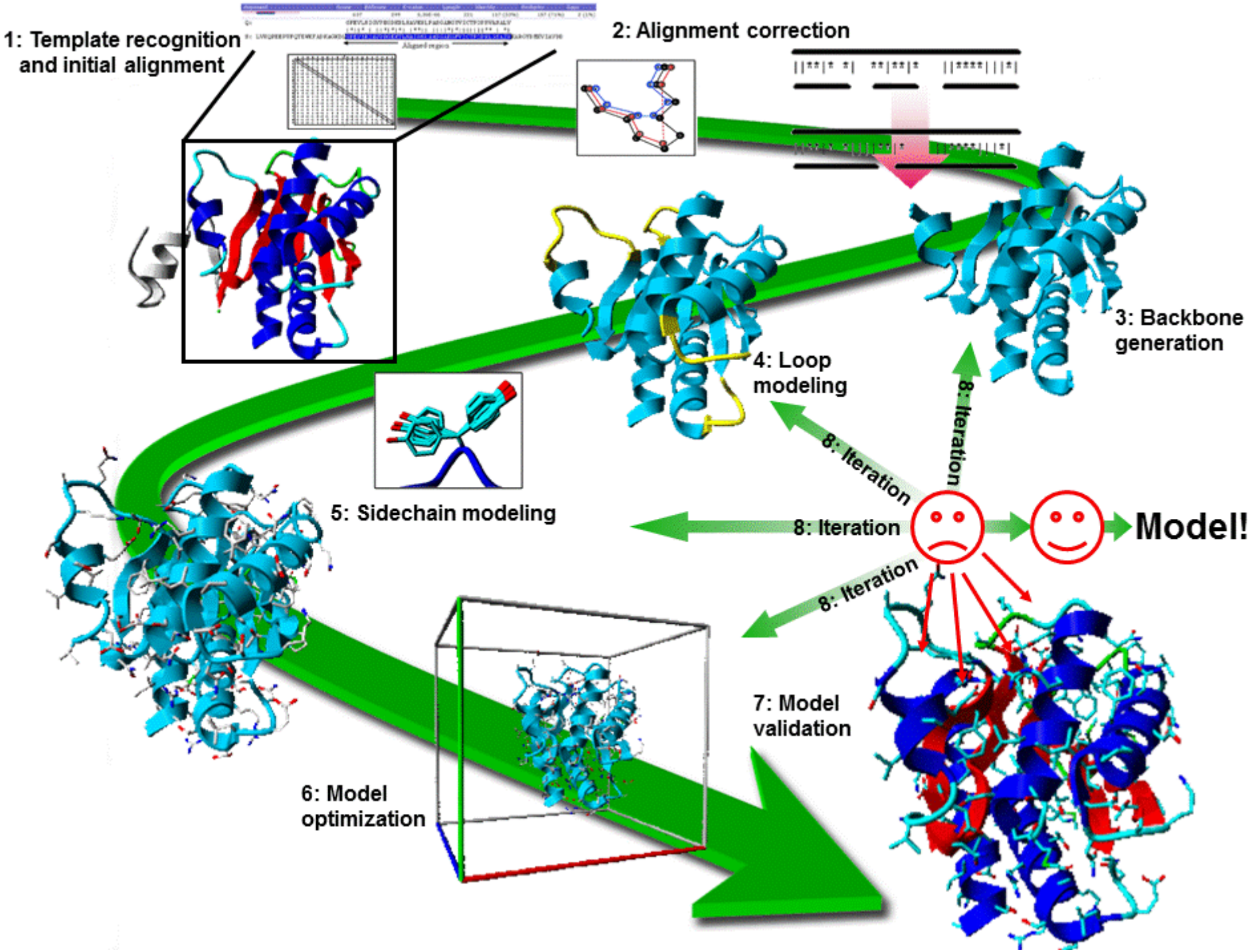


Loop modelling

- Loop libraries
  - –Extract fragments from the PDB
  - –Pick the one that best fullfils geometric restraints
  - –Good for short loops

- *ab initio* methods/loop closure algorithms
  - –Sample the loop conformational space with restraints (backbone dihedral angles)
  - –Apply a loop closure algorithm
  - –good for longer loop




1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Model!

Side-chain modelling

- Corresponding residues virtually retain the same rotameric state (Ponder and Richards 1987, Benedetti et al. 1983)
- Where possible, it is advised to use template side chains

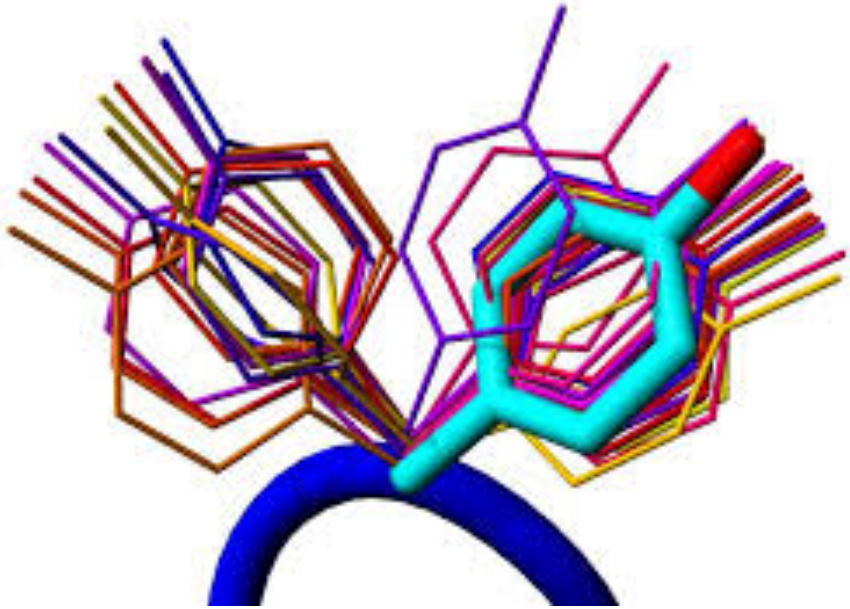
- With two rotable backbone bonds per residue, its very difficult to find the best conformation of a side chain
- Fortunately, statstical studies show side chains adopt only a small number of many possible conformations

Side-chain modelling

Problem: placing the side chains of the residues that are different between model and template

Solution:

- rotamer libraries
- energy calculations



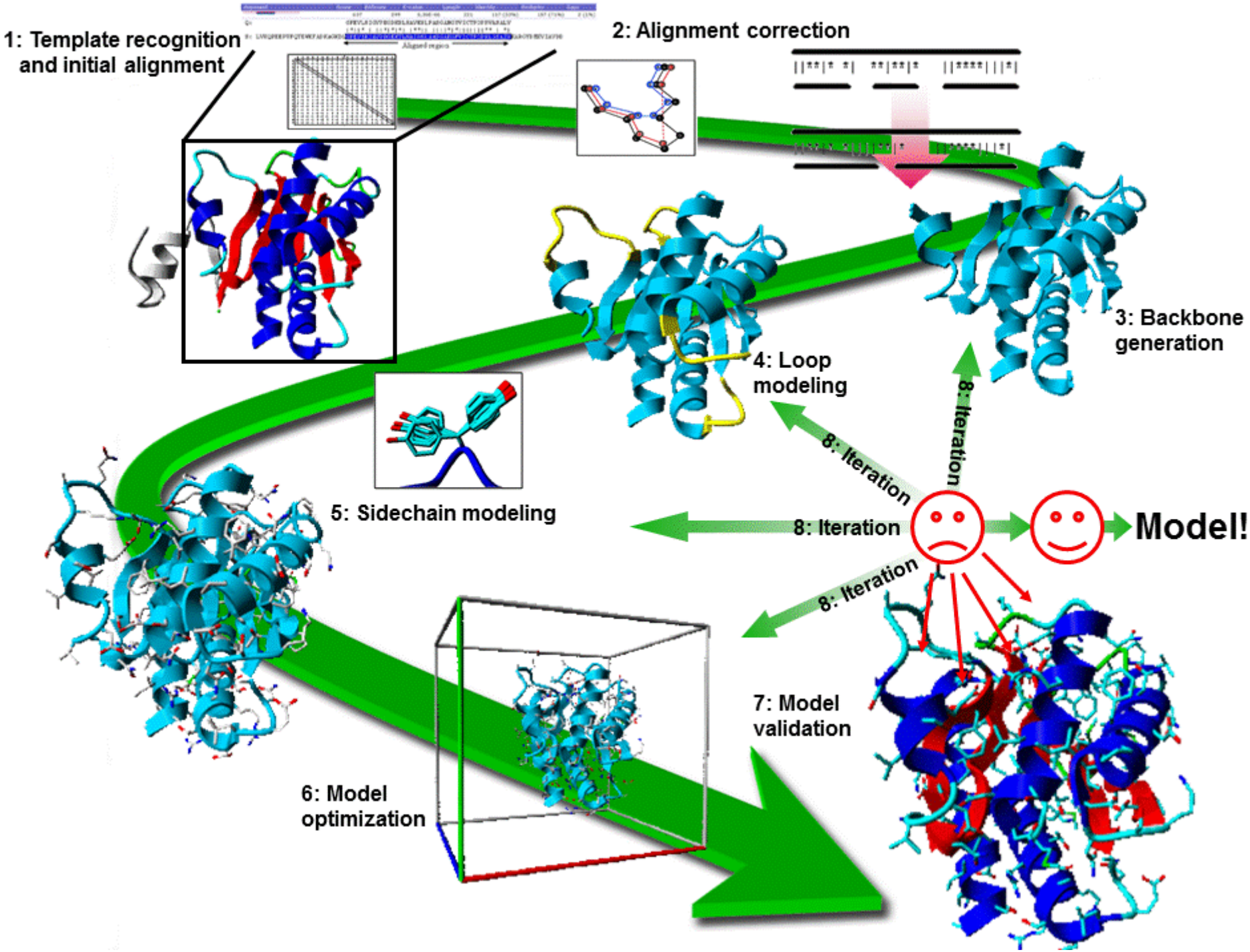
Rotamers: low energy side-chain conformations

Two commonly used rotamer libraries:
Jane & David Richardson: http://kinemage.biochem.duke.edu/databases/rotamer.php
Roland Dunbrack: http://dunbrack.fccc.edu/bbdep/index.php

Rotamer library example

```
SER     59.6    41.0
SER    -62.5    26.4
SER    179.6    32.6
```

```
TYR      63.6     90.5          21.0
TYR      68.5    -89.6          16.4
TYR     170.7     97.8          13.3
TYR    -175.0   -100.7          20.0
TYR     -60.1     96.6          10.0
TYR     -63.0   -101.6          19.3
```

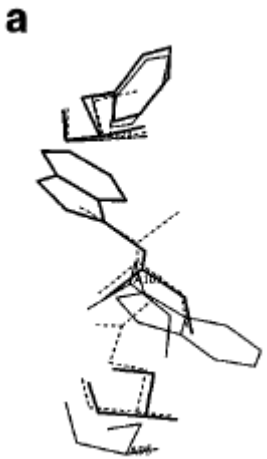

1: Template recognition and initial alignment
2: Alignment correction
3: Backbone generation
4: Loop modeling
5: Sidechain modeling
6: Model optimization
7: Model validation
8: Iteration
8: Iteration
8: Iteration
8: Iteration
Model!

Model optimisation

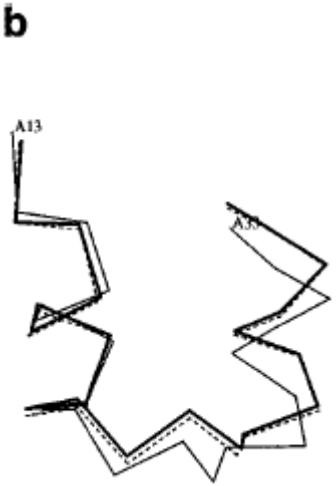
To reduce small errors accumulated during the modelling processing, MD runs can be used to minimise the model energy (e.g., CHARMM or AMBER)

- Can reduce molecular clashes
- Computing demanding
- Do not modify the model significantly
- Most MD programs make models worse rather than better

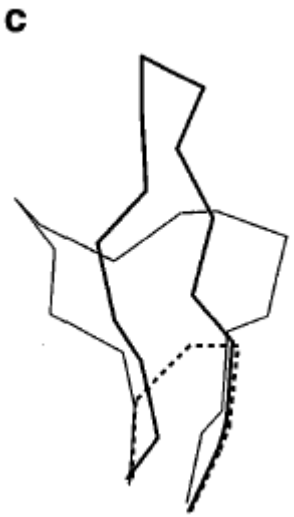
Errors in comparative modeling



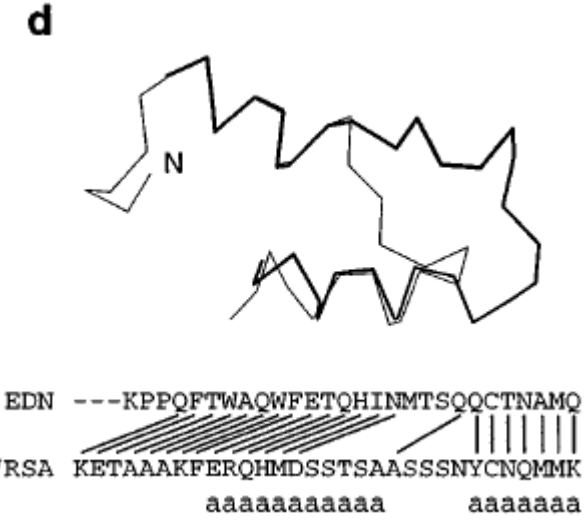
Wrong side chain conformations



Small backbone deviations



Wrong loop modeling



Wrong alignment

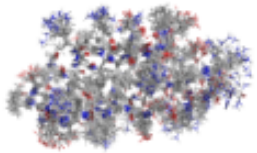
Wrong template

(Marti-Renom & Sali, 2000)

Sources of errors

- experimental errors and uncertainties in X-ray, NMR

1Å
100%

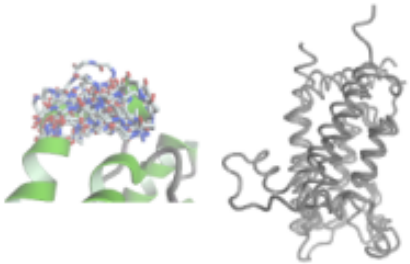


- side-chain packing
- mis-placed side-chains

1.5Å
95%

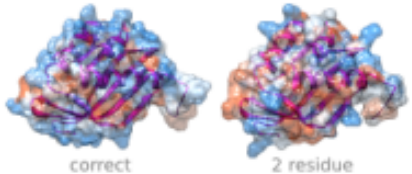
- modeling of loop regions (insertions and deletions)
- distorsions of aligned regions

60%



- alignment errors

3Å
40%



- sub-optimal template selection
- model may even have the wrong fold

>3Å
<30%

Applications

- studying catalytic mechanism / function
- structure-based drug design, ligand docking
- structural support for mutagenesis studies
- molecular replacement
- integrative modeling
- modeling into low-resolution density maps
- domain boundaries
- identification of structural motives

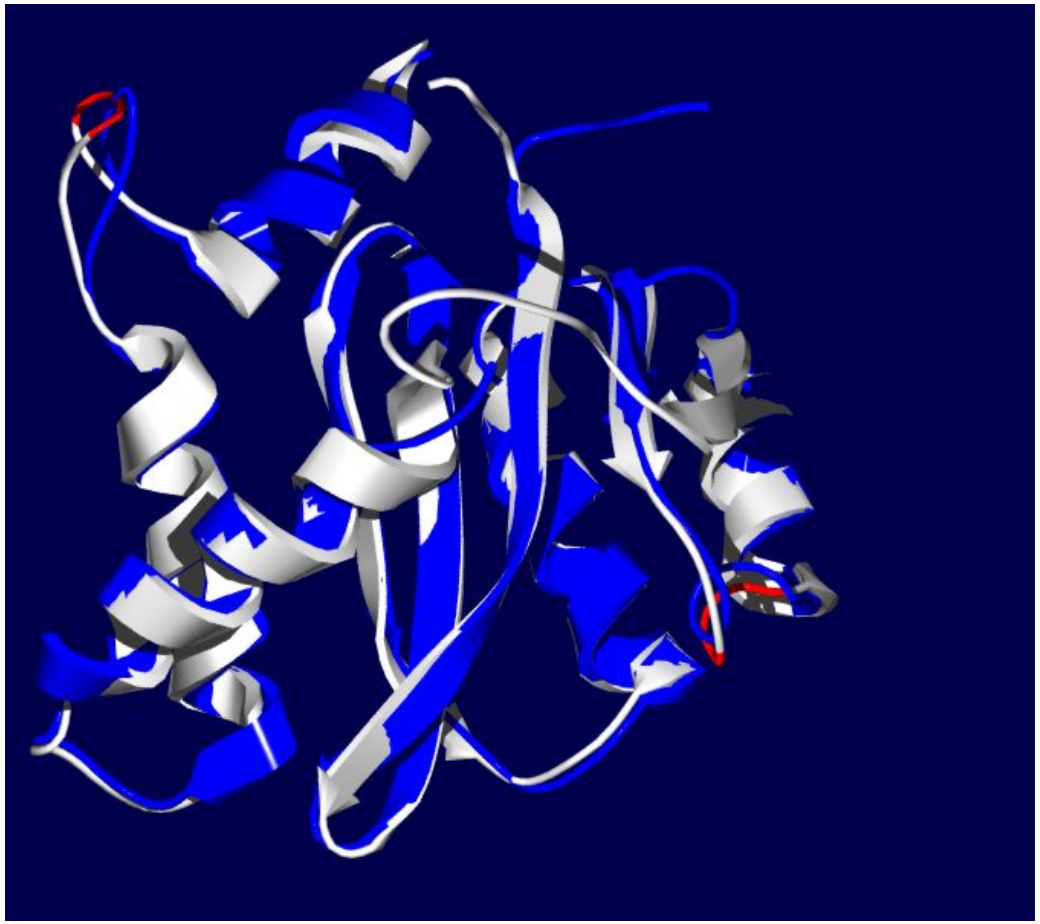
Evaluation of model accuracy

- No matter how good the software and how careful the modeller, at the end the model will contain errors

- Two main reasons:
  - % sequence identity between reference and model
  - The number of errors in templates

- Hence it is essential to check the correctness of overall fold/ structure, errors of localised regions and stereochemical parameters: bond lengths, angles, geometries…

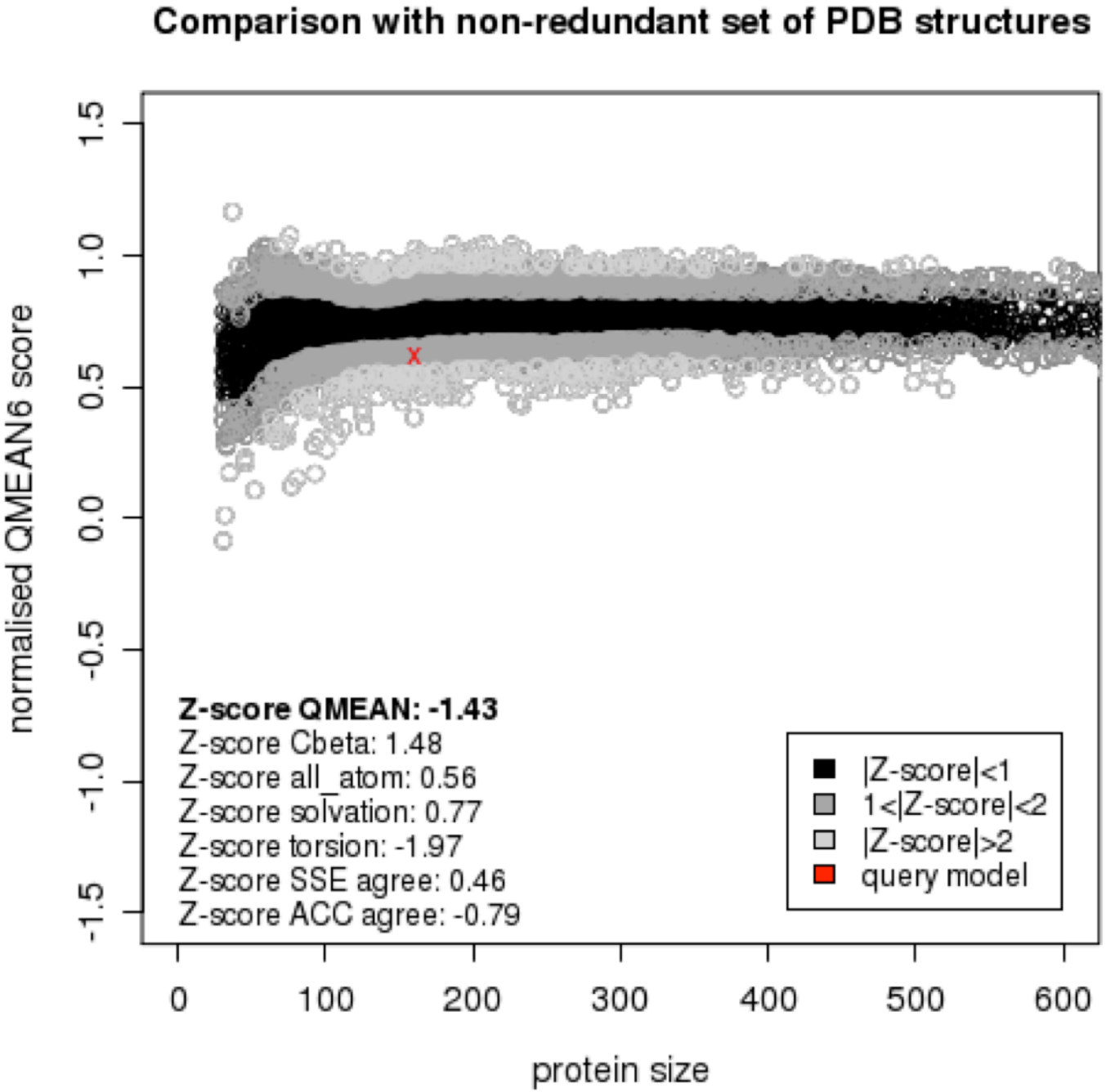
Evaluation of model accuracy



"... a model must be wrong, in some respects -- else it would be the thing itself. The trick is to see ... *where it is right.*"

Henry A. Bent
"Uses (and Abuses) of Models in Teaching Chemistry,"
J. Chem. Ed. ***1984*** *61, 774.*

SIV Model based on:	1BL3 (C) HIV-1 Integrase core domain
Experimental structure:	1C6V (C) SIV Integrase core domain
Seq. Identity:	61 %


Comparison with non-redundant set of PDB structures
normalised QMEAN6 score
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
100
200
300
400
500
600
protein size
Z-score QMEAN: -1.43
Z-score Cbeta: 1.48
Z-score all_atom: 0.56
Z-score solvation: 0.77
Z-score torsion: -1.97
Z-score SSE agree: 0.46
Z-score ACC agree: -0.79
|Z-score|<1
1<|Z-score|<2
|Z-score|>2
query model

Homology modeling - summary

- Homology modeling to high resolution is challenging
- Today models are already better than the template – GOOD NEWS!

➢ Good alignment and template selection are critical

- Sophisticated new approaches have improved homology modeling in recent years
  - Include additional information during template selection, alignment and refinement

Threading (fold recognition)

Protein threading is based on two basic observations:

1. The number of different folds in nature is fairly small (~1400)
2. 90% of the new structures submitted to the PDB in the past three years have similar structural folds to ones already in the PDB

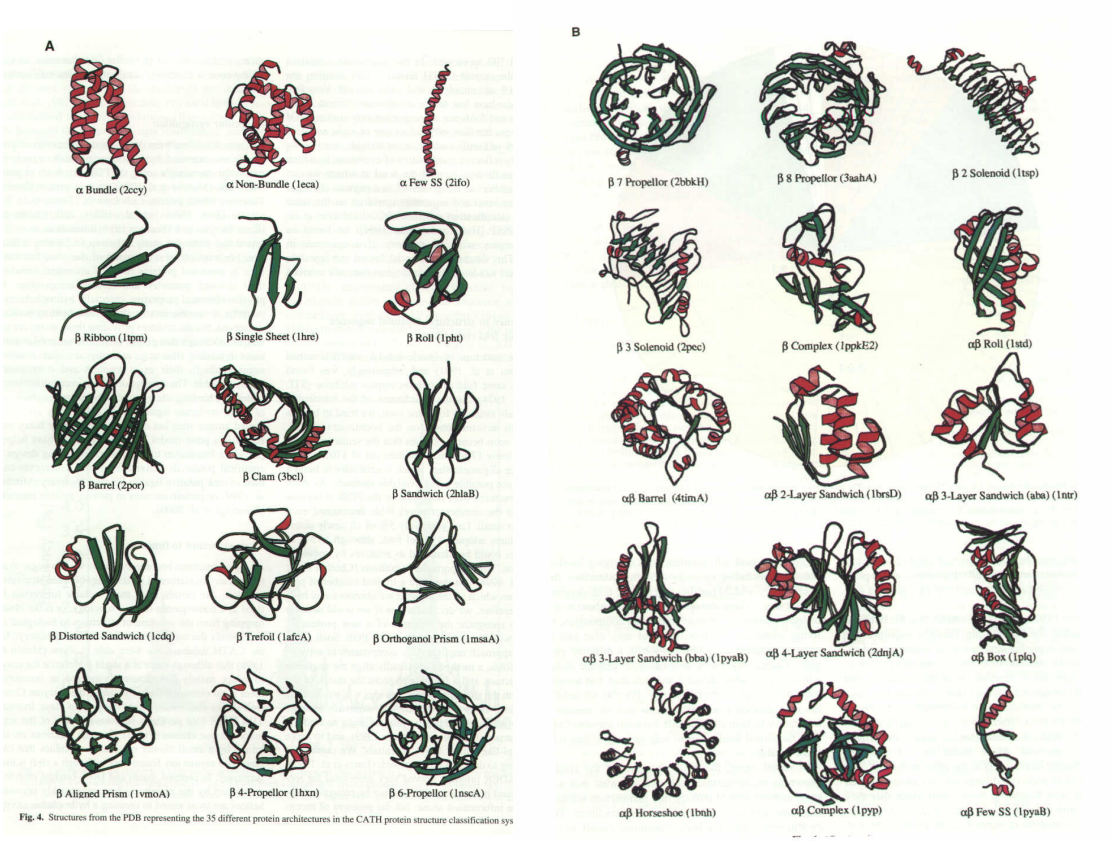
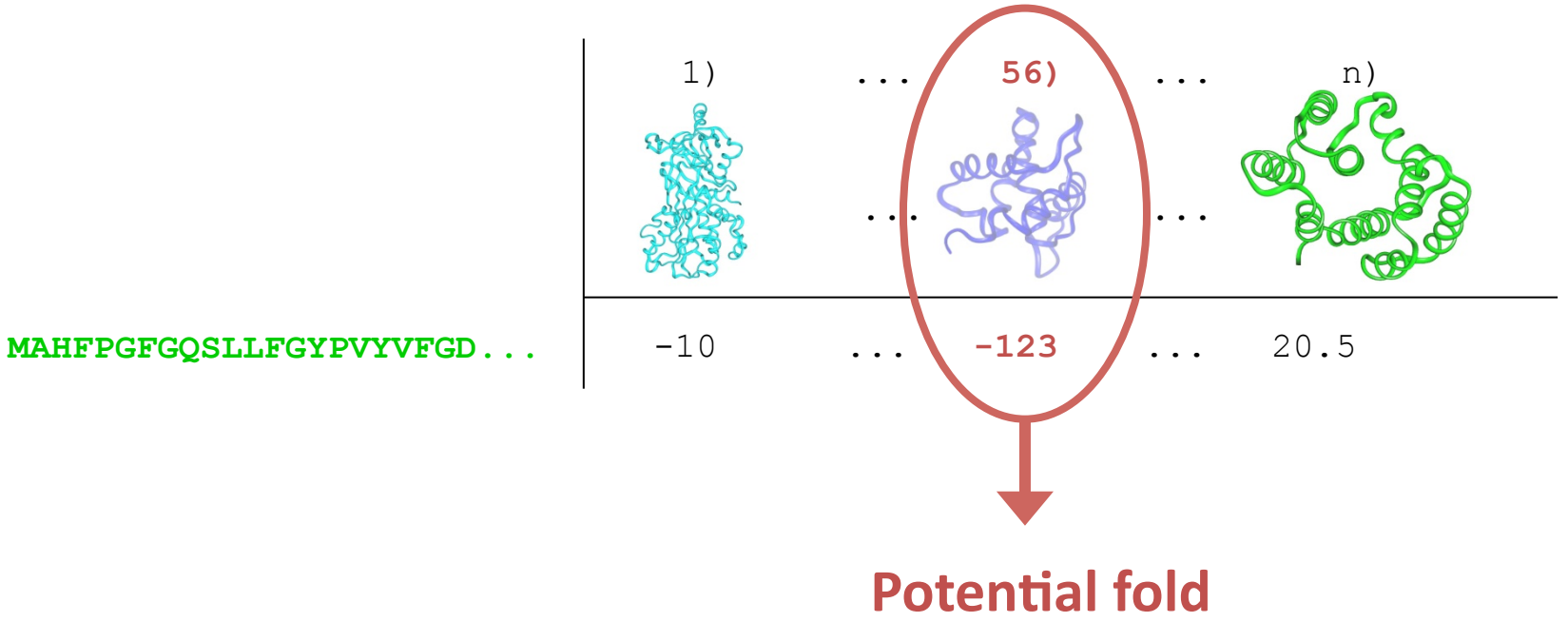


Fig. 4. Structures from the PDB representing the 35 different protein architectures in the CATH protein structure classification sys

Threading (fold recognition): Find best template for given sequence

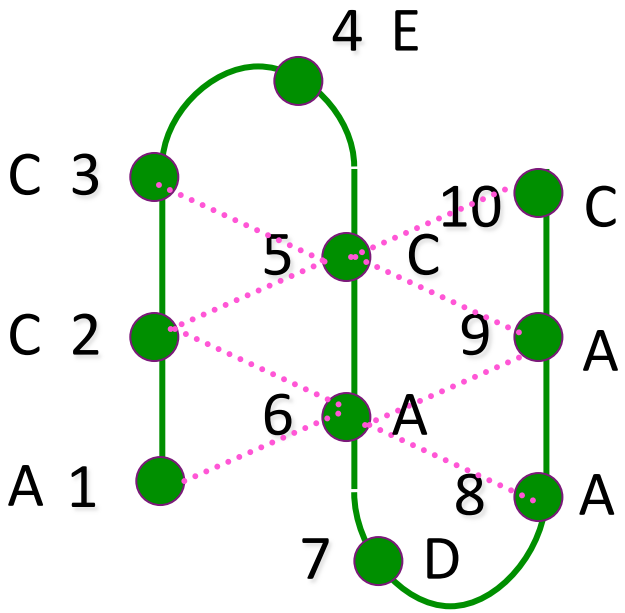
Problem description: Target has little or no sequence similarity with proteins of known structure

Idea: Evaluate compatibility of sequence to representative structure of each fold



How Threading is performed

Evaluate compatibility of sequence with fold, based on pairwise residue potentials



Essential components:

- structural template
- neighbor definition
- energy function

$$E = \sum_{\text{positions } i,j} E_{aibj}$$

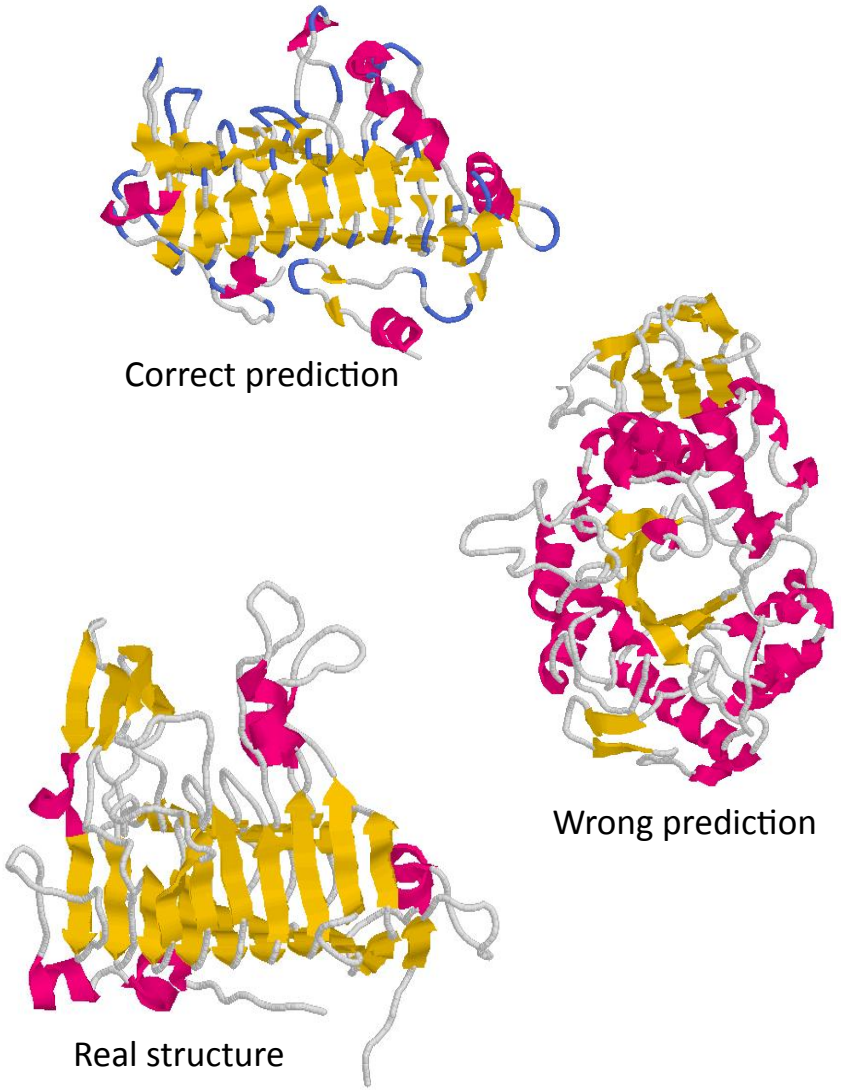
E_{ab}	A	C	D	E .....
A	-3	-1	0	0 ..
C	-1	-4	1	2 ..
D	0	1	5	6 ..
E	0	2	6	7 ..
.	.	.	.	.

ACCECADAAC

-3-1-4-4-1-4-3-3=-23

Threading (fold recognition)

- Once the best "template" has been selected, the model is built using the target-template alignment
- Prediction of the correct fold in 60-70% of cases with no clear homology



Web sites for fold recognition

Profiles:

3D-PSSM - http://www.bmm.icnet.uk/~3dpssm

Libra I - http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/LIBRA_I.html

UCLA DOE - http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

Contact potentials

123D - http://www-Immb.ncifcrf.gov/~nicka/123D.html

Profit - http://lore.came.sbg.ac.at/home.html

RaptorX

State of the art threading method of choice

- Successful for "low-homology" proteins (few homolog sequences)
  - Adjusts reliance on sequence profile Vs. Structure information based on quality of profile and assessed similarity to template .
  - Optimizes use of **several** templates - fix sequence – template alignment errors by looking for a consensus alignment

http://raptorx.uchicago.edu/

Jian Peng and Jinbo Xu. RaptorX: exploiting structure information for protein alignment by statistical inference. PROTEINS, 2011; A multiple-template approach to protein threading. PROTEINS, 2011.

Ab initio methods

- If structure homologues do not exist, or exist but cannot be identified, models have to be constructed from scratch
- Currently, the accuracy of *ab initio* modelling is low and the success is limited to small proteins (<100 residues)
- Typically, *ab initio* modelling **conducts a conformational search under the guidance of a designed energy function**
- This procedure usually generates a number of possible conformations (structure decoys), and final models are selected from them

- ***ab initio*** modelling (Klepeis et al. 2005; Liwo et al. 2005; Wu et al. 2007),
- ***de novo*** modelling (Bradley et al. 2005),
- **physics-based** modelling (Oldziej et al. 2005)
- **free modelling** (Jauch et al. 2007)

Ab initio methods

A successful *ab initio* modelling depends on three factors:

(1) an accurate energy function with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to all possible decoy structures;

(2) an efficient search method which can quickly identify the low-energy states through conformational search;

(3) selection of native-like models from a pool of decoy structures.

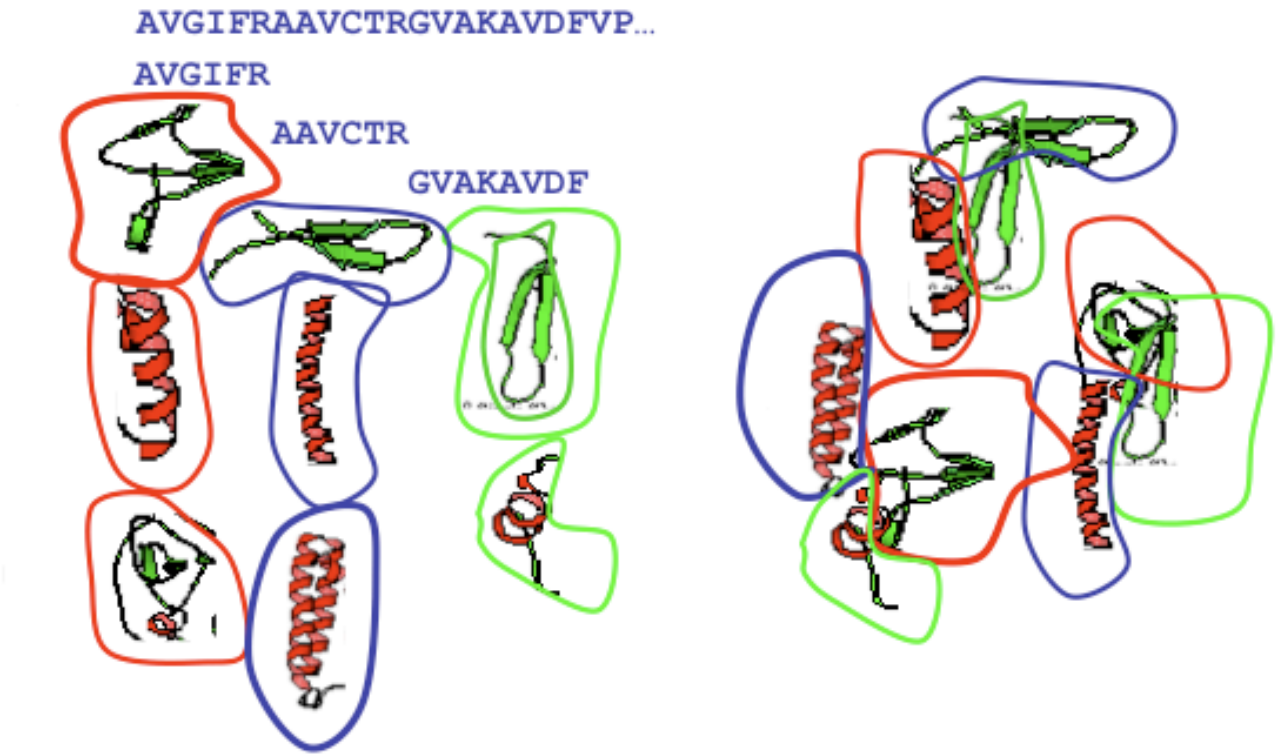
One of the best-known ideas for *ab initio* modelling

- Generation of protein models by assembling small fragments (mainly 9-mers) taken from the PDB library

- Baker and coworkers developed ROSETTA, which was extremely successful for the free modelling targets in CASP experiments and made the fragment assembly approach popular in the field

(Bowie and Eisenberg 1994)
(Simons et al. 1997)

Rosetta



Bystroff and Baker, JMB, 1998

Rosetta

AVGIFRAAVCTRGVAKAVDFVP…

AVGIFR

AAVCTR

GVAKAVDF

Optimize and score

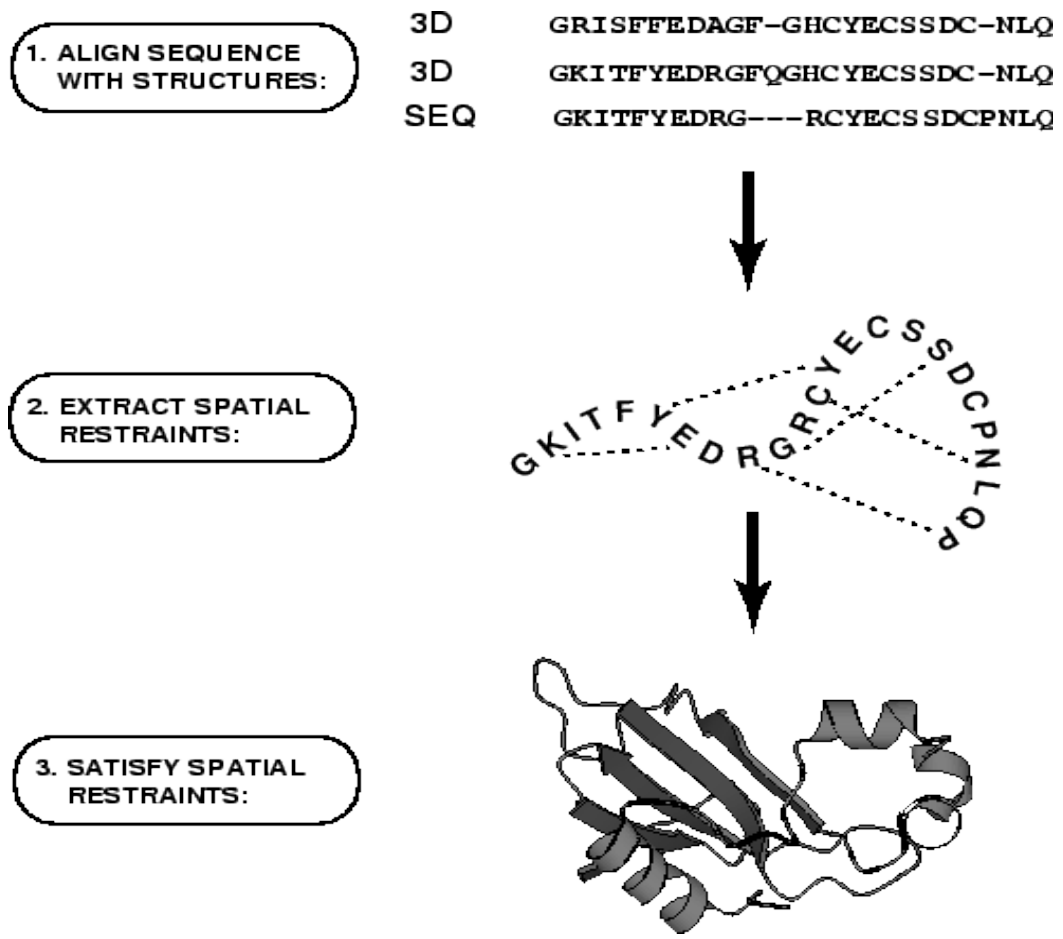
Approaches to modelling

- MODELLER
  - Developed by Sali
  - Based on optimisation that satisfies a list of constraints
- ROSETTA
  - Developed by Rosetta Commons
  - Based on full-atom refinement of the protein model using the Rosetta full-atom energy function. The final step is a selection of the models using clustering.
- I-Tasser:
  - Developed by Zhang and Skolnick
  - Based on threading of parts of sequence onto parts of known structures

Modeller (template-based)

Satisfaction of spatial constraints

Overview



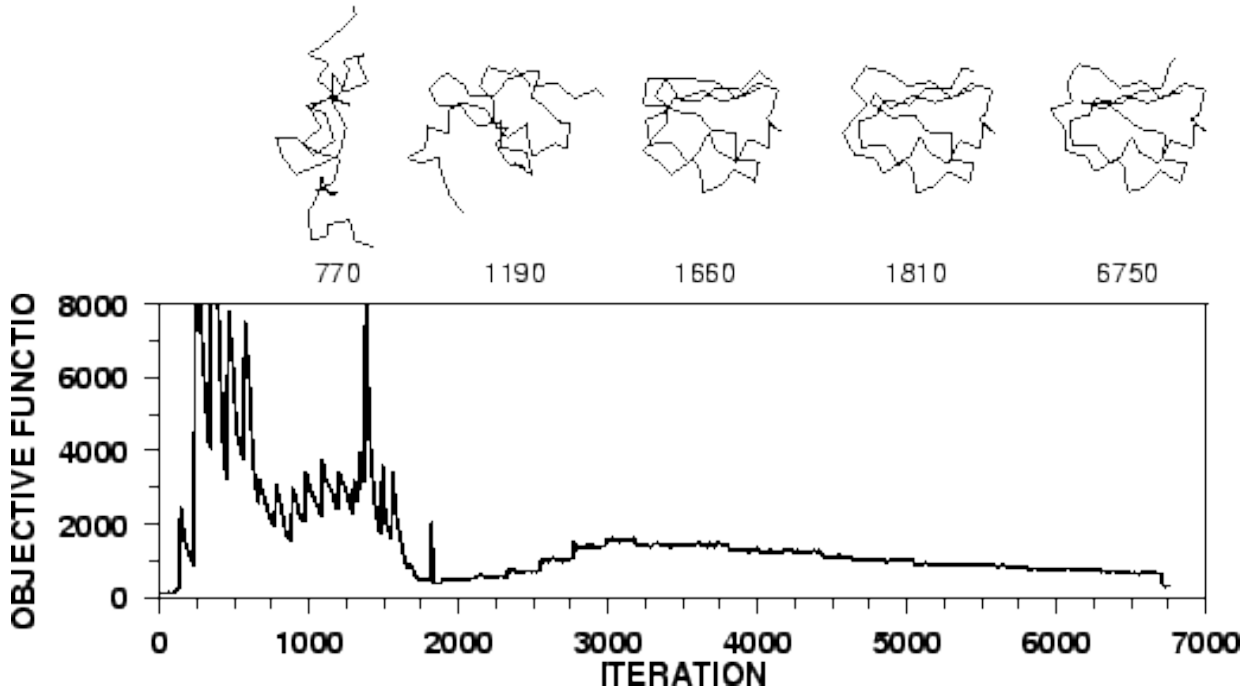
(*Sali*)

Modeller

- Spatial restraints:
  - Ca-Ca distances
  - Dihedral angles
  - Hydrogen bonds
  - *Etc*
- Objective function:
  - spatial restraints and energy terms enforcing proper stereochemistry are combined into an objective function

Modeller:
optimise objective function

- Optimization:
  - Iterative Conjugate gradient minimisation
  - Molecular dynamics with simulated annealing



Homology modeling with Rosetta

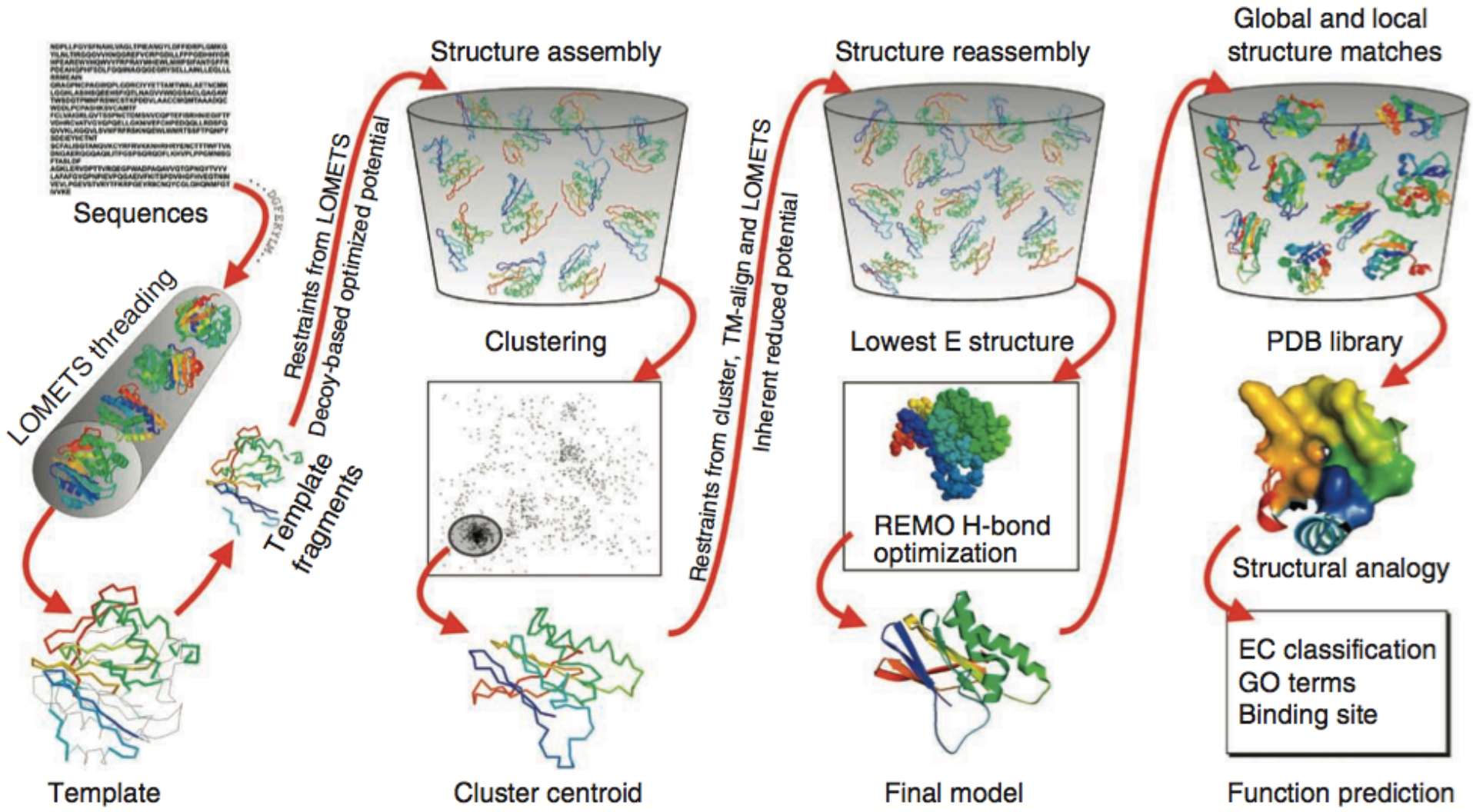
Summary - Basic protocol:

1. Detect template and align sequence: based on HHSEARCH (alignment of two HMMs) or RAPTOR (Threading)
2. Define aligned regions and loop regions; copy aligned regions and complete protein structure with loop modeling (using KIC/CCD)
3. Refine structure with the “relax” protocol

I-Tasser *I*terative *T*hreading ***Asse***mbly ***R***efinement

*(**Zhang**, & Skolnick)*

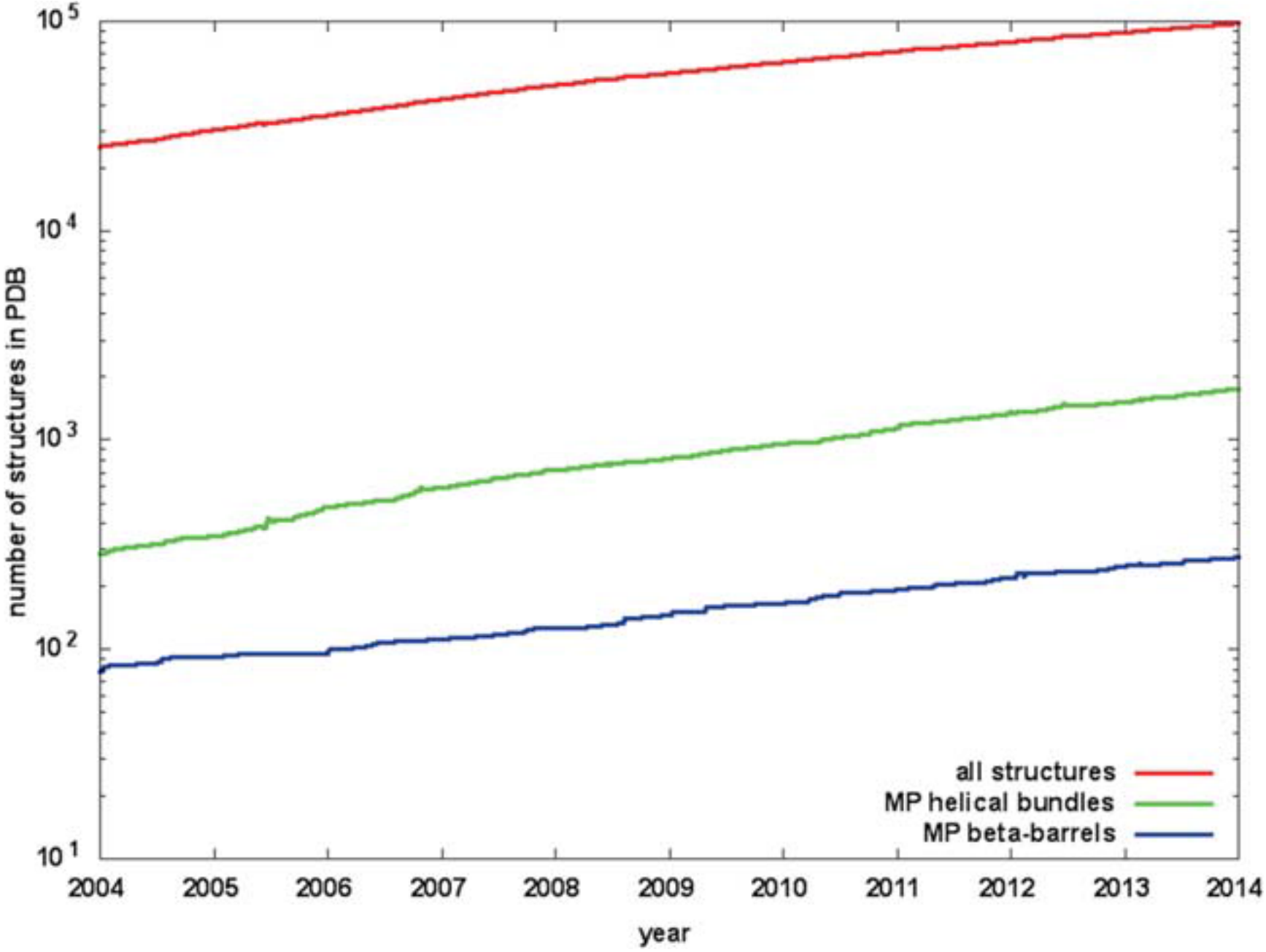
Separate training of protocol for: easy/ medium/ hard targets



Membrane proteins

- up to 30% of the human genome encodes membrane proteins
- Membrane proteins are estimated to be the targets of 50% of drugs that are currently in development
- difficulties in overexpression, reconstitution into membrane mimetics, and subsequent structure determination
- few membrane protein crystal
- Only 1–2% of the PDB (for the last decade)

Leman et al. Proteins 2014

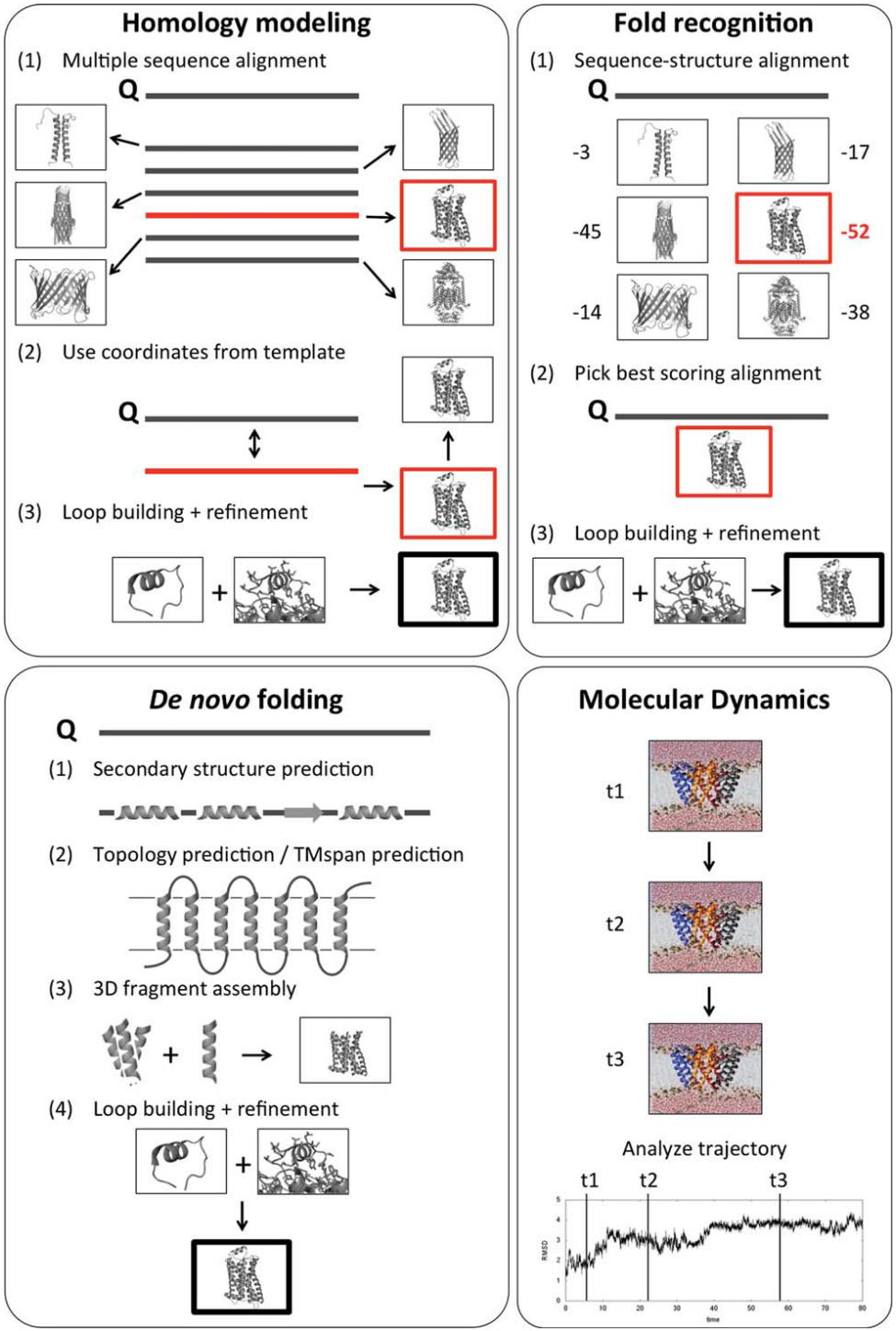


logarithmic scale

Modelling of membrane proteins

- advantages over the prediction of soluble proteins
  - smaller conformational search space
  - only two structural motifs in the bilayer
  - membrane-specific features
- challenges:
  - large size of the proteins
  - derivation of accurate scoring functions and molecular force field parameters to model the membrane

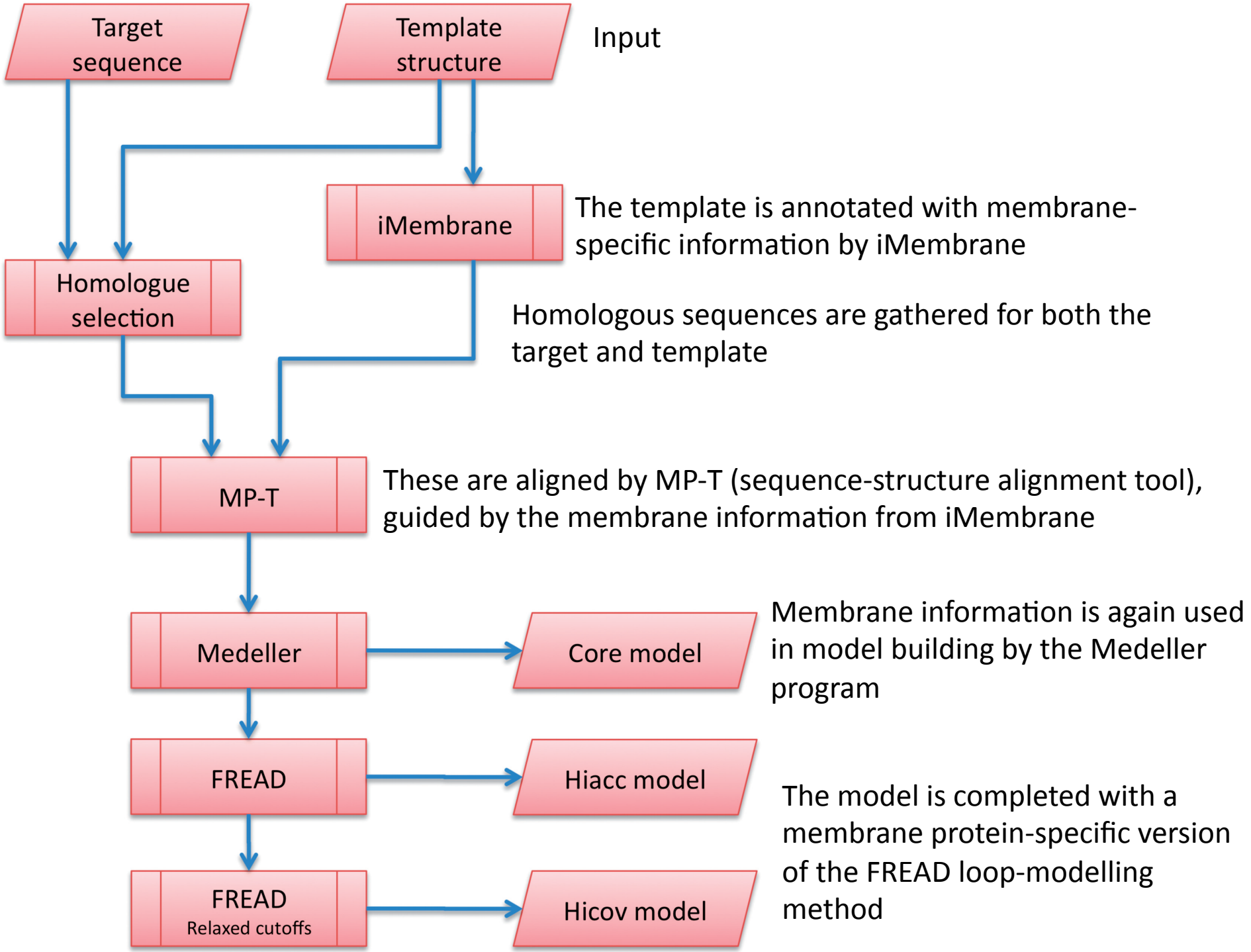
- MP specific substitution matrices
- TM span prediction to improve sequence alignments
- Incorporate environment parameters such as secondary structure, accessibility, and membrane depth



Leman et al. Proteins 2014

Template-based methods

- RosettaMembrane
  - rosettacommons.org
- MODELLER (no membrane mode)
- MEDELLER (outperforms MODELLER)
  - http://medeller.info
- MEMOIR fully automated web server (outperforms HHSearch and SwissModel)
  - http://opig.stats.ox.ac.uk/webapps/memoir


Target sequence
Template structure
Input
iMembrane
The template is annotated with membrane-specific information by iMembrane
Homologue selection
Homologous sequences are gathered for both the target and template
MP-T
These are aligned by MP-T (sequence-structure alignment tool), guided by the membrane information from iMembrane
Medeller
Core model
Membrane information is again used in model building by the Medeller program
FREAD
Hiacc model
FREAD
Relaxed cutoffs
Hicov model
The model is completed with a membrane protein-specific version of the FREAD loop-modelling method

Membrane proteins

- Homology modelling: Memoir (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692111/)
- Rosetta: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479309/
- Ab initio: EVfold_membrane (http://www.cell.com/cell/fulltext/S0092-8674(12)00509-0)

CASP

- Double-blind structure prediction experiment
  allows assessment of different approaches
- *every 2 years; summer 2014: CASP11*
- Steady improvement of methodology

Categories:

- Template based modeling **(TBM)**
- Free modeling **(FM)**
- Refinement of initial models

Identification of "winner strategies":

- Rosetta in CASP4-6
- iTasser in CASP7 & CASP8
- servers
- improved combination of multiple templates in CASP9
- CASP10: refinement with MD
- CASP11: contact prediction methods & contact-assisted modeling

- **New:** prediction of contacts, unstructured regions, ligand binding sites

http://www.predictioncenter.org/casp10/meeting/talks.html
Proteins special issue vol:82, S2

Improvement over the years

Improvement in each round

- CASP7: in *difficult* region
- CASP8: accuracy in template-based modeling (few difficult cases)
- CASP9: intermediate difficulty targets
- CASP10: refinement using MD CASP11: contact predictions

Summary (CASP)

- *steady improvement of structure prediction over the years*
- impressing quality of current *ab initio* modeling
  - efficient combination of appropriate sampling strategies and a tailored energy function
- models now often better than template
- automatic servers outperform now also FM

To which extent can we use homology models in protein-protein docking?

- The most important thing for a successful docking is the correct interface
- Side chain rotamers might differ significantly from the original crystal
- Even small (2Å) local loop deformations on the backbone level can affect the contact formation upon docking

The quality of the model affects the accuracy of docking results

- The use of good quality models is crucial
- Quantitative and robust correlations exist between the **accuracy** of docking results and the model quality (especially in the binding site
- In the most desirable scenario, docking accuracy would be predicted directly from the quality of the protein model

Bordogna et al. J Comput Chem, 2011

Using homology models in protein-protein docking is feasible…

- The sequence identity can be low as long as we can be sure that **the binding site is well conserved**
- Binding site very similar to the original structure
- Accurate modeling of the binding site
- Docking approaches allow flexibility in the binding site
- Use of restraints

Conclusions

- Homology modelling can provide "low-resolution" structures – the alignment step is crucial
- Homology modelling can be used in membrane protein prediction
- CASP is contributing to a steady improvement of structure prediction over the years
- Models can be used in docking experiments provided the binding site is well conserved