Comparative Modeling

Part 1

Jaroslaw Pillardy
Computational Biology Service Unit
Cornell Theory Center

- Function is the most important feature of a protein
- Function is related to structure
- Structure is related to sequence

Exchange of even a few residues can (and often does) change a structure and destabilizes a protein.

Multiple random mutations that preserve structure are statistically unlikely.

Mutations resulting in a structural change are not likely to survive since function is lost …

There is a very significant evolutionary pressure for structure conservation.

RESULT:
STRUCTURE IS MORE CONSERVED THAN SEQUENCE

All naturally evolved proteins of sequence identity of 35% or more have similar structures.

Majority of proteins with sequence similarity of 15% or more still have similar structures, however degree of similarity decreases with sequence similarity.

Using sequence-to-sequence alignment a model can be constructed when structure of one of aligned proteins is known …

Comparative Modeling is an algorithm for constructing a 3-dimensional model of a protein based on a known structure of another protein (template) when optimal alignment of the two sequences is known.

- When sequences are similar BLAST alignment (usually modified) may be used, often leading to a model with quality corresponding to low-resolution experimental (X-ray or NMR) structure. This is the only method for protein structure prediction capable of producing models that close to experiment.
- For distantly related sequences alignment obtained with different methods may be used.
- Quality of an alignment may be assessed by analyzing a model.

Why is 3D model important?

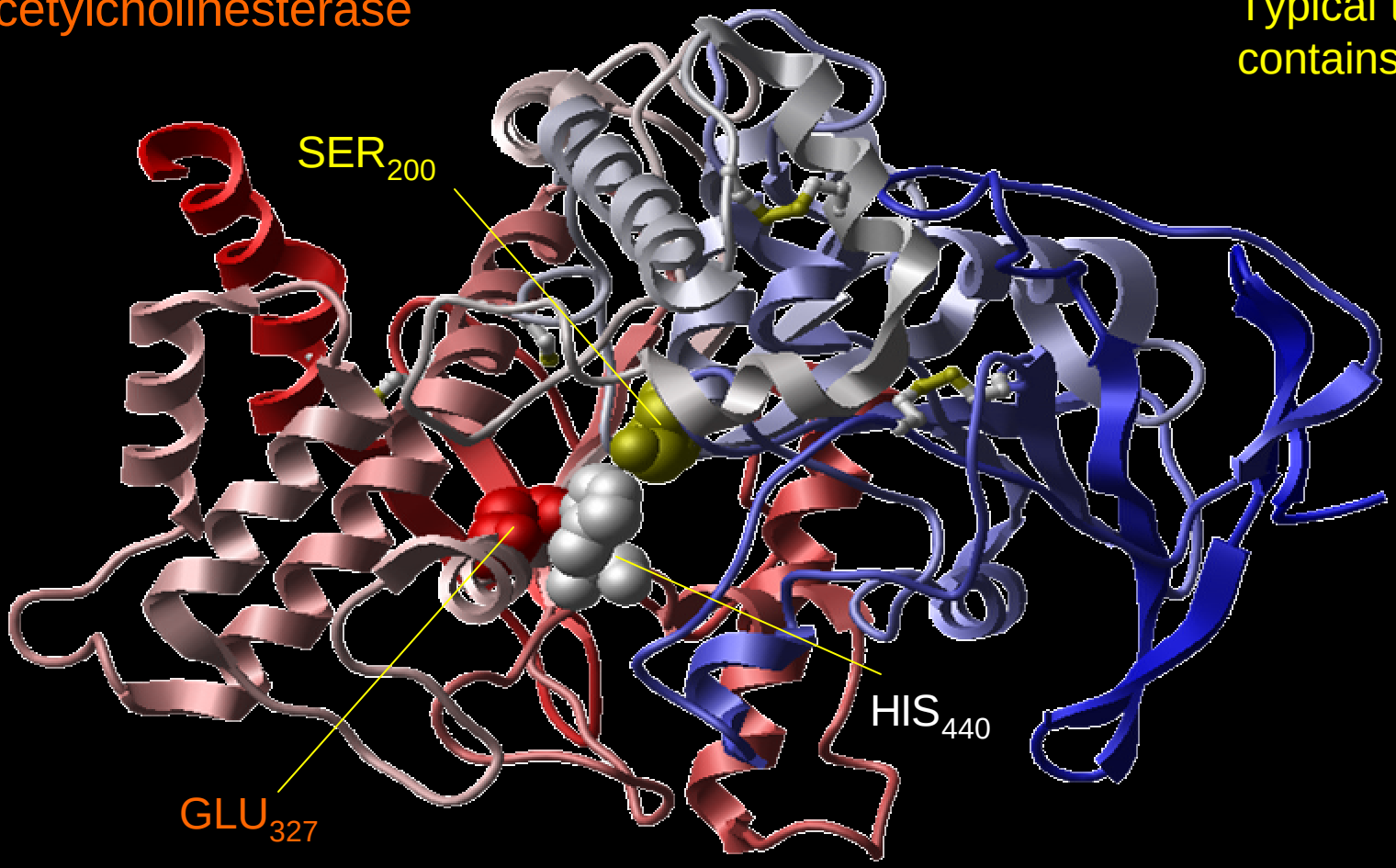
Typical problem: prediction of protein function.

It may be done in some cases by sequence matching to a protein of a known function and/or recognizing block of sequence of known function.

The 3D structure of a protein generally provides much more information about its function than sequence because interactions of a protein with other molecules are determined by amino acids that are close in space but frequently distant in sequence (e.g. serine proteases) …

1EA5 Acetylcholinesterase

Typical triad contains ASP



DDHSELLVNTKSGKVMGTRVPVLSSHISAFLGIPFAEPPVGNMRFRRPEPKKPWSGVWNASTYPNNCQQYVD
EQFPGFSGSEMWNPNREMSEDCLYLNIWSSEEEBSVPSPRPKSTTVMVWIYGGGFYSGSSTLDVYNGKYLAY
TEEVVLVSLSYRVGAFGFLALHGSQEAPGNVGLLDQRMALQWVHDNIQFFGGDPKTVTIFGE**S**AGGASVGMHI
LSPGSRDLFRRAILQSGSPNCPWASVSVAEGRRRAVELGRNLNCNLNSDEELIHCLREKKPQELIDVEWNVLPF
DSIFRFSFVPVIDGEFFPTSLESMLNSGNFKKTQILLGVNKD**E**GSFFLLYGAPGFSKDSESKISREDFMSGVKLSV
PHANDLGLDAVTLQYTDWMDDNNGIKNRDGLDDIVGDHNVICPLMHFVNKYTKFGNGTYLYFFNHRASNLVWP
EWMGVI**H**GYEIEFVFGLPLVKELNYTAEEEALSRRIMHYWATFAKTGNPNEPHSQESKWPLFTTKEQGGTGGG
KFIDLNTEPMKVHQRLRVQMCVFWNQFLPKLLNATAC

Why not to use 3D structure of template?

Model is similar, but not identical to the template.

- Side chains are replaced and backbone conformation modified if necessary. They must be packed optimally in the protein core.
- Gaps on the target side of alignment are filled by constructing and optimizing loops.
- Gaps on the query side of alignment are removed by moving segments of protein in space and relaxing parts of its structure.

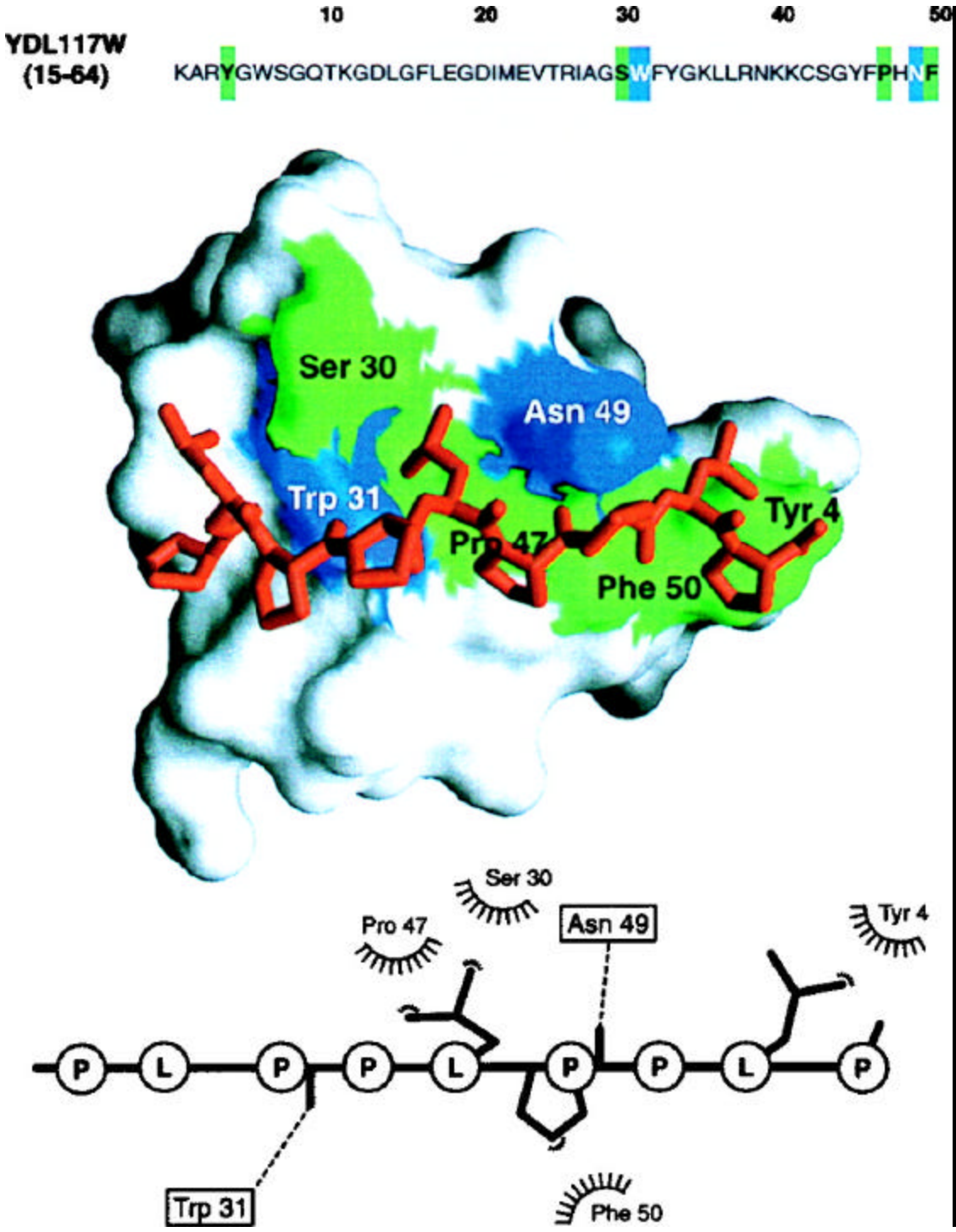
In general model should be a little closer to the target structure than the template.

This is especially true when more than one template is being used – this is because model tends to inherit the best structural features from different templates.

It is always better to use comparative model to represent target rather than its template. This is because errors in alignment (typical source of bad models) affect similarly the use of the template as a representation of the target as well as the comparative model based on the same template.

Various applications for comparative models:

- Designing (site-directed) mutants to test hypotheses about function
- Identifying active and binding sites
- Searching for ligands of a given binding site
- Modeling substrate specificity
- Predicting antigenic epitopes
- Protein-protein docking simulations
- Inferring function from calculated electrostatic potential
- Molecular replacement in X-ray structure refinement
- Testing a given sequence-structure refinement
- Rationalizing known experimental observations
- Planning new experiments



Modeling a putative interaction of a predicted YDL117W SH3 domain with a proline-rich peptide.

A segment in the yeast ORF YDL117W sequence (*Top*) was predicted to be remotely related to the SH3 domains, many of which have known 3D structure.

[R. Sànchez and A. Šali, PNAS Vol. 95, Issue 23, 13597-13602 (1998)]

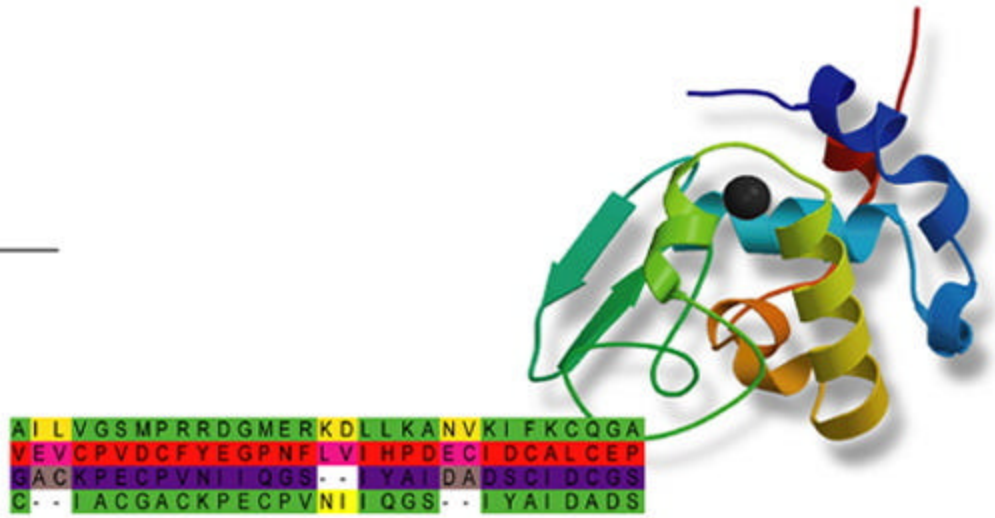
Comparative Modeling is broadly applicable.

It has been estimated that approximately one third of all sequences are related to at least one protein of known structure. Since there are approx 1,000,000 known protein sequences comparative modeling could be applied to approximately 300,000 proteins. It is 17 times more than the number of known protein structures in the PDB (18,000).

The number of possible distinct protein folds may be limited. In this case at some point in the future at least one example of most structural folds will be known making comparative modeling applicable to most protein sequences.

Modeller

Program for Comparative Protein
Structure Modelling by Satisfaction
of Spatial Restraints

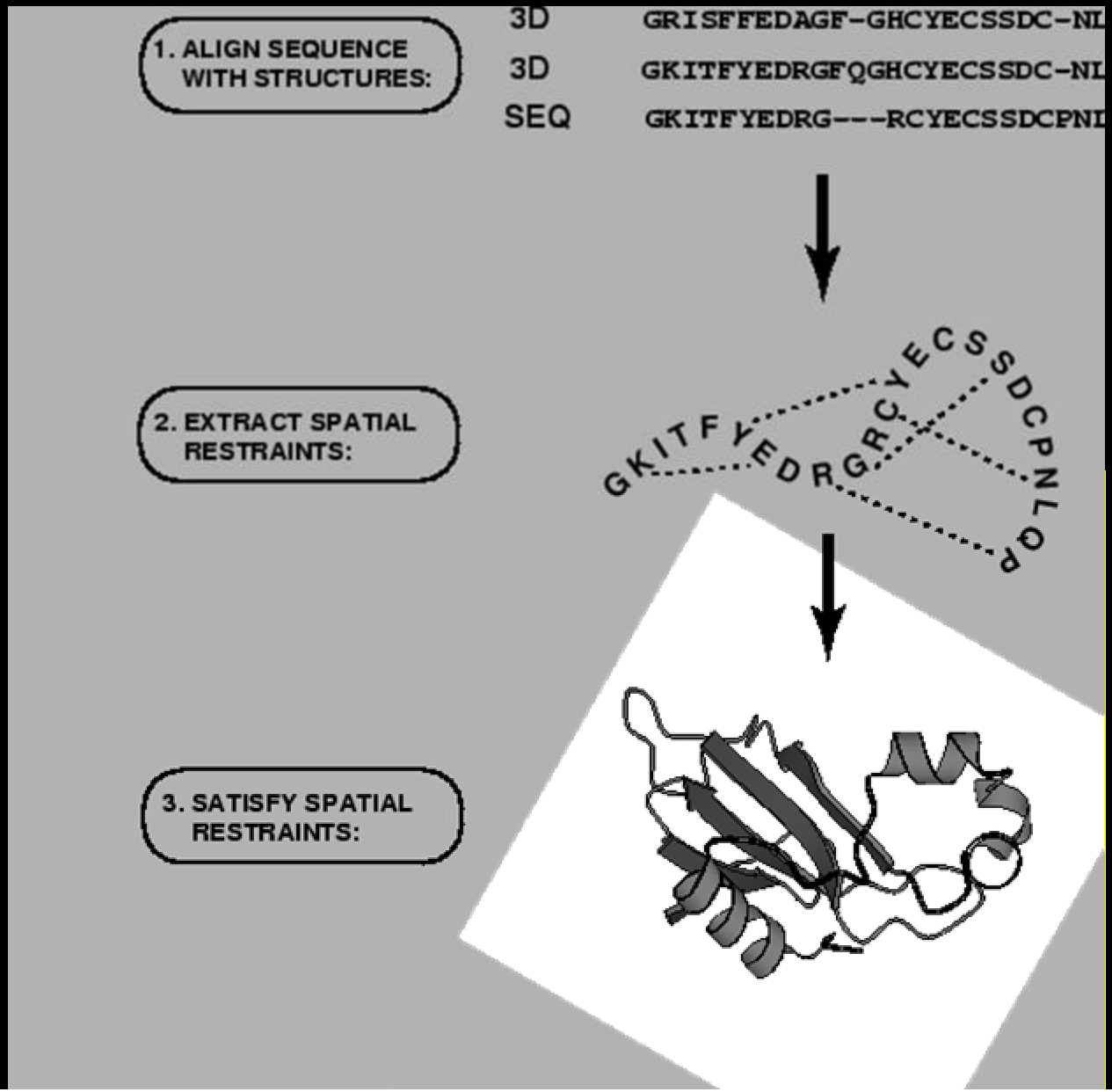


MODELLER is a very popular comparative modeling program.

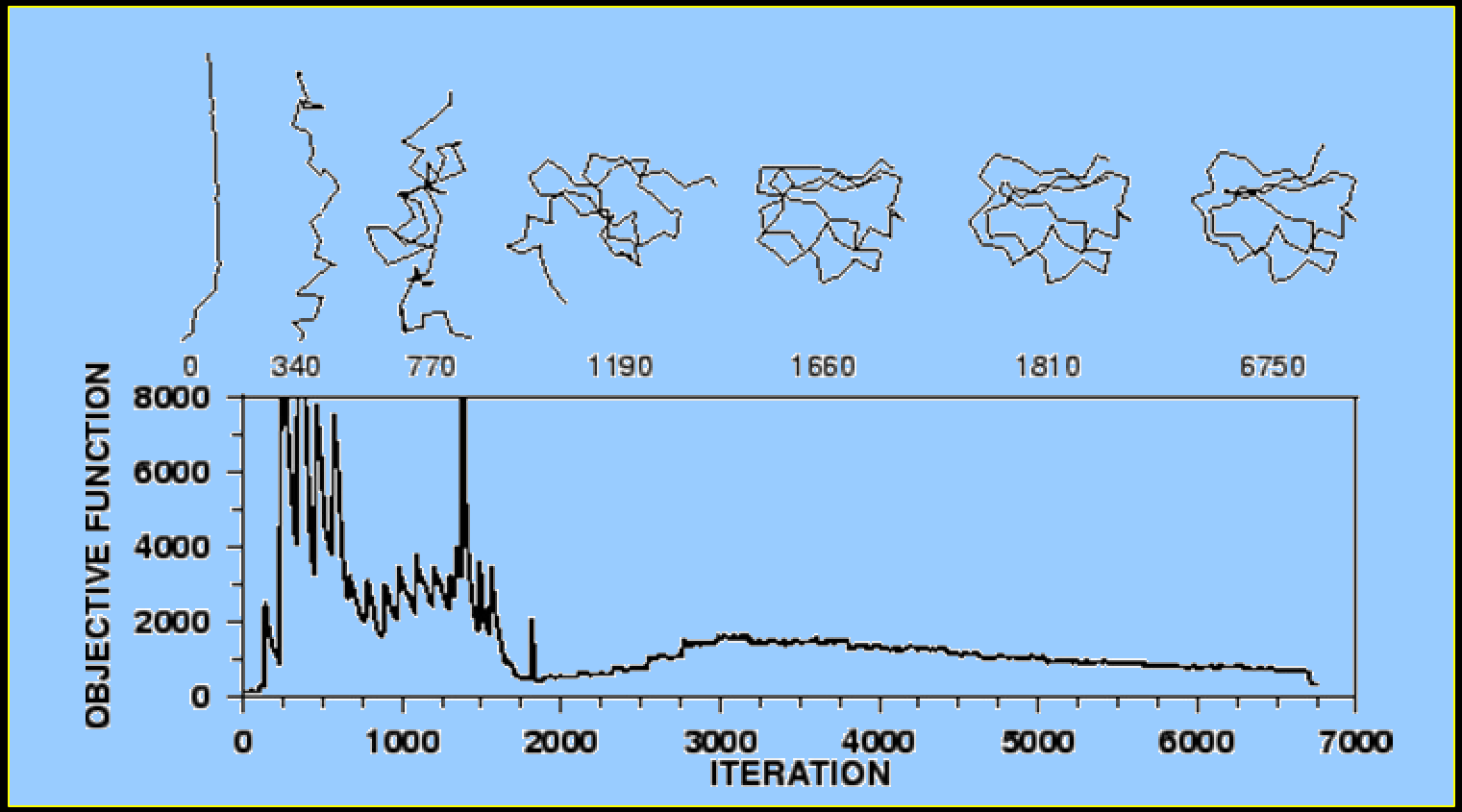
It is a part of many commercial packages
(e.g. InsightII from Accelrys).

It is also available for free from the authors at
http://guitar.rockefeller.edu/modeller/modeller.html

How does the MODELLER work?



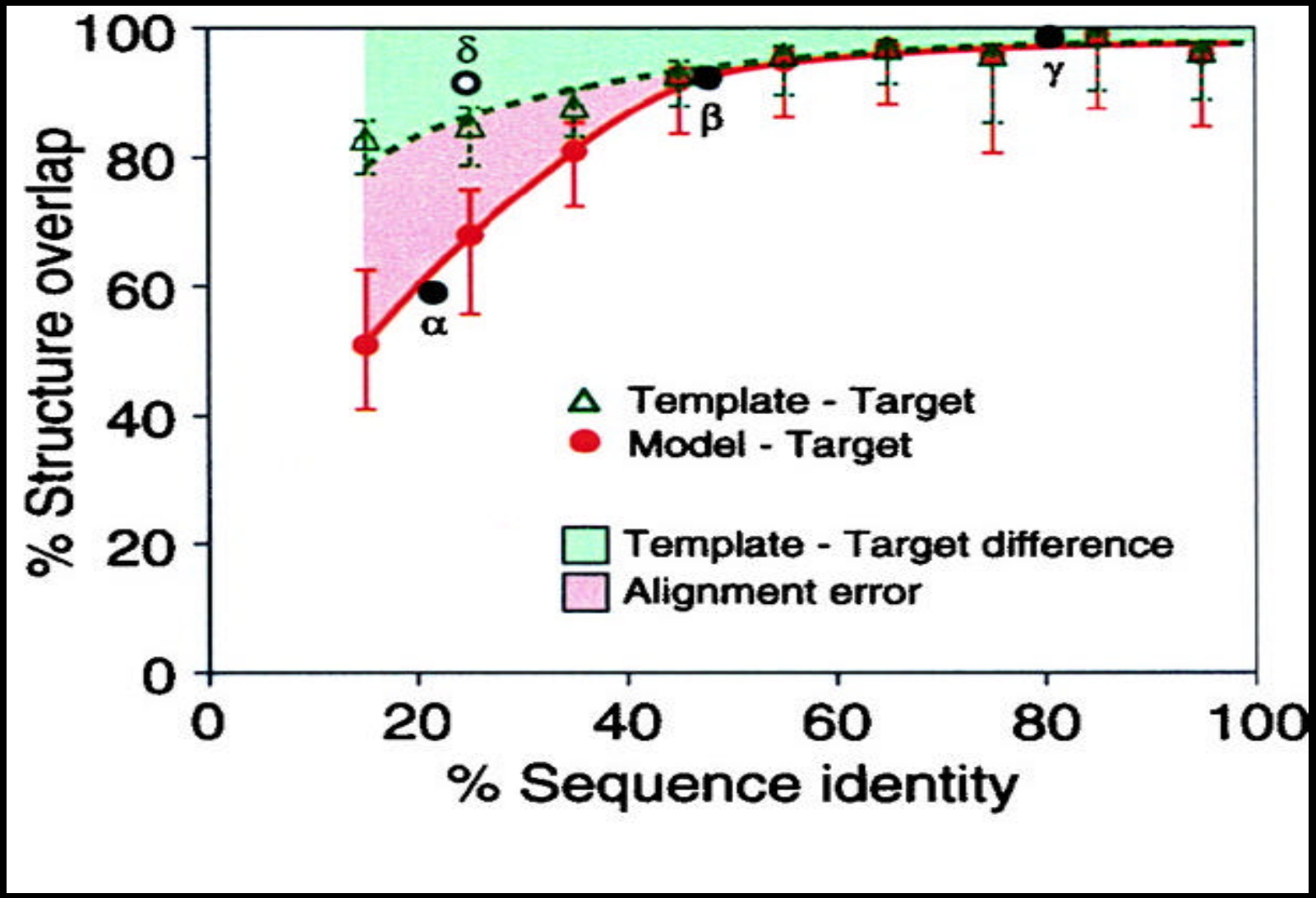
How does the MODELLER work?



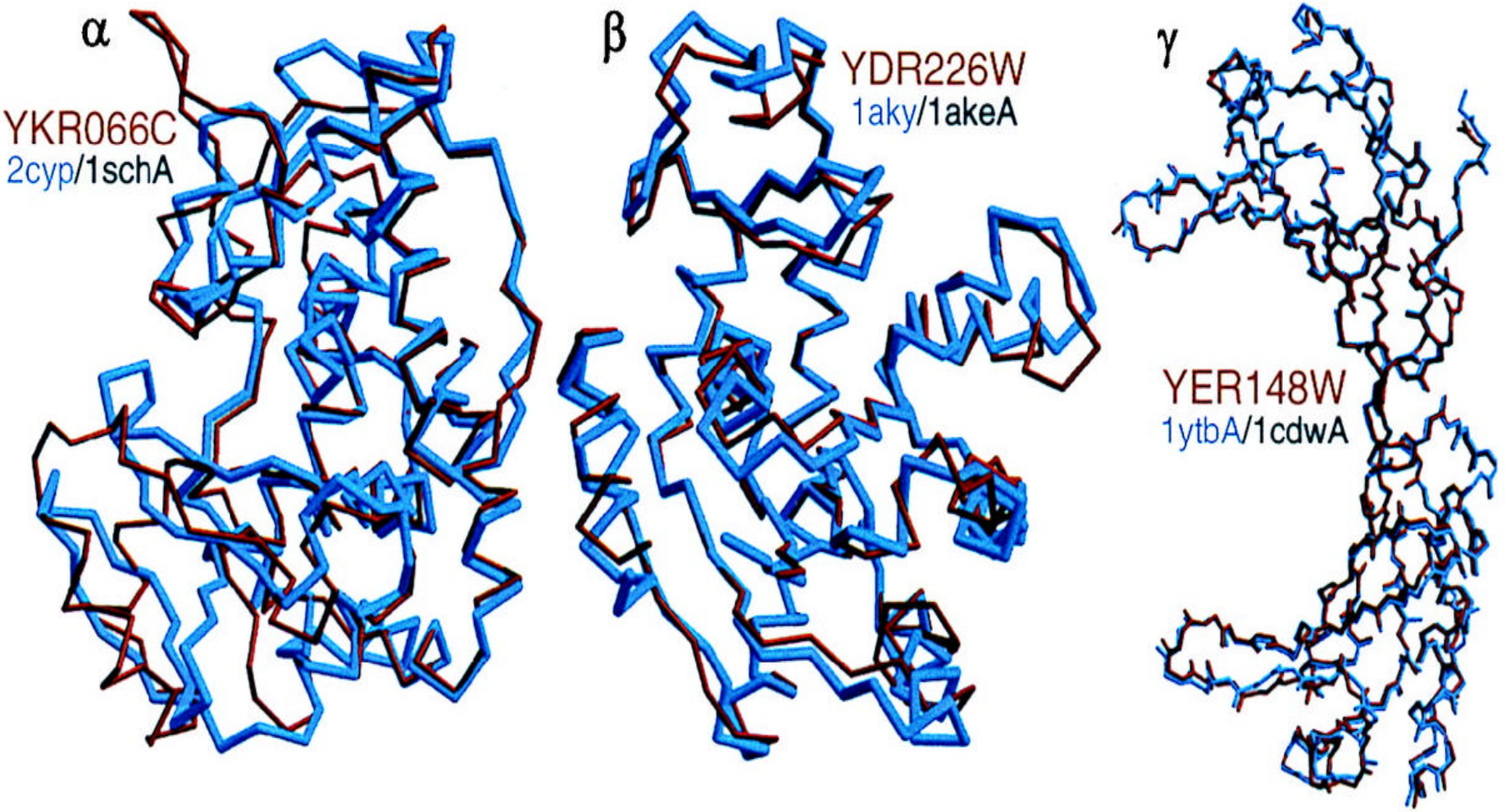
Optimization of the objective function (curve) starts with a distorted average of template structures (not with an extended structure as shown here). The iteration number is indicated below each sample structure. In this run, the first iterations correspond to the variable target function method relying on the conjugate gradients technique. This approach first satisfies sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimized.

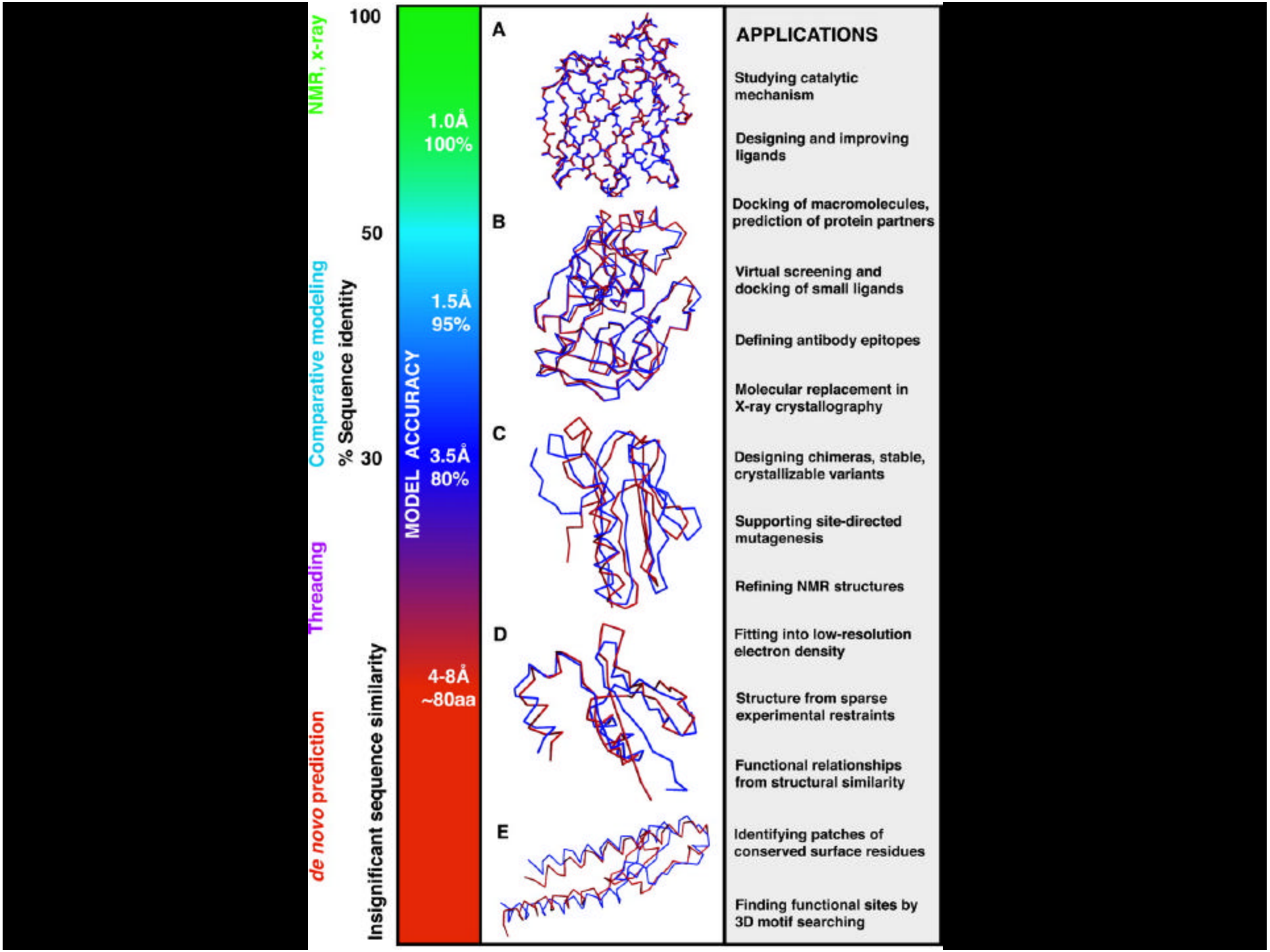
MODELLER's approach (spatial constraints) is very similar to refining of a NMR structure. Other approaches are:

- Rigid body assembly. This group of methods constructs the model from a few core regions and from loops and side chains, which are obtained from dissecting related structures. The assembly involves fitting the rigid bodies on the framework defined as the average of the C^{α} atoms in the conserved regions of the fold.
- Segment matching. Database of short protein fragments is used to calculate the coordinates of atoms based on positions of conserved atoms.

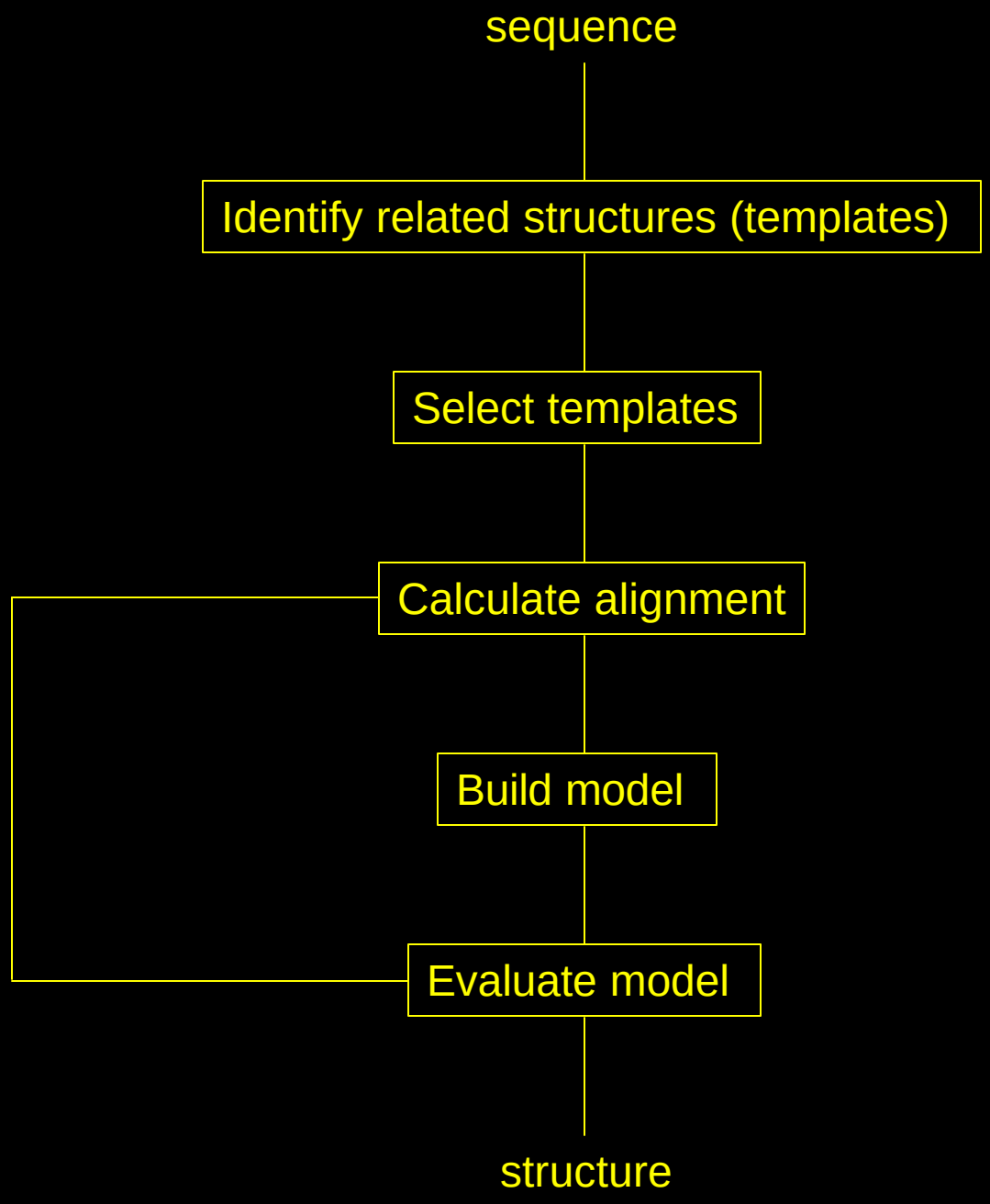


Model accuracy depends on sequence identity.


α
YKR066C
2cyp/1schA
β
YDR226W
1aky/1akeA
γ
YER148W
1ytbA/1cdwA


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%
4-8Å
~80aa
A
B
C
D
E
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
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching

MODELLER flowchart



Template selection

- High sequence similarity
- Select a template from a subfamily phylogenetically closest to query sequence
- Environment of target structure should be as similar as possible to the query. It may be solvent, ligand, pH, quarternary interactions.
- The quality of template structure should be as high as possible as described by resolution for X-ray, and number of constraints (NOE) for NMR.

Alignment.

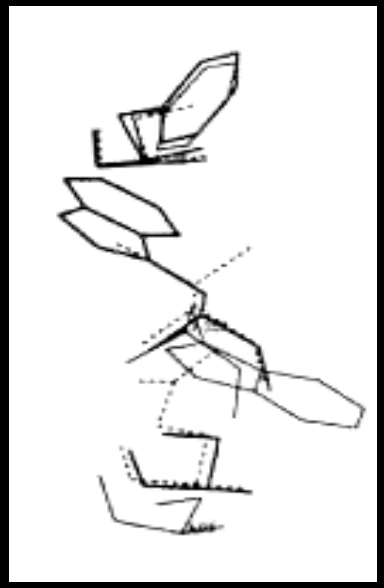
- When sequence similarity is above 40% alignment is simple to obtain.
- If sequence similarity is lower the alignment usually contains gaps and requires manual intervention in order to minimize the number of misaligned residues. Structural information from template usually helps: gaps should be avoided inside secondary structure elements, in buried regions and between residues far apart in space.

Model evaluation.

- Evaluating model is also assessing quality of the template as well as the correctness of an alignment. However in practice it is hard to tell which part is wrong …
- Energy profile may be used, but available force fields are by far not perfect themselves.
- Visual inspection of amino acid packing versus various physical/chemical properties (hydrophobicity, charge distribution etc) is usually the most important tool.
- Stereochemical properties (bond lengths, bond angles, atom-atom overlaps). They can be checked automatically (ProCheck).

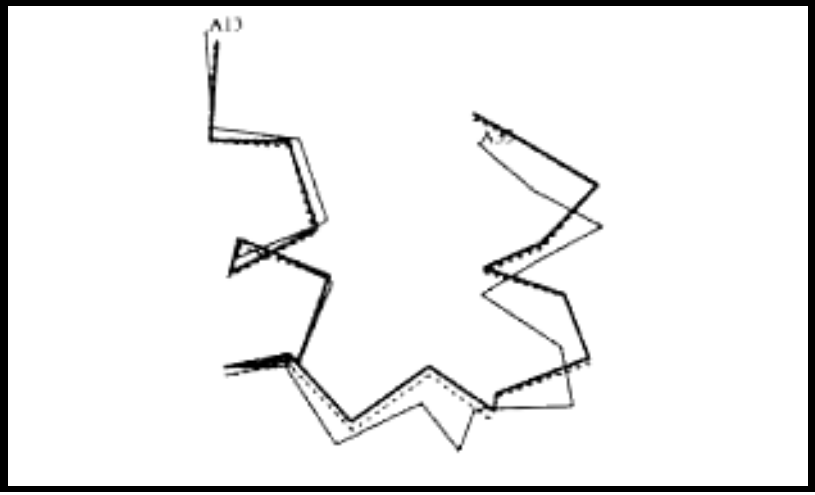
Problems and limitations.

As the similarity between query and template sequences decrease, the errors in model increase.

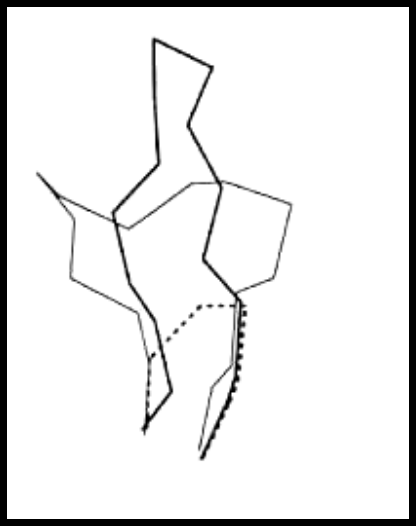


- Sidechain packing. Sometimes even the conformation of identical sidechains is not conserved. This is not very important unless these sidechains are close to the active site.

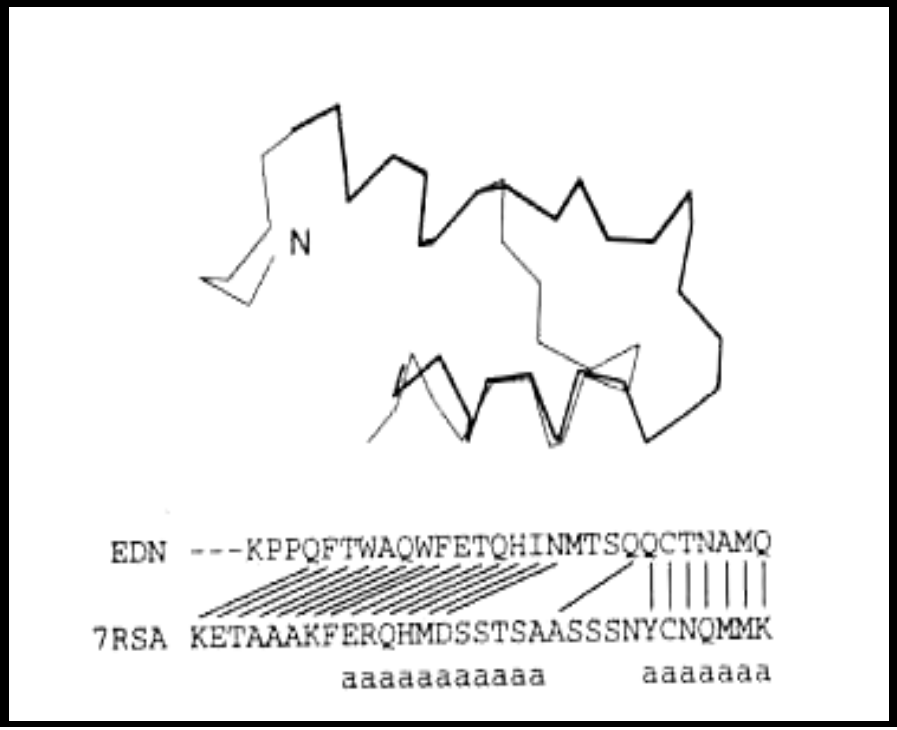
- Distortions and shifts in correctly aligned regions. As a consequence of sequence divergence the backbone conformation also changes even if the overall fold remains the same. This error can be also due to different environment.



Problems and limitations.



- Errors in regions without template (loops). Usually insertions of length 8 or less residues are possible to model (but with lower accuracy). Longer insertions are usually not modeled correctly and should be avoided.
- Misalignments. The largest source of errors, especially below 40% sequence identity. It can be avoided by using multiple sequence alignments instead of pairwise ones, and manual correction after model evaluation.



- Incorrect template. Model evaluation should eliminate improper templates.