(Aspekte der Thermodynamik in der Strukturbiologie)

Einführung in die Bioinformatik

Wintersemester 2012/13

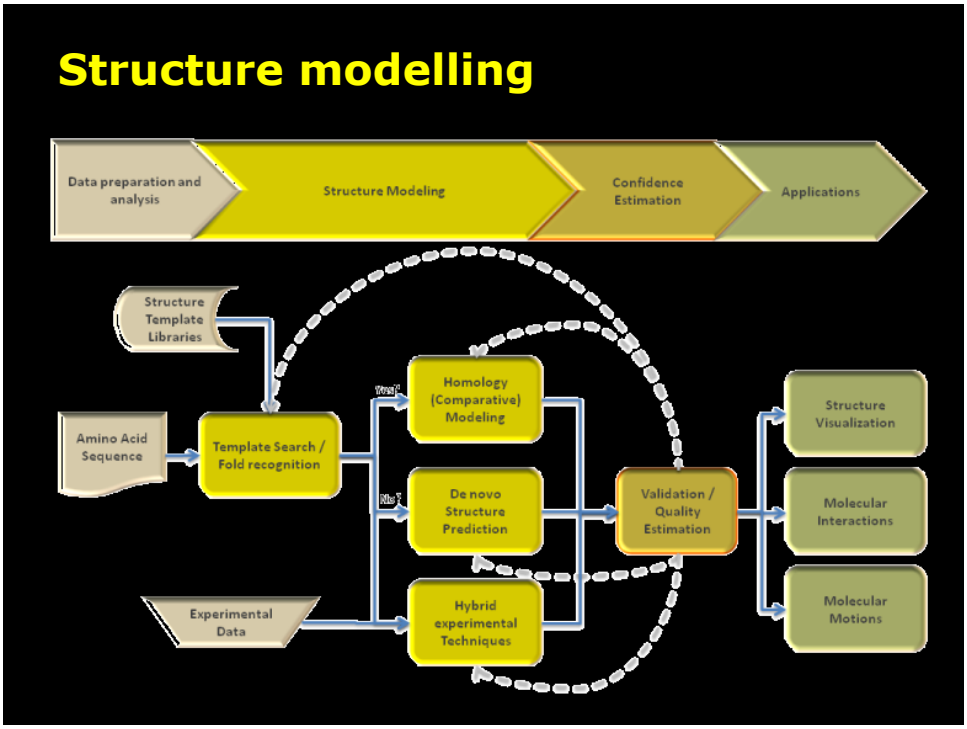
Peter Güntert

Structure modelling

Outline

- Structure modeling:
  - Homology modeling
  - Fold recognition
  - ab initio modeling

Structure modelling



Sequence identity → Structural similarity

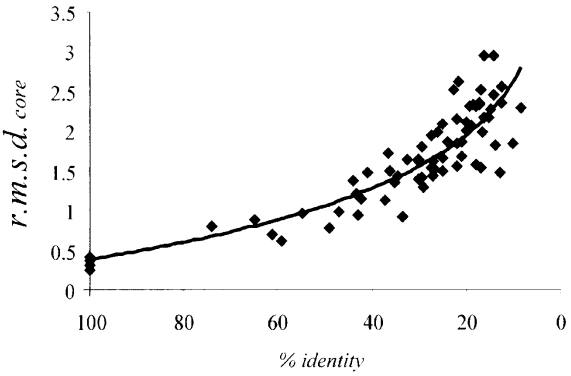


Figure 1.25 Relationships between sequence identity and structural similarity.

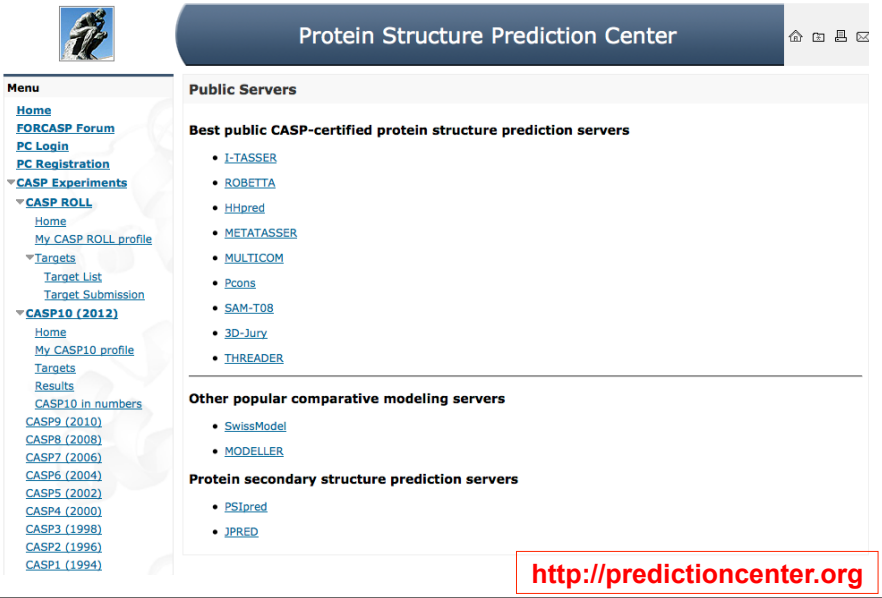
BUT:
Structural similarity ~~→~~ Sequence identity

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein and proteins of known structure:

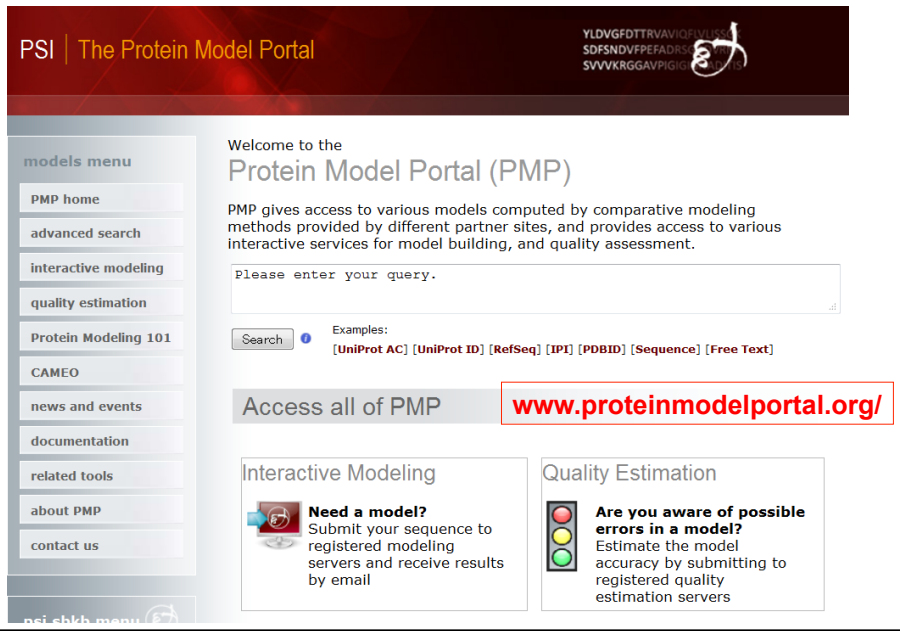
- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

Protein Structure Prediction Servers



http://predictioncenter.org

PSI Protein Model portal (PMP)

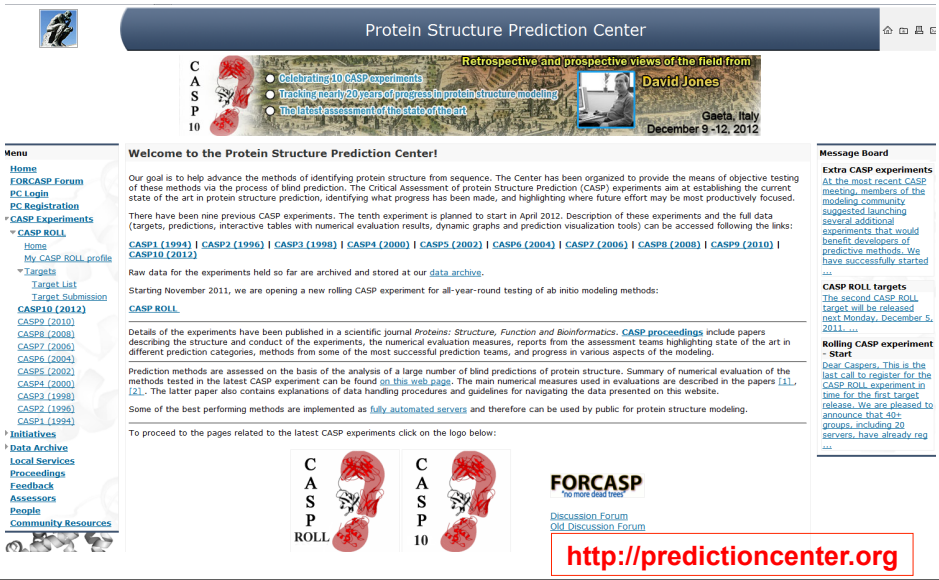


www.proteinmodelportal.org/

PSI Protein Model portal (PMP)

www.proteinmodelportal.org/

CASP: Critical Assessment of Structure Prediction



http://predictioncenter.org

CASP: Critical Assessment of Structure Prediction

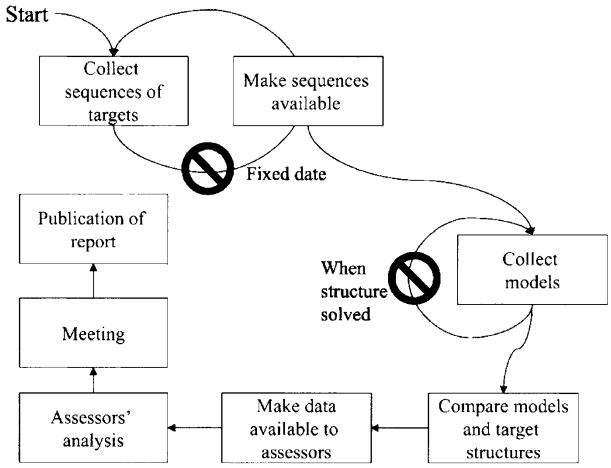


Figure 2.9 The CASP experiment runs every two years. In the spring, approximately, targets are collected from experimenters working on the resolution of their structure. The sequences are made available to predictors who can submit predictions until the structure is solved. Numerical comparison of models and targets is performed by a group of scientists led by John Moult and Krzystof Fidelis. The data are then passed to thee assessors, chosen by the community on the basis of their expertise, who analyze the data and try to derive general conclusions about the state of the art in the prediction field. In approximately December of the same year, predictors, assessors, and organizers convene in a meeting to discuss the results and, later, publish the final reports in the scientific journal *Proteins: Structure, Function and Bioinformatics*.

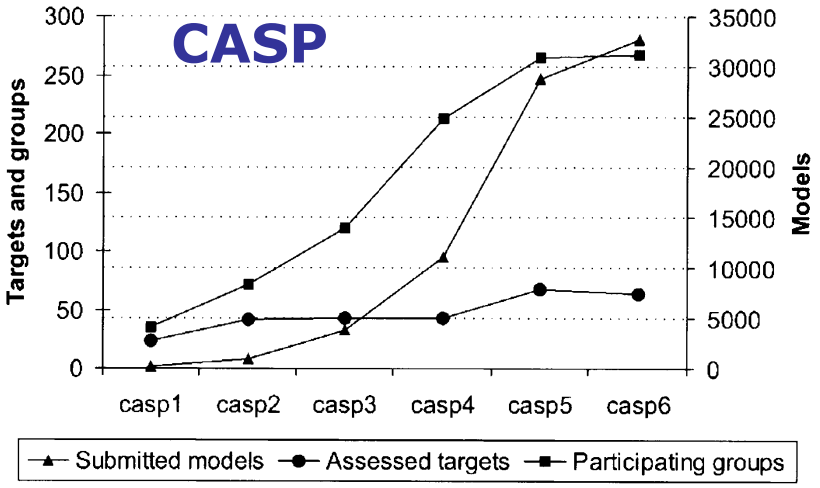
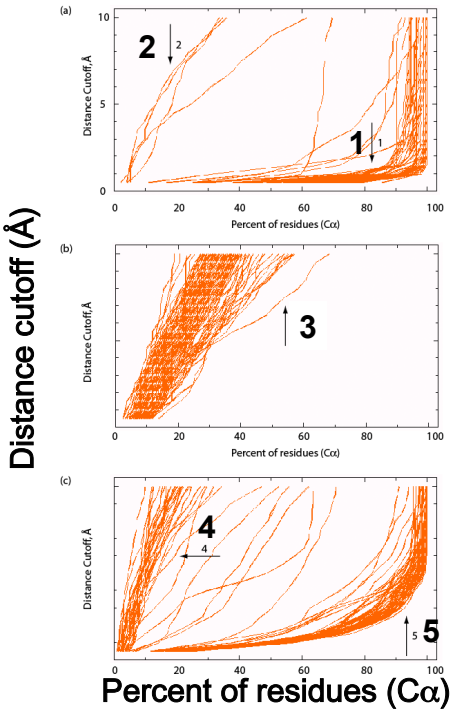


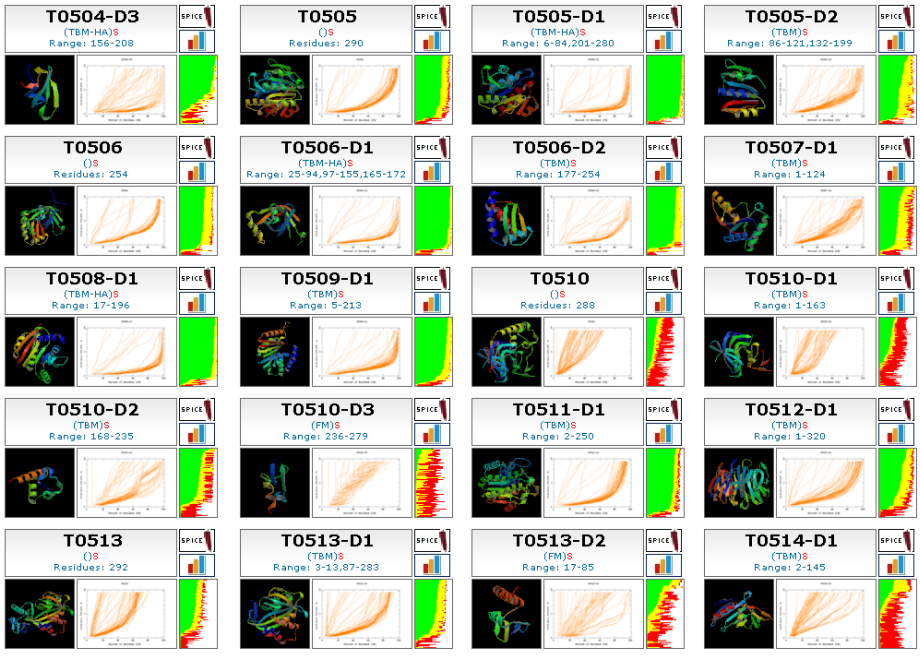
Figure 2.10 The plot shows the numbers of targets, participating groups, and models submitted to each of the editions of CASP from 1994 (CASP1) to 2004 (CASP6). All the thousands of models are publicly available on the CASP web site.



most groups had the right prediction for this structure (but not those at arrow 2)

most groups had the wrong prediction for this structure (but arrow 3 did better)

half the groups got it wrong (arrow 4), half got it right (arrow 5); the key difference is the multiple sequence alignment



Scheme of protein structure predicition

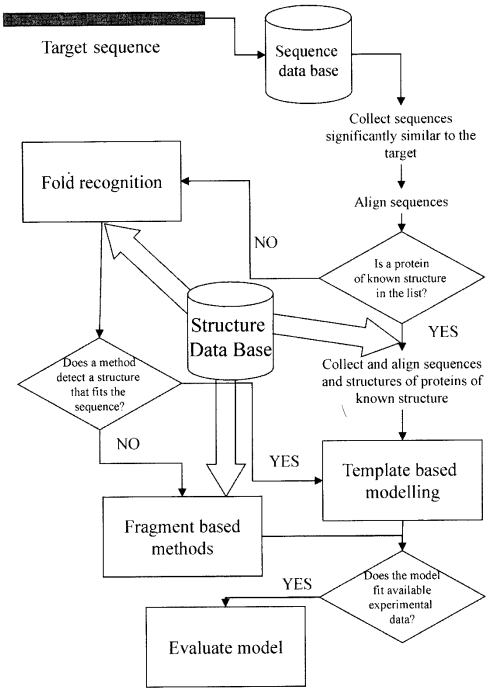


Figure 4.1 A guide to protein-structure prediction. The first step is always a search in the protein sequence database. Comparative modeling should be used when a protein of known structure sharing sequence similarity with the protein under examination is present in the database. If this is not so, fold-recognition methods should be applied and, should they fail, the user should resort to new fold or fragment-based methods. Note the central role played by the structure database in all these heuristic methods.

Comparative protein structure modelling

Classical procedure for construction of a homology model

1. Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
2. If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
3. Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
4. Model the regions outside the conserved core.
5. Model the positions of the side-chains of the target.
6. Optimize the final three-dimensional structure.

Scheme of comparative modelling

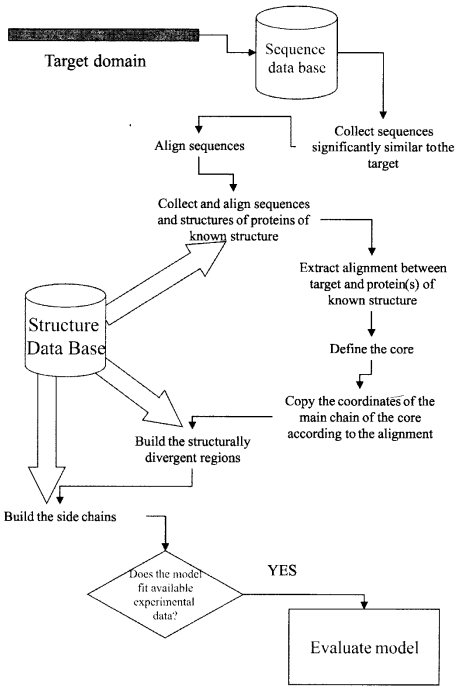


Figure 4.2 Schematic diagram of a typical comparative modeling procedure. The protein of interest should first be split into its domains. For each domain, sequences similar to the target sequences should be collected using a database search tool such as FASTA, BLAST, or PSI-BLAST. The sequences retrieved should be realigned using a multiple sequence alignment program (for example CLUSTAL or T-COFFEE). The implied alignment between the target protein and the protein(s) of known structure will form the basis of construction of the model. This can proceed by first building the main chain of the core regions, then the main chain of the structurally divergent regions, and, finally, the side-chains. The final evaluation of the model should take into account any available information on the protein of interest.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Building structurally divergent regions

- Reinspect alignment, e.g. shift gaps/insertions outside regular secondary structure elements
- Short canonical loops (type I, type II etc.)
- Rely on sequence pattern
- Loops that form compact substructures: internal H-bonds
- Packing inward pointing side-chain between secondary structure elements connected by the loop

Loops with similar conformation

Figure 4.16 The figure shows two loops with similar conformations stabilized by the packing of a central hydrophobic amino acid. Note that one of the loops connects two alpha helices and the other two beta strands.

Similar loops, different environment

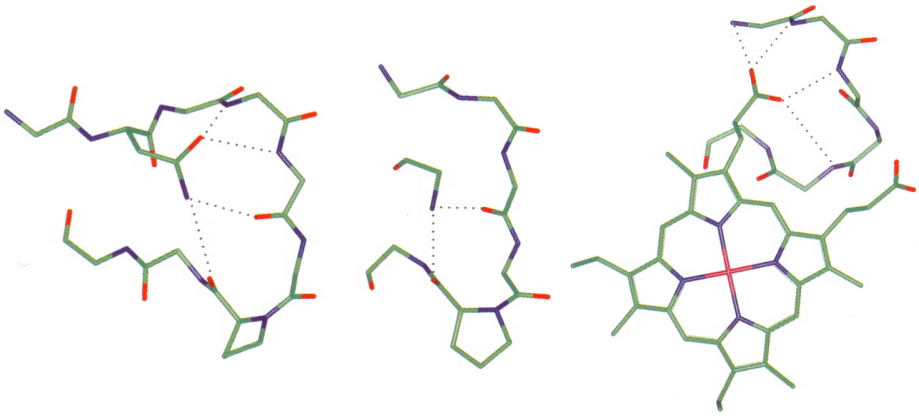


Figure 4.17 The three loops shown in the figure are very similar and stabilized by hydrogen-bonds, however the partners of these interactions are different in the three different proteins (an immunoglobulin, a viral protein, and a cytochrome).

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Classical procedure for construction of a homology model

- Given a protein of unknown structure, identify proteins of known structure that are evolutionarily related to it.
- If they exist, construct a reliable alignment, i.e. deduce the correspondence between related amino acids in the core, i.e. in regions other than those affected by insertions, deletions, and local refolding.
- Assign the coordinates of the backbone atoms of the corresponding amino acids of the target protein according to the sequence alignment.
- Model the regions outside the conserved core.
- Model the positions of the side-chains of the target.
- Optimize the final three-dimensional structure.

Difficulties of comparative modelling

- Identification of domain boundaries
- Identify correct template
- Find correct alignment between target and template sequence
- Prediction of loop structures
- Side-chain conformation prediction
- Energy refinement is not effective in finding a better model.
- Multi-domain proteins when using different templates for individual domains
- Active sites are better modeled than regions with less evolutionary constraints

Prediction accuracy

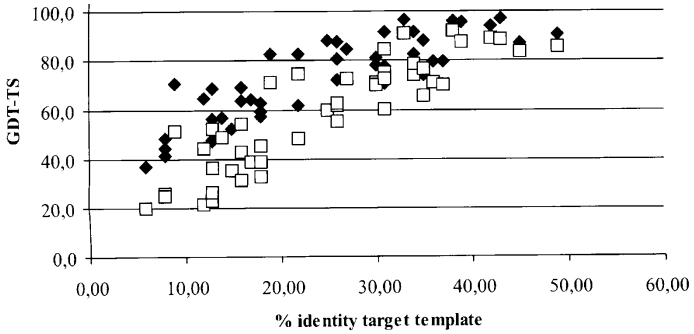


Figure 4.21 The relationship between the GDT-TS of the best (filled symbols) and average (open symbols) models and the sequence identity between the target protein sequence and the sequence of the best structural template. The data are taken from the CASP5 results and indicate that, above 40% sequence identity between target and template sequence, most methods can produce very respectable models. In more difficult examples the best methods can still produce useful results, but the gap between the quality of their results and those that can be obtained on average increases.

Comparative modelling examples

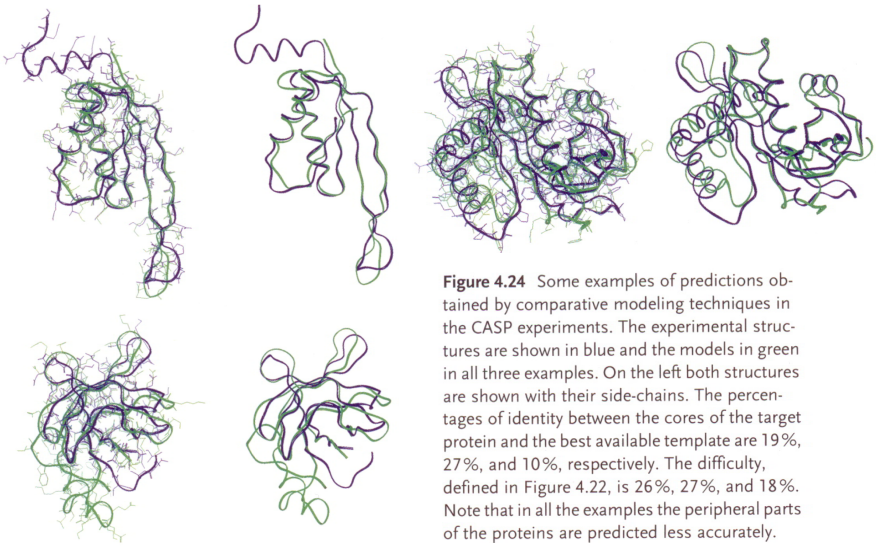


Figure 4.24 Some examples of predictions obtained by comparative modeling techniques in the CASP experiments. The experimental structures are shown in blue and the models in green in all three examples. On the left both structures are shown with their side-chains. The percentages of identity between the cores of the target protein and the best available template are 19%, 27%, and 10%, respectively. The difficulty, defined in Figure 4.22, is 26%, 27%, and 18%. Note that in all the examples the peripheral parts of the proteins are predicted less accurately.

Fold recognition

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein and proteins of known structure:

- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

Structural similarity ↛ Sequence identity

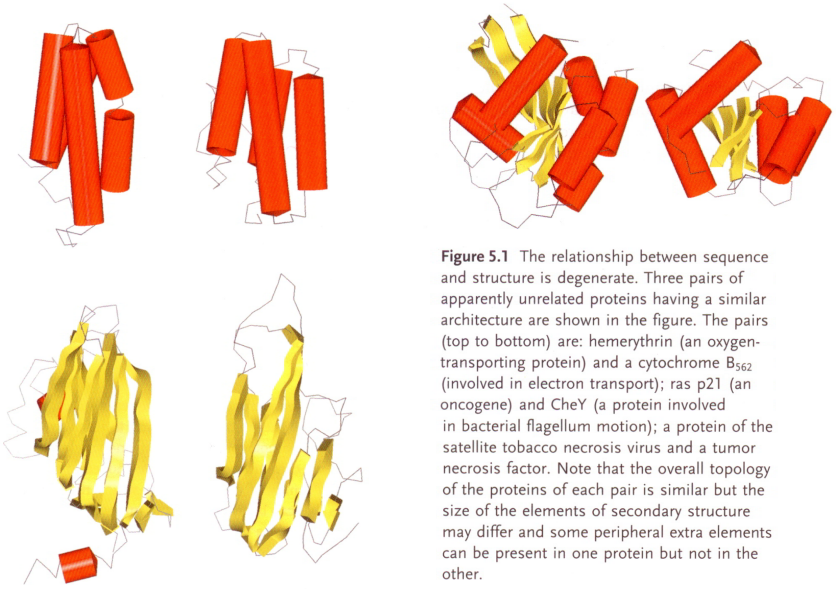


Figure 5.1 The relationship between sequence and structure is degenerate. Three pairs of apparently unrelated proteins having a similar architecture are shown in the figure. The pairs (top to bottom) are: hemerythrin (an oxygen-transporting protein) and a cytochrome B_{562} (involved in electron transport); ras p21 (an oncogene) and CheY (a protein involved in bacterial flagellum motion); a protein of the satellite tobacco necrosis virus and a tumor necrosis factor. Note that the overall topology of the proteins of each pair is similar but the size of the elements of secondary structure may differ and some peripheral extra elements can be present in one protein but not in the other.

Non-uniform distribution of folds

- Few (~10) folds are shared by a large number (~30%) of known proteins
- Large diversity in sequences and functions among members of these "superfolds"

Examples:

- Immunoglobulin fold
- Rossman fold
- TIM barrel fold
- Globin fold

Inverse protein folding problem

Which amino acid sequences fold into a known three-dimensional structure?

Protein folding problem

Which three-dimensional structure is adopted by a given amino acid sequence?

Fold recognition methods

- **3D profile methods**
  Physico-chemical properties of the amino acids of the target protein must "fit" with the environment in which they are placed in the modeled structure.
- **Threading**
  Sequences are fitted directly onto the backbone coordinates of known protein structures.

Profile method for fold recognition

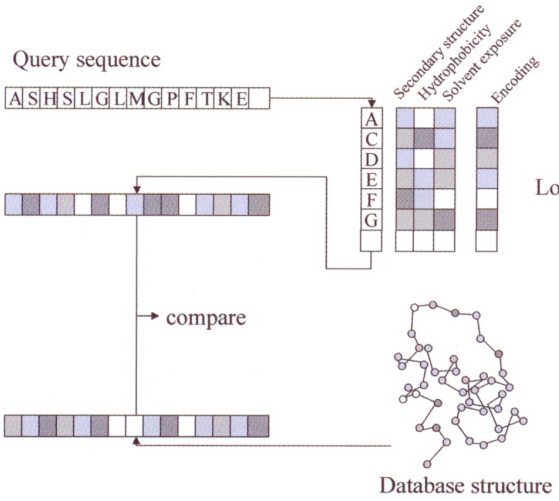


Figure 5.2 Schematic diagram of a possible profile-based method for fold recognition. The amino acids of the query sequence are replaced by a code that summarizes their hydrophobicity and their propensity for secondary structure type and solvent exposure. Each structure in the database is also encoded as a string by assigning a code to each of its amino acid positions. The code reflects their structural environment (secondary structure, solvent accessibility, and hydrophobicity of their envir onment). This does not depend on the actual amino acid present in the position analyzed. The string encoding the query sequence and each of the strings encoding the database structures are aligned and compared.

Bowie, Lüthy & Eisenberg. *Science* 253, 164-170 (1991)

Threading

- Sequences are fitted directly onto the backbone coordinates of known protein structures.
- Matching of sequences to backbone coordinates is performed in 3D space, incorporating specific pair interactions explicitly.

A new approach to protein fold recognition

D. T. Jones*†, W. R. Taylor† & J. M. Thornton*

* Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular Biology,
University College, Gower Street,
London WC1E 6BT, UK
† Laboratory of Mathematical Biology, National Institute for Medical Research,
The Ridgeway, Mill Hill, London, NW7 1AA, UK

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs[1] or folds[2–4] in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

Nature 358, 86-89 (1992)

Threading

- A library of different protein folds is derived from the database of protein structures.
- Each fold is considered as a chain tracing through space; the original sequence being ignored completely.
- The test sequence is then optimally fitted to each library fold, allowing for relative insertions and deletions in loop regions.
- The 'energy' of each possible fit (or threading) is calculated by summing the proposed pairwise interactions and the solvation energy.
- The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match.

Knowledge-based (pair) potentials

$$E(r) = -k_B T \ln[f(r)]$$

r	distance between two atoms (or some other parameter, like dihedral angles or solvent accessible surface)
$E(r)$	is the energy at r
$f(r)$	is the probability density at r
k_B	is the Boltzmann constant
T	is the absolute temperature

Statistically derived potentials

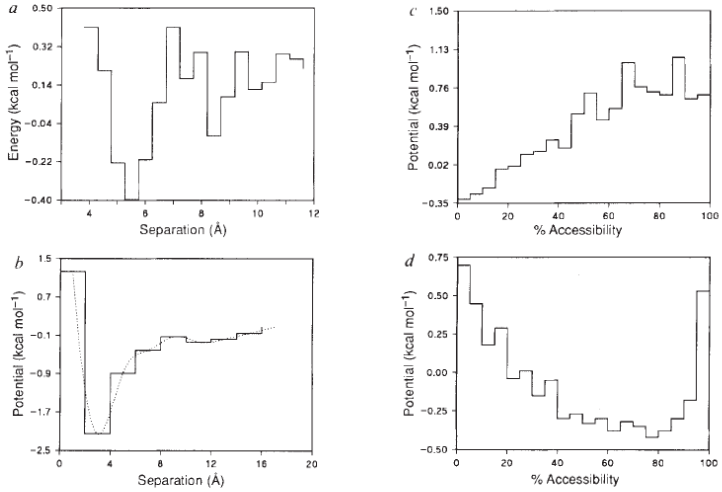


FIG. 1 Samples of the statistically derived potentials are shown. *a*, Short-range ($k = 3$) Ala-Ala $C\beta \rightarrow C\beta$ interaction. Low-energy states are observed for distances around 6 Å, corresponding mainly to α-structure, and 9 Å, corresponding mainly to β-structure. *b*, Long-range ($k > 30$) Cys-Cys $C\beta \rightarrow C\beta$ interaction. The most significant energy minimum around 4 Å corresponds to disulphide bridge formation. *c*, Solvation potential for leucine, and *d*, solvation potential for glutamic acid.

Fold recognition results from CASP

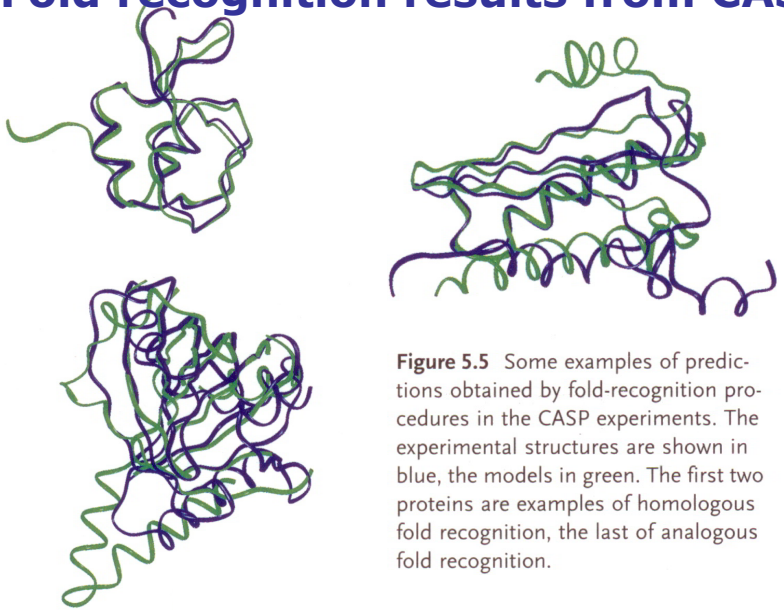


Figure 5.5 Some examples of predictions obtained by fold-recognition procedures in the CASP experiments. The experimental structures are shown in blue, the models in green. The first two proteins are examples of homologous fold recognition, the last of analogous fold recognition.

New fold prediction

Methods for protein structure prediction

Methods are distinguished according to the relationship between the target protein(s) and proteins of known structure:

- **Comparative modelling**: A clear evolutionary relationship between the target and a protein of known structure can be easily detected from the sequence.
- **Fold recognition:** The structure of the target turns out to be related to that of a protein of known structure although the relationship is difficult, or impossible, to detect from the sequences.
- **New fold prediction:** Neither the sequence nor the structure of the target protein are similar to that of a known protein.

CASP: Fragment-based predictions

Figure 6.2 Some examples of fragment-based predictions submitted to CASP experiments.

Fragment-based approaches

- Rosetta (David Baker)
- Fragfold (David Jones)

Toward High-Resolution de Novo Structure Prediction for Small Proteins

Philip Bradley, Kira M. S. Misura, David Baker*

The prediction of protein structure from amino acid sequence is a grand challenge of computational molecular biology. By using a combination of improved low- and high-resolution conformational sampling methods, improved atomically detailed potential functions that capture the jigsaw puzzle–like packing of protein cores, and high-performance computing, high-resolution structure prediction (<1.5 angstroms) can be achieved for small protein domains (<85 residues). The primary bottleneck to consistent high-resolution prediction appears to be conformational sampling.

Science 309, 1868–1871 (2005)

Steps of fragment-based structure prediction

- Split sequence into fragments
- For each fragment, search the database of known structures for regions with a similar sequence ("neighbors")
- Use an optimization technique to find the best combination of fragments

Fragment search

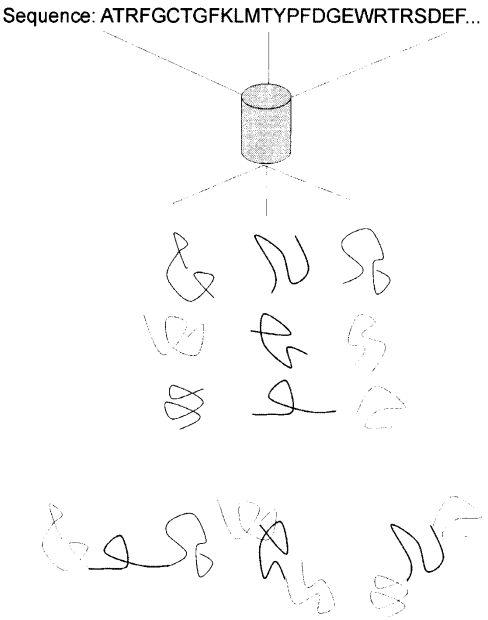
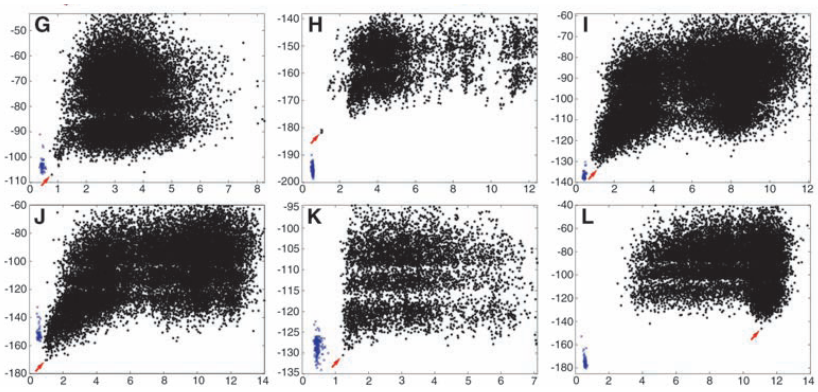


Figure 6.3 Schematic explanation of the first steps of the Rosetta method. The query sequence is split in fragments nine amino acids long. Each fragment sequence is used to search for similar fragments among the sequences of proteins of known structure. Next, the fragments are joined.

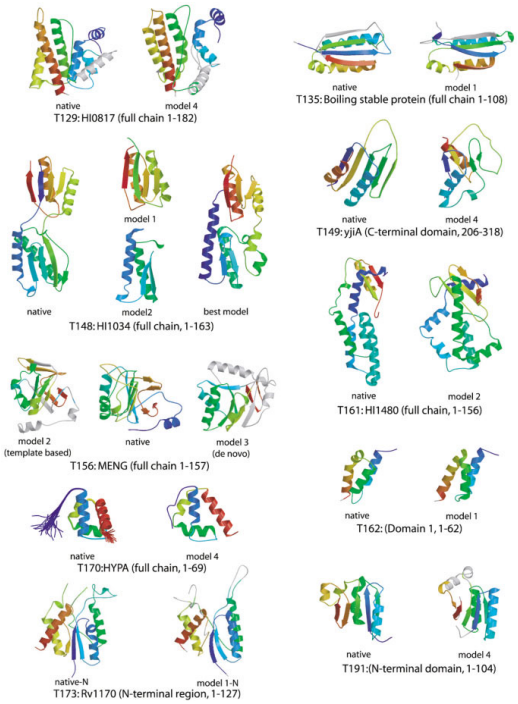
Energy vs. accuracy



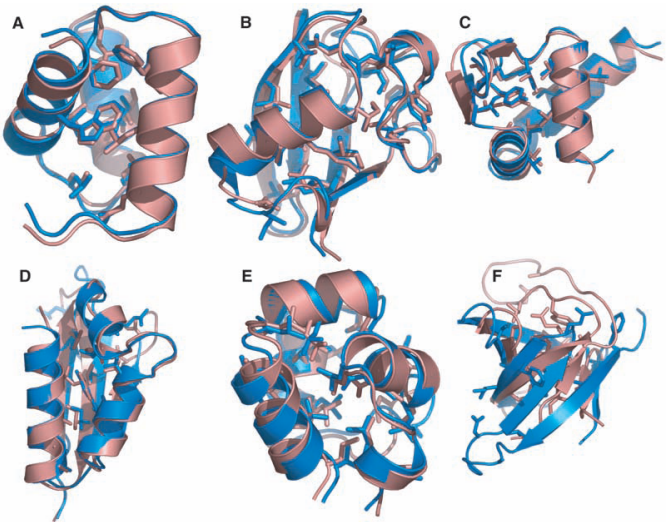
Plots of C^{α}-RMSD (x axis) against all atom energy (y axis) for refined natives (blue points) and the de novo models (black points). Red arrows indicate the lowest energy de novo models.

ROSETTA results in CASP5

Ribbon diagrams of predictions made by using the fragment insertion approach. The native structure and best submitted model are shown colored from the N-terminus (blue) to C-terminus (red). For T148, the best generated model is also shown, and for T156, both template-based and fragment insertion based models are shown. For targets T173, T135, T156, and T191, colored regions deviate from the native structure by <4 Å, and gray regions deviate by >4 Å. For targets T129 and T156, colored regions deviate from the native structure by <6 Å C^{α} RMSD, whereas the gray regions deviate by >6 Å.



High-resolution de novo structure predictions



Superposition of low-energy models (blue) with experimental structures (red) showing core side chains.

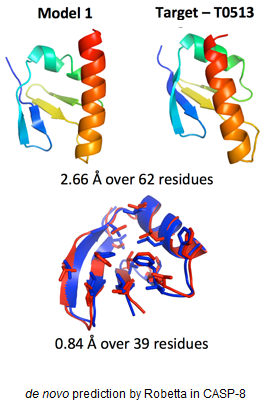
A: Hox-B1
B: Ubiquitin
C: RecA
D: KH domain of Nova-2
E: 434 repressor
F: Fyn tyrosine kinase

ROBETTA BETA

Full-chain Protein Structure Prediction Server

www.bakerlab.org

Robetta protein structure prediction server



de novo prediction by Robetta in CASP-8

http://robetta.bakerlab.org

REGISTRATION
[Register / Update] [Login]

DOCUMENTATION
[Docs / FAQs]

SERVICES
Domain Parsing & 3-D Modeling
[Queue] [Submit]

Interface Alanine Scanning
[Queue] [Submit]

Fragment Libraries
[Queue] [Submit]

DNA Interface Residue Scanning
[Queue] [Submit]

RELATED SITES
RosettaBackrub Server
RosettaAntibody Server
RosettaDesign Server
RosettaDock Server
Rosetta Commons
FoldIt
Rosetta@home
Human Proteome Folding Project

Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching