Computergestützte Strukturbiologie
(Strukturelle Bioinformatik)

Comparative protein structure modeling

Sommersemester 2009

Peter Güntert

Inference of function from structure

One can expect to gain insight into a protein's function from analysis of other, structurally similar proteins. There are at least three difficulties to overcome in this process:

- Homologous proteins might have originated by gene duplication and subsequent evolution and therefore have acquired a different function.
- Some folds are adopted by proteins performing a variety of functions.
- The protein of interest might have a novel, not yet observed fold.

Sequence identity → Structural similarity

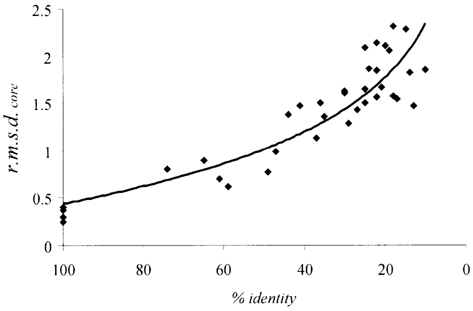


Figure 1.23 Relationship between sequence identity and structural similarity. The plot is obtained using the same set of proteins originally analyzed by Lesk and Chothia.

Sequence identity → Structural similarity

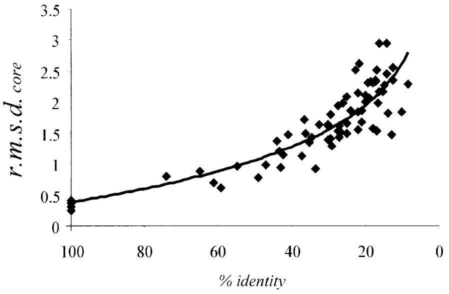


Figure 1.25 Relationships between sequence identity and structural similarity. The plot was obtained by using a larger set of proteins than in Figure 1.23, but the trend is essentially the same.

Methods for protein structure prediction

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

- **Comparative modeling**: 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: Critical Assessment of Structure Prediction

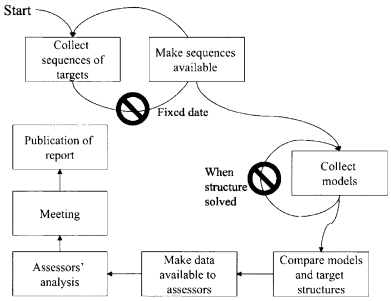


Figure 2.9 The CASP experiment runs every two years. In the spring, approximately, targets are collected from experimenters working on the resolution of their structure. The sequences are made available to predictors who can submit predictions until the structure is solved. Numerical comparison of models 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.*

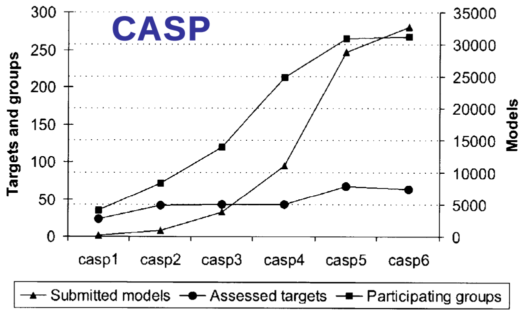


Figure 2.10 The plot shows the numbers of targets, participating groups, and models submitted to each of the editions of CASP from 1994 (CASP1) to 2004 (CASP6). All the thousands of models are publicly available on the CASP web site.

Scheme of protein structure predicition

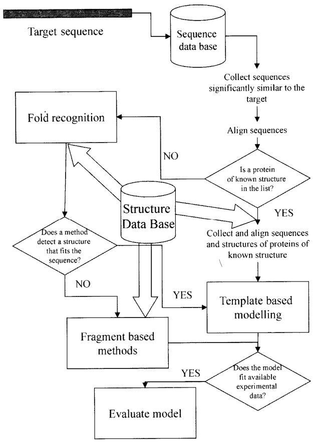

Target sequence
Sequence data base
Collect sequences significantly similar to the target
Align sequences
Is a protein of known structure in the list?
NO
YES
Fold recognition
Structure Data Base
Collect and align sequences and structures of proteins of known structure
Does a method detect a structure that fits the sequence?
NO
YES
Template based modelling
Fragment based methods
Does the model fit available experimental data?
YES
Evaluate model

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.

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.

Scheme of comparative modeling

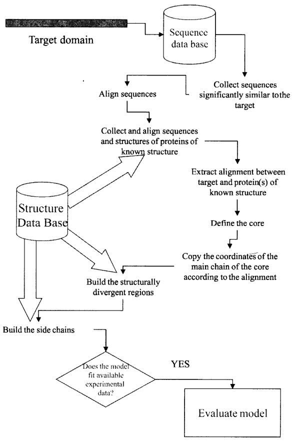

Target domain
Sequence data base
Collect sequences significantly similar tothe target
Align sequences
Collect and align sequences and structures of proteins of known structure
Extract alignment between target and protein(s) of known structure
Define the core
Copy the coordinates of the main chain of the core according to the alignment
Structure Data Base
Build the structurally divergent regions
Build the side chains
Does the model fit available experimental data?
YES
Evaluate model

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.

Refinement of initial alignment

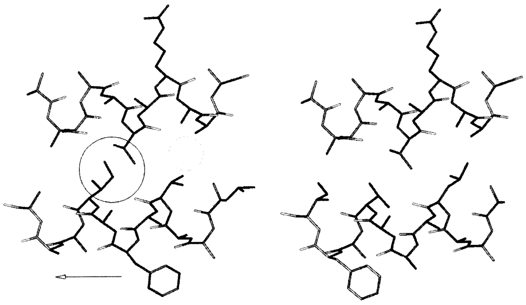
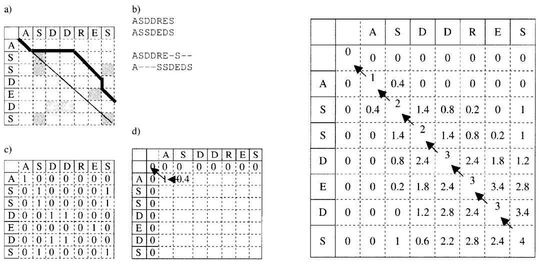


Figure 4.3 The model-building procedure might help refining the initial alignment. In the example shown here it is apparent that a shift of one residue of the sequence towards the carboxy-terminus results in a much better packing of the side-chains.

Needleman-Wunsch alignment algorithm



	A	S	D	D	R	E	S
A	1	0	0	0	0	0	0
S	0	1	0	0	0	0	1
S	0	1	0	0	0	0	1
D	0	0	1	1	0	0	0
E	0	0	0	0	0	1	0
D	0	0	1	1	0	0	0
S	0	1	0	0	0	0	1

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	1	0.4					
S	0							
S	0							
D	0							
E	0							
D	0							
S	0							

		A	S	D	D	R	E	S
	0	0	0	0	0	0	0	0
A	0	1	0.4	0	0	0	0	0
S	0	0.4	2	1.4	0.8	0.2	0	1
S	0	0	1.4	2	1.4	0.8	0.2	1
D	0	0	0.8	2.4	3	2.4	1.8	1.2
E	0	0	0.2	1.8	2.4	3	3.4	2.8
D	0	0	0	1.2	2.8	2.4	3	3.4
S	0	0	1	0.6	2.2	2.8	2.4	4

Figure 4.4 The Needleman and Wunsch alignment algorithm. A path in the matrix corresponds to an alignment. In the example, the thin line in part *a* of the figure corresponds to the first alignment shown in part *b*. The line runs diagonally and therefore corresponds to an alignment where there are no insertions or deletions. The tick line, instead, contains an horizontal line (indicating that the amino acids SDD of the first sequence do not correspond to any amino acid of the second and therefore represent an insertion in the first sequence) and two vertical lines (implying that the amino acid D and the final DS pair of the second sequence do not correspond to any amino acid in the first and is an insertion in the second sequence or, equivalently, a deletion in the first). To compute the optimum alignment we fill the cells of the matrix (part *c*) with a number representing the likelihood that the amino acid in the row is replaced by that in the column. In this example we assign 1 to identical amino acids and 0 to different ones. Part *d* shows the construction of the cumulative matrix as described in the text.

Amino acid substitution matrices

Pam 250

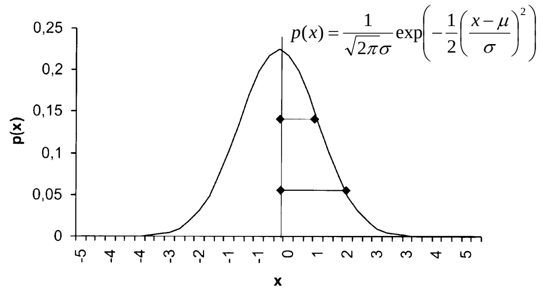
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Blosum 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 4.5 The PAM250 (part *a*) and BLOSUM62 (part *b*) substitution matrices. The values corresponding to pairs of amino acids can be used to fill the alignment matrix (part *c* of Figure 4.4).

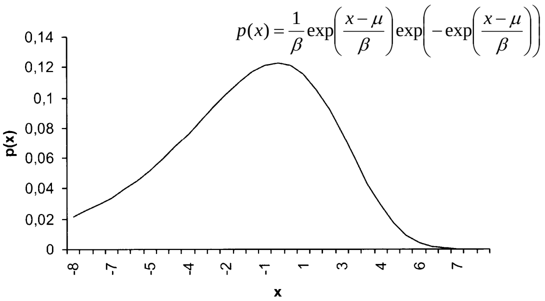
Gaussian probability distribution



$$p(x)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Figure 4.6 A Gaussian distribution with mean = 0 and σ = 1. The two segments correspond to one and two standard deviations.

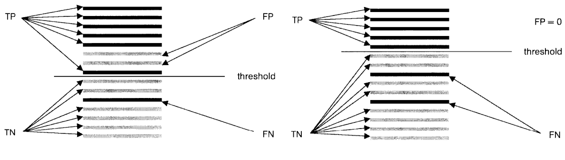
Extreme value distribution



$$p(x)=\frac{1}{\beta}\exp\left(\frac{x-\mu}{\beta}\right)\exp\left(-\exp\left(\frac{x-\mu}{\beta}\right)\right)$$

Figure 4.7 Extreme value distribution with μ = 0 and β = 3. This is the expected distribution for the alignment scores of unrelated sequences.

Sensitivity and specificity



Sensitivity = 6/7 = 0.86
Specificity = 6/8 = 0.75

Sensitivity = 5/7 = 0.71
Specificity = 8/8 = 1.00

Sensitivity = TP / (TP + FN)   Specificity = TN / (TN + FP)

Figure 4.8 Examples of sensitivity and specificity values for a database search method. In the figure, dark and light segments, respectively, represent proteins homologous and unrelated to the query sequence. If we select the threshold as shown in the top part of the figure, two unrelated sequences will be labeled as "homologous" and one homologous one as "unrelated". A more stringent threshold (bottom), will eliminate false positives, but will increase the number of false negatives.

True positives vs. false negatives

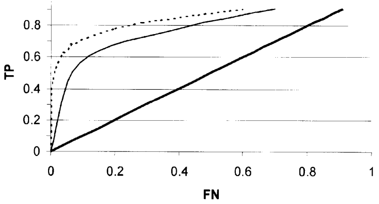


Figure 4.9 Examples of ROC curves. The tick line corresponds to a worthless method, unable to discriminate between positives and negatives. The method represented by the dotted curve is better than that represented by the continuous line: it detects more true positives when finding the same number of false negatives.

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.

Domains (BLAST)

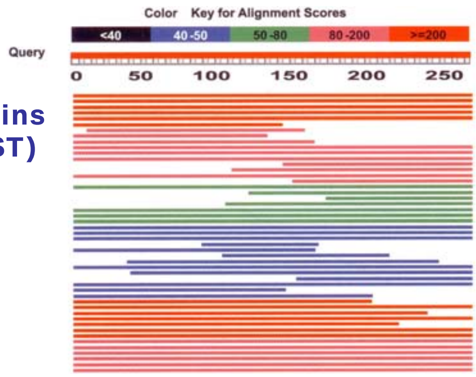
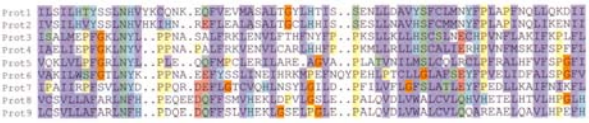


Figure 4.10 Example of the graphical output of BLAST. The example shown suggests that the query protein is formed by two domains, one spanning from the beginning to approximately residue 150, the other from approximately residue 150 to the end of the protein.

Multiple sequence alignment



```
Prot1 ILSILHTYSSLNHVYKCQNK.EQFVEVMASALTGYLHTIS..SENLLDAVYSFCLMNYFPLAPFNQLLQKDII
Prot2 IVSILHVYSSLNHVHKIHN..REFLEALASALTGCLHHIS..SESLLNAVHSFCMMNYFPLAPINQLIKENII
Prot3 ISALMEPFGKLNYL..PPNA.SALFRKLENVLFTHFNYFP..PKSLLKLLHSCSLNECHPVNFLAKIFKPLFL
Prot4 IAELIEPFGKLNYV..PPNA.PALFRKVENVLCARLHHFP..PKMLLRLLHSCALIERHPVNFMSKLFSPFFL
Prot5 VQKLVLPFGRLNYL..PLE..QQFMPCLERILARE.AGVA..PLATVNILMSLCQLRCLPFRALHFVFSPGFI
Prot6 VAKILWSFGTLNYK..PPNA.EEFYSSLINEIHRKMPEFNQYPEHLPTCLLGLAFSEYFPVELIDFALSPGFV
Prot7 IPAIIRPFSVLNYD..PPQR.DEFLGTCVQHLNSYLGILD..PFILVFLGFSLATLEYFPEDLLKAIFNIKFL
Prot8 VCSVLLAFARLNFH..PEQEEDQFFSMVHEKLDPVLGSLE..PALQVDLVWALCVLQHVHETELHTVLHPGLH
Prot9 LCSVLLAFARLNFH..PDQE.DQFFSLVHEKLGSELPGLE..PALQVDLVWALCVLQQAREAELQAVLHPEFH
```

Figure 4.11 A multiple sequence alignment. Note that completely conserved amino acids are easier to detect when more sequences are considered.

Alignment of

PTLRS
LTTRS with PTLR:

	P	T	L	R	S
	L	T	T	R	S
P	(Score (P,P) + score (L,P))/2	(Score (T,P) + score (T,P))/2	…	…	…
T	(Score (P,T) + score (L,T))/2	(Score (T,T) + score (T,T))/2	…	…	…
L	(Score (P,L) + score (L,L))/2	(Score (T,L) + score (T,L))/2	…	…	…
R	(Score (P,R) + score (L,R))/2	(Score (T,R) + score (T,R))/2	…	…	…

	P	T	L	R	S
	L	T	T	R	S
P	(7-3)/2=2	…	…	…	…
T	(-1-1)/2=-1	…	…	…	…
L	(-3+4)/2=0.5	…	…	…	…
R	(-2-2)/2=-2	…	…	…	…

Figure 4.12 The method for aligning a sequence to an alignment. The alignment is written in the first rows of a matrix and the sequence in the first column. Each cell contains the average between the score of each amino acid of the alignment with the corresponding amino acid of the sequence. The alignment strategy, once the matrix is filled, is identical with that outlined in Figure 4.4.

Score of a multiple alignment

Sum of pairs score for the alignment :
PTLRS
LTTRS
PTLRT

P	T	L	R	S
L	T	T	R	S
P	T	L	R	T
Score (P,L) + Score (P,P) + Score (L,P) = -3+7-3=1	Score (T,T) + Score (T,T) + Score (T,T) = 5+5+5=15	Score (L,T) + Score (L,L) + Score (T,L) = -1+4-1=2	Score (R,R) + Score (R,R) + Score (R,R) = 5+5+5=15	Score (S,S) + Score (S,T) + Score (S,T) = 4+1+1=6
Score = 39				

Figure 4.13 The score of a multiple alignment can be computed by averaging the scores of each column, as shown in the figure.

Sequence similarity tree

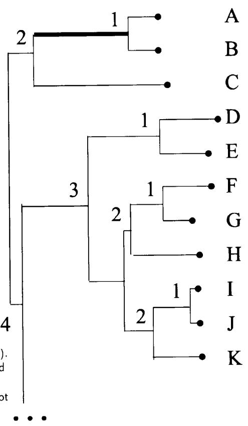


Figure 4.14 A tree constructed on the basis of the sequence similarity among several proteins (indicated by the filled circles). The numbers indicate the order in which the sequences should be iteratively aligned by use of the method described in Figure 4.12, starting from the leaves and proceeding toward the root of the tree.

Hidden Markov model

a)

Multiple sequence alignment:

```
A C C -  E
E C E -  A
A C E A  A
C -  E -  E
```

Counts	Begin-1	1-2	2-3	3-4	4-5	5-end
Match-match	4+1	3+1	3+1	1+1	1+1	4+1
Match-delete	0+1	0+1	0+1	3+1	0+1	0+1
Match-insert	0+1	1+1	0+1	0+1	0+1	0+1
Insert-match	0+1	0+1	0+1	0+1	0+1	0+1
Insert-delete	0+1	0+1	0+1	0+1	0+1	0+1
Insert-insert	0+1	0+1	0+1	0+1	0+1	0+1
Delete-match	0+1	0+1	1+1	0+1	3+1	0+1
Frequencies						
Match-match	0.45	0.36	0.36	0.18	0.18	0.45
Match-delete	0.09	0.09	0.09	0.36	0.09	0.09
Match-insert	0.09	0.18	0.09	0.09	0.09	0.09
Insert-match	0.09	0.09	0.09	0.09	0.09	0.09
Insert-delete	0.09	0.09	0.09	0.09	0.09	0.09
Insert-insert	0.09	0.09	0.09	0.09	0.09	0.09
Delete-match	0.09	0.09	0.18	0.09	0.36	0.09

Counts						
A	2+1	0+1	0+1	1+1	2+1	
C	1+1	3+1	1+1	0+1	1+1	
E	1+1	0+1	3+1	0+1	2+1	
Frequencies						
A	0.43	0.17	0.14	0.5	0.37	
C	0.29	0.67	0.28	0.25	0.25	
E	0.29	0.17	0.57	0.25	0.37	

b)

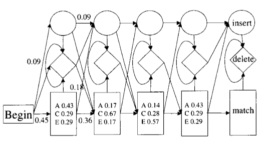


Figure 4.15 Construction of a hidden Markov model. Given a multiple sequence alignment, we first count how many times each transition (Match–match, Match–delete, Match–insert, Insert–match, Insert–delete, Insert–insert, Delete–match) occurs, and add 1 to each. The counts are then transformed into frequencies and used to construct the scheme shown in part *b*. For each match, we also count how many times each amino acid is observed (always adding 1 to the counts) and compute the frequencies. The scheme represents the hidden Markov model of the family of aligned proteins and can be used to calculate the probability that a new sequence belongs to the family, i.e. is generated by the HMM.

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

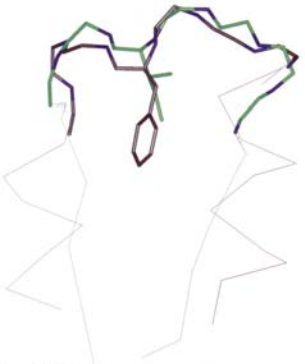


Figure 4.16 The figure shows two loops with similar conformations stabilized by the packing of a central hydrophobic amino acid. Note that one of the loops connects two alpha helices and the other two beta strands.

Similar loops, different environment

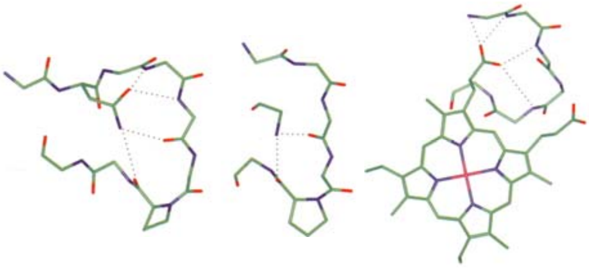


Figure 4.17 The three loops shown in the figure are very similar and stabilized by hydrogen-bonds, however the partners of these interactions are different in the three different proteins (an immunoglobulin, a viral protein, and a cytochrome).

Immunoglobulin antigen binding loops

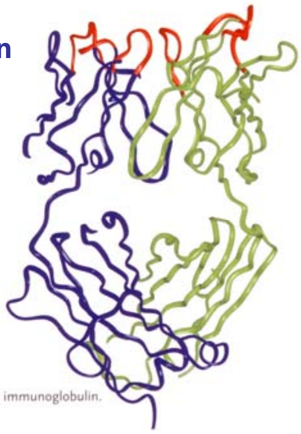


Figure 4.18 The structure of a fragment of an immunoglobulin. The antigen binding loops are shown in red.

Immunoglobulin structures

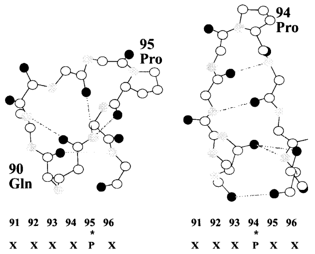


Figure 4.19 The canonical structures of immunoglobulins. The loop shown in the figure is called L3 (it is the third loop of the light (L) chain of antibodies and is part of the antigen-binding site). When the length of the loop is six amino acids, as in the figure, only two main chin conformations are observed. The one on the left occurs when the amino acid in position 95 is a proline. The conformation shown on the right instead occurs when the proline is in position 94. All other residues are free to vary and contribute to the shape of the antigen binding region.

H2 hairpin loops of three immunoglobulins

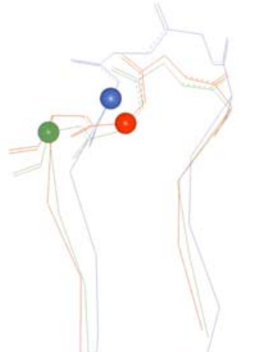


Figure 4.20 Superposition of the H2 hairpin loops of three immunoglobulins. Their conformation does not follow the rules relating sequence and structure in hairpin loops. The determinant of their conformation is the type of amino acid that occupies position 71, and not the position of the glycine (indicated by the sphere in the figure).

1NCD	T	N	T	G
2FBJ	P	D	S	G
2FB4	D	G	S	D

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.

Other approaches

- Construct the complete models on the basis of spatial constraints, i.e. compute a set of distance and dihedral angle probability distributions that must be satisfied by the final models and then build the models that are compatible with these distributions (Modeller).
- Construct several models for each target protein and selecting the most likely only at the end of the complete model-building procedure.

Difficulties of comparative modeling

- 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

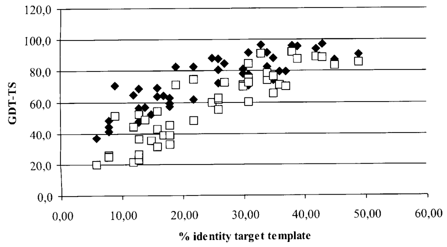


Figure 4.21 The relationship between the GDT-TS of the best (filled symbols) and average (open symbols) models and the sequence identity between the target protein sequence and the sequence of the best structural template. The data are taken from the CASP5 results and indicate that, above 40% sequence 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.

Alignment difficulty measure

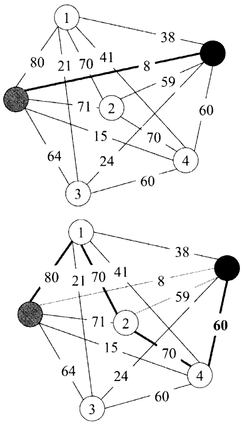


Figure 4.22 Graphical scheme of a method for evaluating the difficulty of aligning two protein sequences when a multiple sequence alignment is available. In the scheme, each circle represents a protein and each edge is labeled with the sequence identity between the two connected proteins. Assume that the gray circle is the target protein and the black circle the template. The sequence identity between the two protein sequences is only 8%. We can, however, progressively align the proteins following the path indicated by the ticker lines in the lower part of the figure. In this instance the most difficult alignment that we are forced to perform is that between the protein labeled "4" and the template sequence.

Best alignment vs. alignment difficulty

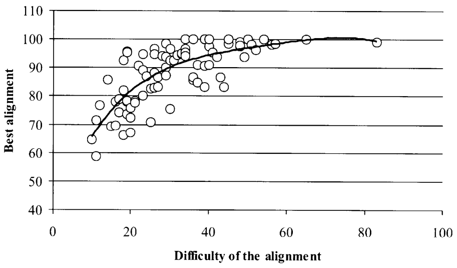


Figure 4.23 Relationship between the difficulty of aligning a target and template protein sequences, computed as described in the legend to Figure 4.22, and the best alignment obtained in the CASP experiments for the same pair of sequences.

Comparative modeling examples

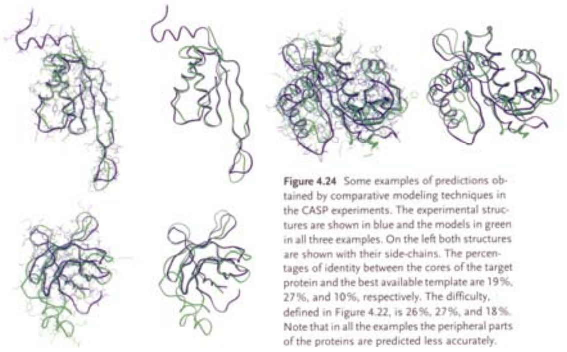


Figure 4.24 Some examples of predictions obtained by comparative modeling techniques in the CASP experiments. The experimental structures are shown in blue and the models in green in all three examples. On the left both structures are shown with their side-chains. The percentages of identity between 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.

Literatur

- Anna Tramontano: *Protein Structure Prediction, Wiley-VCH*, 2006.