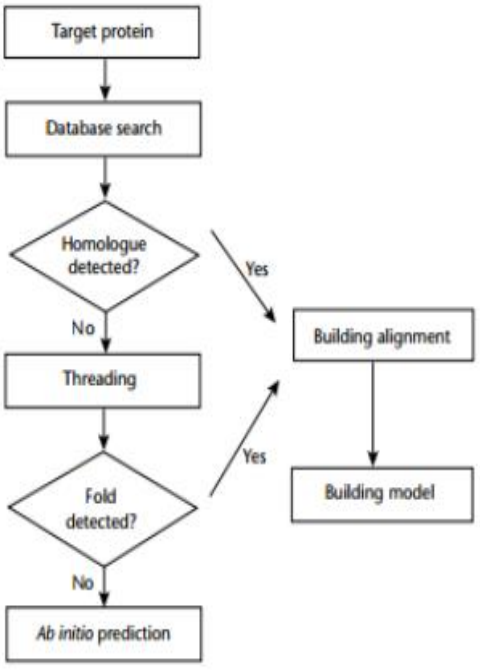
LECTURE 10: PROTEIN TERTIARY STRUCTURES

PROTEIN TERTIARY STRUCTURES: PREDICTION FROM AMINO ACID SEQUENCES

The biological function of a protein is often intimately dependent upon its tertiary structure. X-ray crystallography and nuclear magnetic resonance are the two most mature experimental methods used to provide detailed information about protein structures. However, to date the majority of the proteins still do not have experimentally determined structures available. As at December 2000, there were about 14 000 structures available in the protein data bank (PDB, http://www.pdb.org), and there are about 10 106 000 sequence records sequences in GenBank (http://www.ncbi.nlm.nih.gov/Genbank). Thus theoretical methods are very important tools to help biologists obtain protein structure information. The goal of theoretical research is not only to predict the structures of proteins but also to understand how protein molecules fold into the native structures. The current methods for protein structure prediction can be roughly divided into three major categories: comparative modelling; threading; and ab initio prediction. For a given target protein with unknown structure, the general procedure for predicting its structure is described below



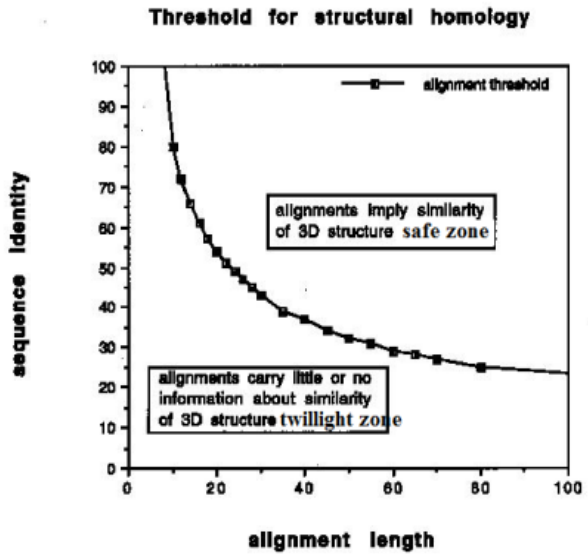
Procedure for predicting a protein structure from its amino acid sequence.

Comparative modelling

It is based on two major observations:

1. The structure of a protein is uniquely determined by its amino acid sequence. Knowing the sequence should, at least in theory, suffice to obtain the structure.

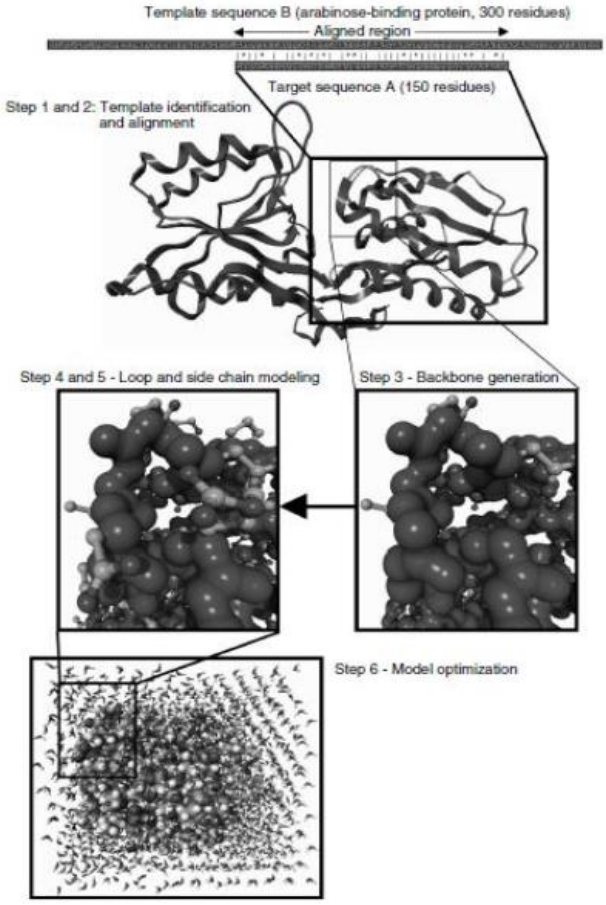
2. During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures. This relationship was first identified by Chothia and Lesk (1986) and later quantified by Sander and Schneider (1991). Thanks to the exponential growth of the Protein Data Bank (PDB), Rost (1999) could recently derive a precise limit for this rule, shown in Figure below. As long as the length of two sequences and the percentage of identical residues fall in the region marked as "safe," the two sequences are practically guaranteed to adopt a similar structure



For a sequence of 100 residues, for example, a sequence identity of 40% is sufficient for structure prediction. When the sequence identity falls in the safe homology modeling zone, we can assume that the 3D-structure of both sequences is the same.

The known structure is called the template, the unknown structure is called the target. Homology modeling of the target structure can be done in 7 steps:

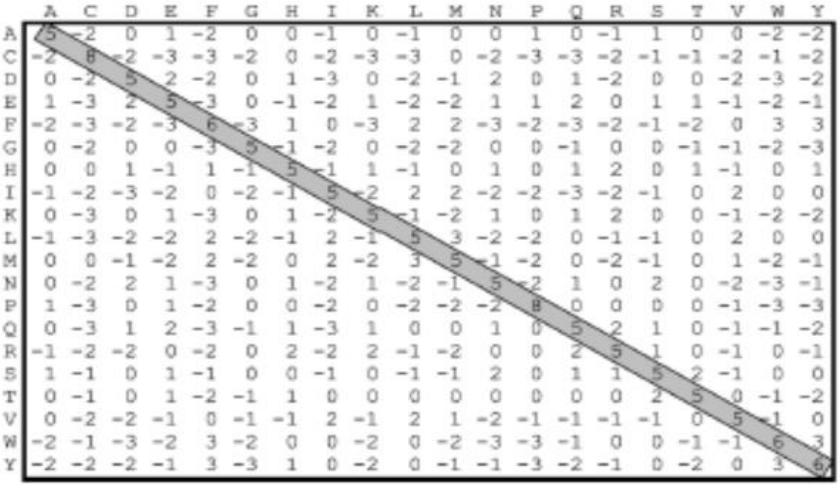
PROTEIN ENGINEERING



1: Template recognition and initial alignment

In the safe homology modeling zone, the percentage identity between the sequence of interest and a possible template is high enough to be detected with simple sequence alignment programs such as BLAST or FASTA. To identify these hits, the program compares the query sequence to all the sequences of known structures in the PDB using mainly two matrices:

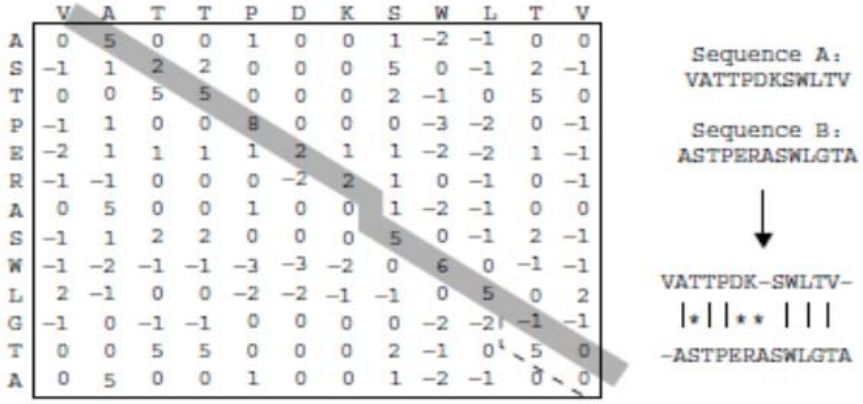
1. A residue exchange matrix (A). The elements of this 20 ∗ 20 matrix define the likelihood that any two of the 20 amino acids ought to be aligned. It is clearly seen that the values along the diagonal (representing conserved residues) are highest, but one can also observe that exchanges between residue types with similar physicochemical properties (for example F → Y) get a better score than exchanges between residue types that widely differ in their properties.



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

A* A typical residue exchange or scoring matrix used by alignment algorithms. Because the score for aligning residues A and B is normally the same as for B and A, this matrix is symmetric.

2. An alignment matrix (B). The axes of this matrix correspond to the two sequences to align, and the matrix elements are simply the values from the



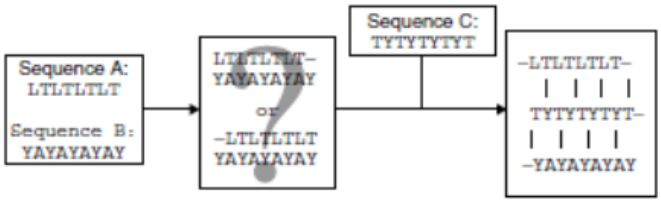
B: The alignment matrix for the sequences VATTPDKSWLTV and ASTPERASWLGTA, using the scores from Figure A. The optimum path corresponding to the alignment on the right side is shown in gray. Residues with similar properties are marked with a star (*). The dashed line marks an alternative alignment that scores more points but requires opening a second gap

residue exchange matrix (Fig. A) for a given pair of residues. During the alignment process, one tries to find the best path through this matrix, starting from a point near the top left, and going down to the bottom right. To make sure that no residue is used twice, one must always take at least one step to the right and one step down. A typical alignment path is shown in Figure B. At first sight, the dashed path in the bottom right corner would have led to a higher score. However, it requires the opening of an additional gap in sequence A (Gly of sequence B is skipped). By comparing thousands of sequences and sequence families, it became clear that the opening of gaps is about as unlikely as at least a couple of nonidentical residues in a row. The jump roughly in the middle of the matrix, however, is justified, because after the jump we earn lots of points (5,6,5), which would have been (1,0,0) without the jump. The alignment algorithm therefore subtracts an "opening penalty" for every new gap and a much smaller "gap extension penalty" for every residue that is skipped in the alignment. The gap extension penalty is smaller simply because one gap of three residues is much more likely than three gaps of one residue each. In practice, one just feeds the query sequence to one of the countless BLAST servers on the web, selects a search of the PDB, and obtains a list of hits—the modeling templates and corresponding alignments.

2: Alignment correction

Having identified one or more possible modeling templates using the fast methods described above, it is time to consider more sophisticated methods to arrive at a better alignment. Sometimes it may be difficult to align two sequences in a region where the percentage sequence identity is very low.

One can then use other sequences from homologous proteins to find a solution. A pathological example is shown in C:



C: A pathological alignment problem. Sequences A and B are impossible to align, unless one considers a third sequence C from a homologous protein.

Suppose you want to align the sequence LTLTLTLT with YAYAYAYAY. There are two equally poor possibilities, and only a third sequence, TYTYTYTYT, that aligns easily to both of them can solve the issue.

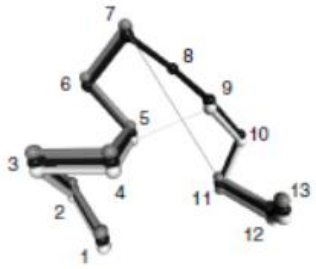
The example above introduced a very powerful concept called "multiple sequence alignment." Many programs are available to align a number of related sequences, for example CLUSTALW, and the resulting alignment contains a lot of additional information.

Think about an Ala → Glu mutation. Relying on the matrix in Figure A, this exchange always gets a score of 1. In the 3D structure of the protein, it is however very unlikely to see such an Ala → Glu exchange in the hydrophobic core, but on the surface this mutation is perfectly normal. The multiple sequence alignment implicitly contains information about this structural context. If at a certain position only exchanges between hydrophobic residues are observed, it is highly likely that this residue is buried. To consider this knowledge during the alignment, one uses the multiple sequence alignment to derive positionspecific scoring matrices, also called profiles. When building a homology model, we are in the fortunate situation of having an almost perfect profile—the known structure of the template. We simply know that a certain alanine sits in the protein core and must therefore not be aligned with a glutamate. Multiple sequence alignments are nevertheless useful in homology modeling, for example, to place deletions (missing residues in the model) or insertions (additional residues in the model) only in areas where the sequences are strongly divergent.

A typical example for correcting an alignment with the help of the template is shown in Figures D and E. Although a simple sequence alignment gives the highest score for the wrong answer (alignment 1 in Fig. D), a simple look at the structure of the template reveals that alignment 2 is correct, because it leads to a small gap, compared to a huge hole associated with alignment 1.

		1	2	3	4	5	6	7	8	9	10	11	12	13
Template		**PHE**	**ASP**	**ILE**	**CYS**	**ARG**	**LEU**	**PRO**	**GLY**	**SER**	**ALA**	**GLU**	**ALA**	**VAL**
Model (bad)	1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model (good)	2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE

D: Example of a sequence alignment where a three-residue deletion must be modeled. While the first alignment appears better when considering just the sequences (a matching proline at position 7), a look at the structure of the template leads to a different conclusion (Figure E)



E: Correcting an alignment based on the structure of the modeling template (Cα-trace shown in black). While the alignment with the highest score (dark gray, also in Figure D) leads to a gap of 7.5 A between residues 7 and 11, the second option (white) creates only a tiny hole of ˚ 1.3 A between residues 5 and 9. This can easily be accommodated by small backbone shifts. (The ˚ normal Cα−Cα distance of 3.8 A has been subtracted).

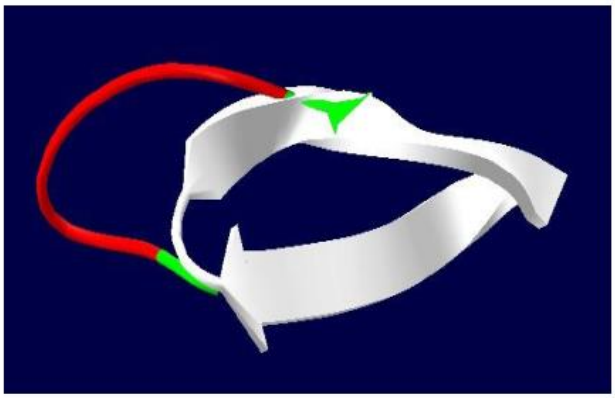
3: Backbone generation

When the alignment is correct, the backbone of the target can be created. The coordinates of the template-backbone are copied to the target. When the residues are identical, the side-chain coordinates are also copied. Because a PDB-file can always contain some errors, it can be useful to make use of multiple templates.

4: Loop modeling

Often the alignment will contain gaps as a result of deletions and insertions. When the target sequence contains a gap, one can simply delete the corresponding residues in the template. This creates a hole in the model, this has already been discussed in step 2. When there is an insertion in the target, shown in Figure B, the template will contain a gap and there are no backbone coordinates known for these residues in the model. The backbone from the template has to be cut to insert these residues. Such large changes cannot be modeled in secondary structure elements and therefore have to be placed in loops and strands. Surface loops are, however, flexible and difficult to predict. One way to handle loops is to take some residues before and after the insertion as "anchor" residues and search the PDB for loops with the same anchor-residues. The best loop is simply copied in the model. This is shown in Figure G. The two residues which are colored green in Figure G are used as anchor, the best loop with the inserted resisdues was found in the database and placed in the model.

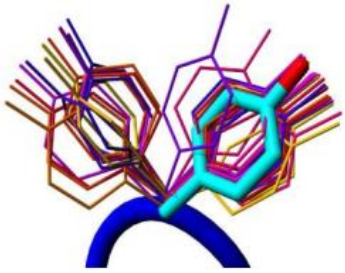
F: Target sequence (green) with insertion (grey box) results in a gap in the template.



F: The red loop is modeled with the green residues as anchor residues. The insertion of 2 residues results in a longer loop.

5: Side-chain modelling

Now it is time to add side-chains to the backbone of the model. Conserved residues were already copied completely. The torsion angle between C-alpha and C-beta of the other residues can also be copied to the model because these rotamers tend to be conserved in similar proteins. It is also possible to predict the rotamer because many backbone configurations strongly prefer a specific rotamer. As shown in Figure G, the backbone of this tyrosine strongly prefers two rotamers and the real side-chain fits in one of them. There are libraries based upon the backbone of the residues flanking the residue of interest. By using these libraries the best rotamer can be predicted. This last method is used by Yasara.



G*: Prefered rotamers of this tyrosin (colored sticks) the real side-chain (cyan) fits in one of them.*

PROTEIN ENGINEERING

6: Model optimization

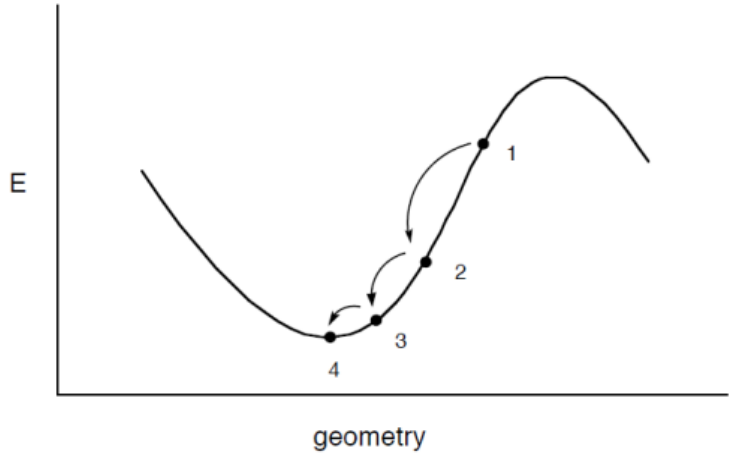
The model has to be optimized because Many structural artifacts can be introduced while the model protein is being built

- Substitution of large side chains for small ones
- Strained peptide bonds between segments taken from difference reference proteins
- Non optimum conformation of loops

Energy Minimisation is used to produce a chemically and conformationally reasonable model protein structure

Two mainly used optimisation algorithms are

- Steepest Descent
- Conjugate Gradients



The process of energy minimization changes the geometry of the molecule in a step-wise fashion until a minimum is reached.

Molecular Dynamics is used to explore the conformational space a molecule could visit, Molecular dynamics (MD) is a computer simulation method for studying the physical movements of atoms and molecules

PROTEIN ENGINEERING

7: Model validation

The models we obtain may contain errors. These errors mainly depend upon two values.

1. The percentage identity between the template and the target.

If the value is > 90% then accuracy can be compared to crystallography, except for a few individual side chains. If its value ranges between 50-90 % r.m.s.d. error can be as large as 1.5 Å, with considerably more errors. If the value is <25% the alignment turns out to be difficult for homology modeling, often leading to quite larger errors.

2. The number of errors in the template.

Errors in a model become less of a problem if they can be localized. Therefore, an essential step in the homology modeling process is the verification of the model. The errors can be estimated by calculating the model's energy based on a force field. This method checks to see if the bond lengths and angles are in a normal range. However, this method cannot judge if the model is correctly folded. The 3D distribution functions can also easily identify misfolded proteins and are good indicators of local model building problems.

Modeller

Modeller is a program for comparative protein structure modelling by satisfaction of spatial restraints. It can be described as "Modeling by satisfaction of restraints" uses a set of restraints derived from an alignment and the model is obtained by minimization of these restraints. These restraints can be from related protein structures or NMR experiments. User gives an alignment of sequences to be modelled with known structures. Modeller calculates a model with all non hydrogen atoms. It also performs comparison of protein structures or sequences, clustering of proteins, searching of sequence databases.