Is the similarity significant, or could it be due to chance?

Even if two proteins are unrelated, we would expect some similarity simply by chance.

Is the alignment score significantly higher than random?

Align random permutations of the sequences, and find the mean and standard deviation of the resulting distribution.

The z-score reflects the significance of a global similarity score.

$$z\text{-}score = \frac{score - mean}{standard\ deviation}$$

Larger values imply greater significance.

Possible substitution matrices for DNA

	A	C	G	T
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

	A	C	G	T
A	2	-2	-1	-2
C	-2	2	-2	-1
G	-1	-2	2	-2
T	-2	-1	-2	2

More realistic similarity measures

Not all substitutions are equally likely.

- A transition between two purines (A, G) or between two pyrimidines (C, T/U) is more common than a purine-pyrimidine transversion.
- Replacement of one amino acid residue by another with similar size or physiochemical properties is more common than replacement by a dissimilar amino acid residue.

Insertion/deletion of N contiguous amino acid residues or nucleotides is more likely than N independent insertion/deletion events.

Thus, we should have different penalties for opening gap and for extending a gap.

Substitution matrices

PAM and BLOSUM are two widely used families of substitution matrices.

Derived from observations of actual mutations in alignments of related proteins.

$$score_{ab} = \log \frac{observed\ ab\ mutation\ rate}{mutation\ rate\ expected\ from\ amino\ acid\ frequencies}$$

PAM and BLOSUM differ in how the alignments, from which observations are made, are selected.

Substitution matrices near one end of the series are most suitable when comparing proteins separated by a short evolutionary distance; those near the other end are most suitable for comparing distantly related proteins.

PAM250

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

BLOSUM62

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

Structural clues from multiple sequence alignments

— Residues at highly conserved positions often have important functional or structural roles.

— Insertions and deletions can be accommodated most easily in surface loops.

— Conserved patterns of hydrophobic residues can suggest secondary structure.

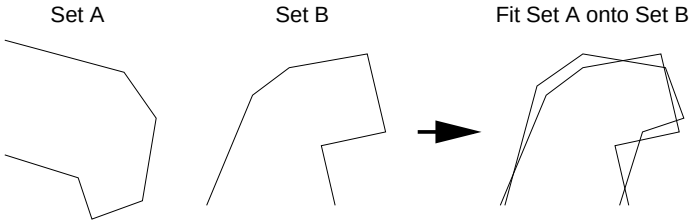
— The root mean square deviation between pairs of homologous proteins generally increases as the percent residue identity decreases.

Heuristics for manual secondary structure prediction

- Many α-helices are amphipathic. Conserved hydrophobic residues at positions i, i+3, i+4, i+7, etc. are highly indicative of an α-helix.
- Half-buried strands will tend to have hydrophobic and hydrophilic residues at alternate positions.
- In proteins containing both α-helices and strands the strands are often completely buried and tend to contain only hydrophobic residues.

For more details and references, see:
http://www.bmm.icnet.uk/people/rob/CCP11BBS/secstrucpred.html

Superposing structures



- 3-D transformation to map Set A onto Set B
- Root Mean Square (RMS) distance

"HSSP-curve"

— Shows the length-dependent threshold for significant sequence identity.

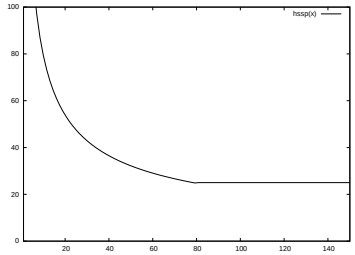
— Proposed by Sander and Schneider (1991) and revised by Rost (1999).

— Above the curve, identifing true positives is easy.

— Just below the curve, the number of false positives rises rapidly; distinguishing between true and false positives in the "twilight zone" is difficult.

(HSSP stands for "Homology-derived Secondary Structure of Proteins")

HSSP-curve



[Sander C. and Schneider, R., Proteins: Structure, Function and Genetics, 1991, 9:55-68]

Why build model structures?

Knowledge of a protein's three-dimensional structure is vital to a full understanding of the molecular basis for its biological function.

We want to understand the function of all proteins encoded by a genome, therefore we would like to know all of their 3-D structures.

Experimental techniques for determining protein structure are relatively slow and expensive, so we look to modelling as a way of extending the set of 3-D structures.

Modelling can also be used in protein engineering when designing proteins for therapeutic applications.

Comparative modelling and fold recognition

Comparative modelling (homology modelling):

Given:

- sequence of target protein with unknown structure
- known structure of a related protein

Predict:

- three-dimensional structure of target protein

Fold recognition:

Given:

- sequence of target protein with unknown structure
- library of known folds

Predict:

- known fold that is most compatible with the target protein's sequence

Fold recognition

The idea behind "threading":

Imagine a wire wound into the shape of a known protein's main chain "fold".

Imagine next that our new sequence is represented by beads that are "threaded", in order, onto the wire, and are pushed along the wire.

At each step, a score is calculated based on which residues are adjacent in space, which residues are buried, etc.

Repeat this process for each different known fold.

A high score indicates that the sequence is compatible with that fold.

Comparative modelling strategy

- identify a known structure that is predicted to be similar;
- align sequences;
- predict structurally conserved regions, and locations of insertions and deletions (sometimes called "indels");
- build model backbone structure
  - copy predicted conserved main chain regions from template structure,
  - remodel loops with insertions or deletions;
- add side chains to the modelled main chain;
- evaluate and refine model.

Further studies in bioinformatics

Sequence

MVE360 - Bioinformatics

Structure

TDA507 - Computational methods in bioinformatics

Systems

KMG060 - Systems biology