Sequence, could be your novel sequence

Assign function to the sequence?

How similar? Can sequence similarity be related to functional (3D) similarity?

How can I compare sequences?

What kind of alignments do I need?
(global: End-to-end; or local (identifying islands))

Local Alignments (statistically sound theory) not Global alignment

What do I need to compare sequences?
Scoring Matrix

Does the scoring matrix (20 x 20 or 4 x 4) have to follow certain rules?

Yes, what are they?

	A	T	G	C
A	+2	-3	-3	-3
T	-3	+2	-3	-3
G	-3	-3	+2	-3
C	-3	-3	-3	+2

- At least one of the scores have to be a positive number

- **Expected score** for aligning <u>random pair of amino acids</u> has to be NEGATIVE

$$\sum_{i,j} p_i p_j s_{i,j} < 0$$

- You give any matrix that satisfies this, it will be a scoring matrix

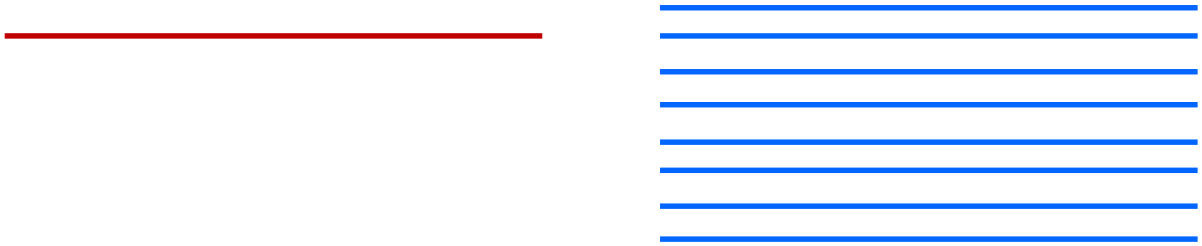
$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

p_{ab} : Target frequencies: "The probability that we expect to observe residues *a* and *b* aligned in homologous sequence alignments" (a quote from Eddy's BLOSUM matrix paper)

The denominator (fa fb) is the likelihood of a null hypothesis: that the 2 residues are uncorrelated and unrelated occurring independently (a quote from Eddy's BLOSUM matrix paper)

Now we know the rules of comparing two sequences. How about comparing a list of sequences ?



Database of sequences (could be nr or RefSeq)

There is another aspect of alignment that we haven't talked about?

How are we going to do the alignments using the scoring scheme that we have developed?
Dynamic Programming

Will DP be used in real-life alignments?
Approximations → BLAST