CSE 527, Lecture 5, 10/10/2007 using the slides for Lectures 4-5. Ryan Stephens

Slide	Notes
9	There are two reasons for the varying, and sometimes negative, values for mismatch in the BLOSUM62 matrix: - some substitutions are rare, so the scores are negative (or just low) - other substitutions (with higher scores for mismatches) have little effect on the structure of the protein, and are almost interchangeable, so these mismatches are not penalized much The table is built empirically; the scores make sense intuitively if you have a feel for the chemical properties of each amino acid
11	You know a “good” alignment score by comparing it to a null model.
12	In short, this technique calculates the probability for the alternative and null models, then computes the ratio between them (alt/null). This gives you a likelihood ratio. A ratio of 1 tells you nothing, since it means that both hypotheses are equally likely.
13	Using the log of the likelihood ratio is often more convenient, e.g. $\log(1) = 0$. Neyman-Pearson says that LRT is sufficient in most cases.
14	The “p-value” is the probability that the data you observed could have happened by chance. You can (usually) publish with a p-value of <0.05. Philosophically, you don’t *accept* the alternative model, you *reject* the null model. Usually the null model is an existing, competing theory and reject it in favor of your model proves only that it (the null model) is less likely than the alternative model, not that the alternative model is right or certain.
15	“Homologous” implies shared ancestry, though the term is frequently overloaded. A simple way to do an LRT for sequence alignment is to calculate the sum of the log of the ratio of the individual probabilities that a residue aligns in homologous organisms to the probability that two residues align at random.
16	The BLOCKS DB shows different blocks (highly conserved parts) of proteins that align in different organisms.

17	You can replace the BLOSUM62 matrix with a random matrix and, surprisingly, it can still be interpreted as giving likelihood ratio scores. So if, e.g., you think A -> G should score +99 for some ad hoc reason, that’s equivalent to saying G is x times more likely than A in your target compared to background.
19	Should the score be the only measure of a good alignment? - use additional, orthogonal criteria - take top n matches and use other discerning criteria to find the best among that set instead of just the one with the highest score
20	Extreme value distribution works well practically.
21	Another way to measure the score is to generate random sequences and compare a given sequence to them - generating realistic random sequences is tricky, though - if you have an uncommon sequence then using a naïve randomly generated sequence is not a good idea because it will not look like your target sequence (and therefore always give low scores)
22	This slide has pseudocode for generating random sequences.