> One of of the requirements for scoring for local pairwise alignments is that "The expected alignment score of a pair of randomly generated sequences (i.e., sequences sampled from a background distribution) must be negative."
>
> What is the expected alignment score and why is it important that it be negative?

The goal of local alignment is to find similar regions in a pair of sequences, $s_1, s_2 \in \Sigma^*$. The optimal local alignment obtained from the dynamic programming algorithm will depend on the function used to score matches and mismatches. We seek a scoring function that will yield local alignments that correspond to biologically meaningful features. However, even unrelated sequences, when aligned, will match at a few positions. In order to have a strong presumption that a high-scoring local alignment is biologically meaningful, the match and mismatch scores should be chosen in such a way that chance matches contribute little to the alignment score.

If a pair of symbols, sampled at random, were appended to the end of an existing alignment, how much would the alignment score increase? The increase, on average, is given by the expected alignment score. Let x be a symbol in Σ and let p_x be the background frequency of x in the genomes from which sequences s_1 and s_2 were sampled. Then the expected alignment score is

$$\overline{\mathcal{S}} = \sum_{x\in\Sigma} p_x^2 \cdot M + \sum_{x\in\Sigma} \sum_{y\in\Sigma, y\neq x} p_x p_y \cdot m,$$

where M and m are the match and mismatch scores, respectively.

For example, suppose that s_1 and s_2 are DNA sequences with uniform nucleotide frequencies; i.e., $p_A = p_G = p_C = p_T = 0.25$. Since the nucleotide frequencies are uniform, the nucleotide pairs also have uniform frequencies: $p_x p_y = 0.25^2, \forall x, y$. There are 16 possible pairs of nucleotides; four pairs consist of the same nucleotide (AA, GG, CC, and TT) and 12 pairs are made up of different nucleotides. In this case, the expected alignment score is

$$\begin{aligned}\bar{\mathcal{S}} &= 4 \cdot 0.25^2 \cdot M + 12 \cdot 0.25^2 \cdot m \\ &= 0.25 \cdot M + 0.75 \cdot m.\end{aligned}$$

Why is it important that the expected alignment score be negative? If the expected score were positive, then extending a local alignment with unrelated pairs of symbols would increase its score. Such an alignment would have a higher score because it is longer, not because it contains stronger evidence of a shared biological relationship.

In the nucleotide example given above, M and m should be selected so that

$$0.25 \cdot M + 0.75 \cdot m < 0.$$

In this simple example, we assumed that $p_x = 0.25, x \in \{A, G, C, T\}$. In general, obtaining an estimate of $p_x \in \Sigma$ is not always straightforward. In Problem Set 1, you are asked to align English words. For this purpose, you might use letter frequencies in the English language. The Wikipedia page on letter frequency discusses various ways these frequencies are estimated and gives a table with two sets of frequencies, estimated using different methods. Either one would serve for homework problems.

We will explore these issues in greater depth later in the semester when we discuss amino acid substitution matrices and database searching.