6.3 Database similarity search -BLAST and FASTA

BLAST is the acronym for Basic Local Alignment Search Tool. It uses the method of Altschul *et al.* (JMB 215:403-410, 1990) to pick out sequences already collected in a database that are similar to the query sequence. BLAST takes the query sequence input by the user and compares it with each entry in the database, looking for segments of high degrees of similarities. It picks out from the database those sequences that contain a segment so similar to part or all of the query that such similarity is deemed statistically significant (i.e., unlikely to occur by chance).

Exercise BLAST is available at the NCBI web site. Before you go on, it may be helpful to visit http://www.ncbi.nlm.nih.gov/blast/ to take a look at the BLAST overview and go through the exercise in the BLAST tutorial there.

The algorithm used in the current version of BLAST at NCBI can be summarized in three main steps:

Step 1. Finding high-scoring segment pairs: For each sequence in the database, BLAST will compare it with the query. BLAST first seeks from the sequence pair, equal length sequence segments, which have maximal aggregate similarity score that cannot be increased by extension or trimming. Such locally optimal alignments are called "high-scoring segment pairs" or HSP's. The current version of BLAST requires that each HSP must contain at least two non-overlapping pairs of words of length W (these word pairs are called "hits" in BLAST jargon, default values for W are 3 for amino acid, and 11 for nucleotide sequences) satisfying certain requirements:

a) Their similarity score exceeds a threshold value T.
b) The offset of the two word pairs are equal. If a word pair occurs at position x_1 of the first sequence and position x_2 of the second sequence, the offset of the word pair is defined to be $x_1 - x_2$
c) The distance between the word pairs is no more than a preset upper limit A. The distance between two word pairs (x_1, x_2) and (x'_1, x'_2) is defined to be the difference between their first coordinates $x_1 - x'_1$.

The rationale behind these criteria for finding HSP's is based on the observation that an HSP with a large enough similarity score to eventually generate a statistically significant local alignment is very likely to contain multiple hits with the same offset and within a relatively short distance of one another. The chances of missing any HSP's of interest using this procedure is relatively small.

Step 2. Gapped extensions of HSP's: BLAST will only retain those HSP's that exceed a moderate score S_g, and further attempt to extend the alignment in both the leftward and rightward directions while allowing gaps to be introduced. S_g is controlled so that no more than about one gapped extension is invoked per 50 database sequences. A dynamic type algorithm, with modifications to improve efficiency, is used. Whenever a gap is opened or extended, a penalty will be imposed according to a gap penalty function of the form $w(k) = a+bk$ with k being the length of the gap. The alignment(s) with the maximal score will be assessed for statistical significance.

Step 3. Assess statistical significance of the maximal alignment score: If the fully extended gapped alignment is deemed significant, the database sequence will be picked and described in the output. The evaluation of statistical significance is based on comparison with the rolling-die random sequence models described before (in Chapter 3). For example, the random amino acid sequence model will be generated by rolling an icosahedral (20 faced) die for a number of times equal to the length of the sequences under comparison. The die is loaded according to the relative frequencies of occurrence of the amino acids in the database.

The maximal alignment score M for two random sequences is a random variable. Asymptotically, it follows an extreme value distribution when the lengths m and n of the sequences $\rightarrow \infty$. In reality, when m and n large, the asymptotic distribution yields a good approximation that can be used to calculate the probability of the maximal local alignment score to exceed any given level. We shall discuss this more fully in the next section.

From the probability distribution of the maximal alignment scores, one can determine the probability of getting an alignment as good as the one observed. If this probability is small (say < 0.05), the alignment is deemed statistically significant. In the BLAST output, this probability p is converted to a *bit score* equal to $-\log_2 p$. The smaller the probability, the larger the bit score.

One can also calculate the expected number E of times an alignment with such a score would occur in a database of the same size as the one searched. If this expected number is high, it means that the alignment can occur quite frequently by chance. On the other hand, a low value of E indicates that alignment is expected to occur very rarely and hence is worth further examination. BLAST lets you specify a parameter which discards those alignments expected to occur more than certain number (default is 10) of times.