CS1820: Algorithmic Foundations of Computational Biology
Prof. Sorin Istrail · April 2014

BLAST

April 2014

BLAST is a useful tool for determining which DNA or protein sequences in a large database have significant similarity to a given query sequence.

1 Comparison of Two Aligned Sequences

Consider an ungapped global alignment of two DNA sequences of length N

```
GGAGACTGTAGACAGCTAATGCTATA
| |  |  ||| ||        |||
GAACGCCCTAGCCACGAGCCCTTATC
```

BLAST is a program that searches for high-scoring local alignments between two sequences, and then tests for significance using P-values. The P-value depends on the lengths of the two sequences. Clearly we need to take this into account, since the longer the sequences get, the more likely it is to have local homology purely by chance. We usually have one “query” sequence, and a database of sequences that we wish to search against. When calculating P-values, we take the size of the database into account.

BLAST is very fast. There are two approaches used to speed up the alignment calculations.

1) Algorithmic heuristics which avoid having to do a full search through all ungapped local alignments. There are exponentially many such local alignments. Because this is just a heuristic, this means that it’s possible for BLAST to miss the absolute highest score alignment.

2) The calculation of the P-value uses sophisticated approximations to achieve very fast calculations.

2 Similarity Matrices

Similarity matrices are a central part of BLAST theory. There is a close relationship between the alternative hypothesis probabilities f(j,k) and the way the entries of a substution matrix are constructed using log likelihood.

In a similarity matrix, identities and conservative replacements have positive scores, while unlikely replacements have negative scores. For proteins we usually use the PAM120 similarity matrix. For DNA, we usually use +5 for matches and -4 for mismatches.

When aligning two sequences of the same length, the alignment score is the sum of the pairwise similarity scores.

3 Maximum Segment Pairs

What we are most interested in is the maximum segment pair or MSP. The MSP is the highest-scoring pair of identical length segments from the two sequences being aligned. The boundaries of the MSP are chosen in order to maximize its score. This means that the score of the MSP cannot be improved by either lengthening the segments or by shortening them. The score of the MSP is therefore a measure of the local similarity between the two segments.

Note that because we have both positive and negative numbers in our similarity matrix, it's possible that lengthening (or shortening) a segment could either raise *or* lower it's score.

BLAST looks for all the MSPs whose score is above a certain threshold. Statistical theory allows us to calculate the highest MSP score that is likely to appear by chance. BLAST minimizes the search time by ignoring areas where the MSP score of the query is very unlikely to exceed the threshold score.

The BLAST algorithm is based on quickly finding a set of "hits", which it then extends to see if it can exceed the threshold MSP score. A hit is a fixed-size word (usually 3 or 4 letters) that occurs in the query string, and which has a score of at least T. The lower the value of T, the greater the chance of any random string being counted as a hit. Therefore, if T is set too low, then BLAST will have to search through a large number of hits, which will take a long time. On the other hand, if T is set too high, then legitimate MSPs may end up being ignored. Thus there is a trade-off between time and accuracy when determining how to set T.

4 Algorithm sketch

1. Break down the query string into all possible **words** of length k. Now consider each word, w, individually.

2. Create a list of **search terms**, comprised of w and all other words that, when compared against w, have a score of at least T.

3. Identify all the places in the database where one of search terms appears, and extend in each direction to find the maximum segment pair (MSP).

4. If the MSP has a score of at least S, report it as a match.