27-05-2020

Francesco Penasa 215038

Algorithms for Bioinformatics – Theoretical exam

1) Discuss the main differences between PAM and BLOSUM matrices.

The first difference that we can observe is that PAM matrices are based on an explicit evolutionary model and BLOSUM matrices are based on an implicit model of evolution and the data used to obtain such matrices, in BLOSUM is derived from conserved Blocks in protein while in PAM the data comes from multiple sequence alignments of related sequences.

In PAM, the matrix is derived from a simplified phylogenetic tree structure while in BLOSUM some clustered sequences are used (blocks) and each cluster is weighted as a single sequence in counting pairs.

In PAM we use transition and log odds scoring matrices and in BLOSUM only the log odds scoring matrix.

The evolutionary distance in PAM is calculated from Markov models while in BLOSUM such distance is computed from clustering of sequences.

In PAM when the parameter increases the evolutionary distance increase too, for example PAM150 has higher evolutionary distance than PAM10. In BLOSUM the exact opposite happens, for example we can observe that BLOSUM90 has a smaller evolutionary distance than BLOSUM62. Since both PAM and BLOSUM are different methods for showing the same scoring information, the two can be compared but due to the very different method of obtaining this score, a PAM100 does not equal a BLOSUM100.

2) What is the problem of multiple sequence alignment? Define the problem and illustrate one of the systems that have been proposed as a solution

The problem of multiple sequence alignment is the high computational complexity in fact it is a NP-complete combinatorial optimization problem, with space complexity O(n^k) and time complexity O($2^k n^k$) for *k* sequences with length *n*.

One of the systems that have been proposed as a solution to this problem is the progressive alignment method. The progressive alignment method is a fast heuristic multiple alignment method that is motivated by the tree alignment approach. The most popular multiple alignment programs follow this strategy.

The basic idea of progressive alignment is that the multiple alignment is computed in a progressive way. In its simplest version, the given sequences s1,s2,...,sk are added one after the other to the growing multiple alignment, first an alignment of s1 and s2 is

computed, then s3 is added, then s4, and so on. Often, the most similar sequences are aligned first in order to start with a well-supported, error-free alignment. The progressive method is used in CLUSTAL W.