Position-Specific Iterated (PSI) BLAST

Paul C. Spiegel

March 7, 2000

1 Website

http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast

2 Introduction

BLAST is an algorithm that has been used extensively to search protein and DNA databases for similarities to specific sequence queries. BLAST employs a substitution matrix that specifies a score $S(i, j)$ for aligning each pair of amino acids i and j. In this original program, BLAST seeks equal length segments within the two sequences that, when aligned to one another without gaps, have maximal aggregate score. Since the development of the BLAST algorithm, there have been attempts to modify it to make it more effective. This has been done by adding a two-hit filtering method, a gapped alignment method, and finally a position-specific iteration method (which will be explained further).

By using iterative runs of the original algorithm, BLAST, a position-specific score matrix can be generated from significant alignments found in round i for use in round $i+1$. The position-specific matrix can be thought of as a consensus sequence used to detect more distant relationships not able to be found by BLAST. These iterations are done until there is no longer a significant difference in the scoring matrix between i and $i-1$.

3 Score Matrix Architecture

The alignments of simple sequences with a position-specific score matrix is quite similar to the alignment of two simple sequences. The difference between the two is that the score for aligning a letter with a pattern position is given by the matrix itself, instead of a reference to a substitution matrix. For a BLAST search of proteins, a query of length L and a substitution matrix of dimension 20×20 is thus replaced by a matrix of dimension $L \times 20$ for each iteration. This improved method for the estimation of amino acid residue possibilities at individual positions creates a more sensitive scoring algorithm.

Another important point to consider that may become a limitation to the PSI-BLAST protocol is that each matrix constructed is precisely equal to that of the original query sequence. With this scoring matrix, the protocol then seeks for local alignments.

4 Multiple Alignment Construction

To construct the multiple sequence alignment from the BLAST output, the algorithm simply collects all database sequence segments that have been aligned to the query with E-values below a threshold. The multiple sequence alignment is also made non-redundant by purging any rows that have greater than 98% similarity to any other sequence hit. Gaps are also ignored so that the position-specific score matrix can stay the exact same length as the query.

5 Sequence Weights

The construction of the score matrix must also have sequence weights for varying residues within each column. In a large set of closely related sequences, a redundant residue could possibly outvote a small number of divergent sequences. The sequence weighting method of Henikoff and Henikoff is used in this algorithm with slight modifications due to its speed and simplicity.

6 Advantages of PSI-BLAST

1. Weaker pairwise alignments can be improved and extended before they are incorporated into the evolving multiple sequence alignment.

2. Unrelated sequences with high scores (false-positives) can have their scores decreased with the improved score matrix. This could, in turn, exclude them from the alignment.

3. Related sequences that received relatively high alignment scores but missed the cutoff could be included with an improved score matrix.

7 Reference

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”, Nucleic Acids Res. 25:3389-3402.