Algorithms for bioinformatics

Giacomo Fantoni
telegram: @GiacomoFantoni

Github: https://github.com/giacThePhantom/algorithms-for-bioinformatics

March 29, 2022

Contents

Chapter 1

Needleman Wunsch

1.1 Introduction

Direct comparison of two sequences based on the presence in both of the corresponding amino acids in an identical array is insufficient to establish the full genetic relationship between two proteins. Allowance for gaps multiplies the number of comparisons that can be made but introduces unnecessary and partial comparisons.

1.2 A general method for sequence comparison

The maximum match can be defined as the largest number of amino acids of one protein that can be matched with those of another protein while allowing for all possible deletions. It can be determined by representing in a matrix all possible pair combinations that can be constructed from the amino acid sequences of the protein being compared. So A_j is the jth amino acids of protein A and B_i is the ith amino acids of protein B. A_j are the columns and B_i all the rows of the matrix MAT. Then A_{ij} represent a pair combination with amino acids A_j and B_i. Every possible comparison can be represented by pathway through the matrix. A pathway is signified by a line connecting cells of the array. Complete diagonals contain no gaps. A necessary pathway begins at a cell in the first column of row. Either i or j must increase by only one, while the other may increase by one or more, leading to the next cell in a pathway. This is repeated until i, j or both reach their limiting value. Every partial or unnecessary pathway will be contained in at least one necessary pathway. The values in the matrix are computed as:

$$MAT_{ij} = \max(MAT_{i-1,j-1} + \alpha\delta(A_j, B_i), MAT_{i-i,j} + d, MAT_{i,j-1} + d)$$

Where d is the penalty factor, a number subtracted for every gap made, may be defined as a barrier for allowing the gap. And α can be a function that can represent any theory with the significance of a pair of amino acids. No gap would be allowed in the operation unless the benefit from allowing that gap would exceed the barrier. This method can be expanded for allowing the comparison of n sequences through and n-dimensional matrix. The maximum-match pathway can be obtained by beginning at the terminals of the sequences and proceeding towards the origin, first by adding to the value of each cell possessing indices $i = y - 1$ and or $j = z - 1$. The process is iterated until all cells in the matrix have been operated upon. Each cell in the outer row or column will contain the maximum number of matches that can be obtained by originating any pathway at that cell and the largest number in that row or column is equal to the maximum match. The cells of the array which contributed to the maximum match may be determined by recording the origin of the number that was added to each cell when the array was operated upon.

1.3 Evaluating the significance of the maximum match

To accomplish the estimate of if a result found differs significantly from a match between random sequences two sets of random sequences can be constructed, each one from the set of amino acid composition of each of the proteins. If the value found for the real proteins is significantly different the difference a function of of the sequences alone and not of the composition.

1.4 Cell values and weighting factors

Cells can be weighted in accordance with the maximum number of corresponding bases in codons of the represented amino acids, to make the comparison more accurate. Also the significance of the maximum match is enhanced by decreasing the weight of those pathways containing a large number of gaps through the penalty factor.

Chapter 2

Smith Watermann

2.1 Introduction

The Smith Watermann algorithm extends the one of Needleman and Wunsch to find a pair of segment, one from each of two long sequences, such that there is no other pair of segments with greater similarity. This similarity measure allows for deletion and insertion of arbitrary length.

2.2 Algorithm

Consider two molecular sequences $A = a_1 a_2 \dots a_n$ and $B = b_1 b_2 \dots b_m$. Given a similarity $s(a, b)$ of elements of the sequence and W_k the weight of deletions of length k, to find pairs of segments with high degrees of similarity, a matrix H is set up such that:

$$H_{k0} = H_{0l} = 0 \qquad \forall 0 \leq k \leq n \land 0 \leq l \leq m$$

H_{ij} si the maximum similarity of two segments ending in a_i and b_j. H_{ij} is computed such that:

$$H_{ij} = \max(H_{i-1,j-1} + s(a_i, b_j), \max_{k \geq 1}(H_{i-k,j} - W_k), \max_{l \geq 1}(H_{i,j-l} - W_k), 0)$$

With $1 \leq i \leq n$ and $1 \leq j \leq m$. So H_{ij} is:

- $H_{i-1,j-1} + s(a_i, b_j)$ If a_i and b_j are associated.
- $H_{i-k,j} - w_k$ if a_i is at the end of a deletion of length k.
- $H_{i-k,j} - W_l$ if b_j is at the end of a deletion of length l.
- 0 is used to prevent calculated negative similarity, indicating no similarity up to a_i and b_j.

The pair of segments with maximum similarity is found first by locating the maximum elemnt of H. The other elements are determined sequentially with a traceback procedure ending with an element of H equal to 0. This procedure other than identifying the elements produces their alignment. The parameters where:

$$s(a_i, b_j) = \begin{cases} 1 & a_i = b_j \\ 0 & a_i \neq b_j \end{cases}$$

And

$$W_k = \frac{1}{3}k$$

This algorithm in particular allows for the alignment of sequences that contained both mismatches and internal deletions.

Chapter 3

PAM

Chapter 4

BLOSUM

Chapter 5

FASTA

Chapter 6

BLAST

Chapter 7

How to BLAST