Algorithms for Bioinformatics - summary

Francesco Penasa

March 9, 2020

https://www.youtube.com/watch?v=PdyARRNwi7I

1 Global sequence alignment

https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

1.1 Exercise

1. We have got two sequences s_1 and s_2 and the following weights $match = 1$ $mismatch = -1$ $gap = -2$;
2. Put the sequences in a matrix like in table 1;
3. Init the first row and the first column like in table 2
4. Starting from the 0 on the top-left (with $i = 1$ and $j = 1$) find the max between the following equations:
   (a) $if\ M[i+1][label] == M[label][j+1]\ then\ M[i][j] + match$
   (b) $if\ M[i+1][label] \neq M[label][j+1]\ then\ M[i][j] + mismatch$
   (c) $M[i][j] + gap$
5. Repeat row to row
6. Since it is global the best alignment will result in the max score at the bottom right of the matrix.

Table 1: Init table $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-								
G								
A								
T								
T								
A								
C								
A								

Table 2: Init first row and first column $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2							
A	-4							
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

Table 3: First row iteration $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2	$\nwarrow 1$	$\leftarrow -1$	$\leftarrow -3$	$\leftarrow -5$	$\leftarrow\nwarrow -7$	$\leftarrow -9$	$\leftarrow -11$
A	-4							
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

Table 4: Second row iteration $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2	$\nwarrow 1$	$\leftarrow -1$	$\leftarrow -3$	$\leftarrow -5$	$\leftarrow\nwarrow -7$	$\leftarrow -9$	$\leftarrow -11$
A	-4	$\nwarrow\uparrow -1$	$\nwarrow 0$	$\nwarrow 0$	$\leftarrow -2$	$\leftarrow -4$	$\leftarrow -6$	$\leftarrow -8$
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

2 Local sequence alignment

`https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm`
Indian explains things: `https://www.youtube.com/watch?v=QphFHG9tmOY`
As Needleman Wunsch for global alignment we use a table. There are only three differences:

1. The table initialization is done with all 0 as in table 5
2. We have gap penality to incentivize not starting the alignment. To implement this gap penality we will use an additional option as we can see in the following equation.
3. We search for the max number in the table, we don't look only at the last cell.

$$H = matrix$$

$$H_{i,0} = \emptyset \quad 0 \leq i \leq n$$

$$H_{0,j} = \emptyset \quad 0 \leq j \leq m$$

$$H(i,j) = max \begin{Bmatrix} 0 & \\ H_{i-1,j-1} + w(a_i, b_i) & Match/Mismatch \\ H_{i-1,j} + w(a_i, -) & Insertion \\ H_{i,j-1} + w(-, b_i) & Deletion \end{Bmatrix}$$

$$w = gap\_weigth$$

In the original paper

$$w = 1 + \frac{1}{3} - k$$

Table 5: Smith Waterson Init first row and first column $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	0	0	0	0	0	0	0
G	0							
A	0							
T	0							
T	0							
A	0							
C	0							
A	0							

Table 6: Smith Waterson first row compute $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	0	0	0	0	0	0	0
G	0	↖ 1	0	0	0	↖ 1	0	0
A	0							
T	0							
T	0							
A	0							
C	0							
A	0							

3 Substitution Matrices

https://en.wikipedia.org/wiki/Substitution_matrix In bioinformatics and evolutionary biology, a substitution matrix describes the rate at which one character in a sequence changes to other character states over time. Substitution matrices are usually seen in the context of amino acid or DNA sequence alignments, where the similarity between sequences depends on their divergence time and the substitution rates as represented in the matrix

3.1 PAM

https://en.wikipedia.org/wiki/Point_accepted_mutation A point accepted mutation (PAM) is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection. A PAM matrix is a matrix where **each** column and row represents one of the **twenty** standard amino acids. PAM matrices are use as substitutionm matrices to score **sequence alignments for proteins**.

Each entry in a PAM matrix indicates the **likelihood of the amino acid** of that row **being replaced with the amino acid** of that column **through** a series of one or more point accepted mutations during a specified evolutionary interval, **rather than** these two amino acids **being aligned due to chance**. In proteins the alphabet goes to 20 (from the 4 of DNA).

3.1.1 Construction of PAM matrices

Each PAM matrix is designed to compare two sequences which are a specific number of PAM units apart. Two sequences $S1$ and $S2$ are at evolutionary distance of PAM1, if $S1$ has converted to $S2$ with an average of one amino acid substitution per 100 amino acids.

For the jth amino acid, the values $m(j)$ and $f(j)$ are its mutability and frequency. The mutability of an amino acid is the ratio of the number of mutations it is involved acceptably with A as the matrix.

$$m(j) = \frac{\sum_{i=1, i\neq j}^{20} A(i,j)}{n(j)}$$

The frequency of the amino acids are normalised so that they sum to 1. If the total number of occurrences of the jth amino acid is $n(j)$ and N is the total number of all amino acids, then

$$f(j) = \frac{n(j)}{N}$$

The mutation matrix M is composed as follow.

$$M(i,j) = \frac{\lambda A(i,j)}{N f(j)}$$

$$M(j,j) = 1 - \lambda m(j)$$

for the diagonals. The PAM matrix tells us if something is more probable than random to mutate or not. $PAM2 = PAM1 \times PAM1$ simple matrix multiplication

Table 7: PAM matrix

	C	S	T	...	W
C					
S					
T					
...					
W					

3.2 BLOSUM

`https://en.wikipedia.org/wiki/BLOSUM`
`https://www.youtube.com/watch?v=xDUzRTx12ZE`
Protein scoring matrices ¡ statistics of conservation, substitution of characters in nature

consisten with natures. Identical amino acids ¿ any substitution conservative substitutions ¿ non-conservative substitutions

Higher identity == lower evolutionary distance

Provide a likelihood for a character substitution

4 Main steps in the construction of a BLOSUM matrix

1. Eliminate repeated sequences in each block of the cluster of blocks whose percentage of identity is at most K.
2. Pair column by column, perserving the order of the block. (each pair is counted twice), store the column in a table
3. Store the total count of pairs in a matrix. $C(X, Y) =$ total number of XY pairs in the sample block.
4. Calculate the frequency of each pair, each number is to be divided by the total number of possible pairs. number of all possible pairs $T = 0.5[ColumnsN \times RowsN(RowsN - 1)]$
5. compute the (frequency) matrix Q. $Q(X, Y) = C(X, Y)/T$
6. find the expected probability $P(X) = Q(X, X) + 0.5 \sum_{Y \neq X} Q(X, Y)$
7. calculate the expected frequencies $E(X, X) = P(X)^2$ $E(X, Y) = 2P(X)P(Y)$
8. the log-odds ration $L(X, Y) = log_2[Q(X, Y)/E(X, Y)]$

Table 8: step 1 and step 2

Seq1	K	P	T
Seq2	K	P	V
Seq3	V	P	A
Seq4	T	P	V
Seq5	T	P	K

Table 9: step 3

Pairs	**Pairs in column 1**	**Total**
KK	1	5
KT or TK	4	7
KV or VK	2	9
TT	1	6
TV or VT	2	8