How to find minimum number of mismatches

Feb 2, 2018

Definitions:

We define the two DNA sequences as S^1 and S^2, and we present S_i as the ith DNA in the sequence S.

We define the maximum length of common subsequence as LM(S^1, S^2). Obviously, LM(S^1, S^2)= 0 means there is no similar DNA between S^1 and S^2.

In another example, let

$$S^1 = \text{BBCC}, \text{length} = l_1,$$
$$S^2 = \text{ABC}, \text{length} = l_2,$$

then LM(S^1, S^2)= 2, as "BC" is the longest common subsequence between S^1 and S^2.

Insights:

Thinking during the interview: If we want to find the minimum number of mismatches, first of all, we have to find the maximum length, i.e., LM(S^1, S^2) of common subsequence of two DNA sequences. Then, according to *Needleman-Wunsch algorithm* we can **get the minimum number of mismatches by tracing back the process of obtaining the maximum length of common subsequence.**

Find LM(S^1, S^2):

First of all, we initialize a matrix M to calculate the LM(S^1, S^2), where M_{ij} denotes to LM($S^1_{0...i}$, $S^2_{0...j}$). We initialize the M in Table 1. Then, according to following rules, we fill the M as shown in Table 2.

(a) If S^1_i=S^2_j, then M_{ij}=$M_{i-1,j-1}$+1.

(b) If $S^1_i \neq S^2_j$, then M_{ij}=Max($M_{i-1,j-1}$,$M_{i,j-1}$,$M_{i-1,j}$).

Trace back:

In this part, we find the minimum number of mismatches by tracing back the process of filling M from M_{l_1,l_2}, i.e., the maximum length of common subsequence by going through following steps. **Remarks:** The red numbers in Table 2 help to illustrate how to choose directions in the process of tracing back.

Step 1: Set the moving direction.

(a) If current location is at $M_{0,0}$, quit the loop.

-	0	A	B	C
0	0	0	0	0
B	0	-	-	-
B	0	-	-	-
C	0	-	-	-
C	0	-	-	-

Table 1: Initialization of M

-	0	A	B	C
0	0	0	0	0
B	0	1	1	1
B	0	1	1	1
C	0	1	2	2
C	0	1	2	3

Table 2: Process of filling M and direction choosing when restructuring DNA sequences

(b) Else if current location is at the first row of M, set the moving direction to LEFT, as now what only we can do is to add gaps in S^1 for completing the length.

(c) Else if current location is at the first row of M, set the moving direction to UP, as now what we can do is to add gaps in S^2.

(d) Else if at current location M_{ij}, S_i^1=S_j^2, set the moving direction to LEFT_UP, as now what we are satisfied with the match.

(e) Else at current location M_{ij}, $S_i^1 \neq S_j^2$, we check that moving to which direction we can get maxium length of common subsequece and then moving to that dimension. (If table cells have same value, we firstly choose LEFT_UP, then UP and finally LEFT.)

Step 2: Move to chosen direction.

(a) If the chosen direction is LEFT_UP, we add S_i^1 to new restructured sequence s^1 and add S_j^2 to s^2.

(b) Else if chosen direction is LEFT, we add a gap to new restructured sequence s^1 and add S_j^2 to s^2.

(c) Else (chosen direction is UP), we add S_i^1 to new restructured sequence s^1

and add a gap to s^2.

Step 3: According to the restructed sequences, we can easily calculate the minimum number of mismatches between two DNA sequences.

Lingkun,
klk316980786@sjtu.edu.cn