BCB BCB/GDCB/STAT/COM S 568 Spring 2011

Homework 1

January 18, 2011

1) **Review of BCB 567 concepts and algorithms.**

A global alignment of two sequences A=$a_1a_2\ldots a_M$ and B=$b_1b_2\ldots b_N$ can be represented by the set of index pairs P = $\{(i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)\}, 1 \le i_1 < i_2 < \ldots < i_k \le M, 1 \le j_1 < j_2 < \ldots < j_k \le N$, where the index pairs (i_x, j_x) indicate that a_{i_x} is aligned with b_{j_x}.

a) The Needleman-Wunsch algorithm imposes the restriction $i_x - i_{x-1} = 1$ and/or $j_x - j_{x-1} = 1$ for $x = 1, 2, \ldots, k+1$ where $i_0 = j_0 = 0$, $i_{k+1} = M+1$, and $j_{k+1} = N+1$ (avoidance of "double gaps"). Prove that the optimal score of a global alignment with end-gap penalties can be calculated as S_{MN}, where S_{ij} is derived recursively at each step as

$$S_{ij} = \max \begin{cases} S_{i-1,j-1} + \sigma(a_i, b_j) & \\ S_{i-1,j-1-p} + \sigma(a_i, b_{j-p}) + w(p) & p = 1, 2, ..., j-1 \\ S_{i-1-q,j-1} + \sigma(a_{i-q}, b_j) + w(q) & q = 1, 2, ..., i-1 \end{cases}$$

provided one specifies correct initial values of $S_{00}, S_{0j}, j = 1, 2, \ldots, N$, and $S_{i0}, i = 1, 2, \ldots, M$ (here (a_i, b_i) is the score for matching a_i with b_j, and $w(x)$ is the gap penalty for a gap of size x).

Solution:
S_{ij} represents the maximal score of alignments of the prefixes $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$, as the maximization is over all possible ways of extending an alignment of shorter prefixes.

a-i) Indicate to what values S_{00}, S_{0j}, and S_{i0} should be set for the recursion to work and how you would obtain an optimal alignment.

Solution:
$S_{00} = 0; S_{0j} = w(j); S_{i0} = w(i)$
To obtain an optimal alignment, one would need to trace back from the cell MN to the 00 cell and record a path that led to the optimal score.

a-ii) How would you change the algorithm to calculate the optimal score for a global alignment without end-gap penalties?

Solution:
$S_{00} = 0; S_{0j} = 0; S_{i0} = 0$
The trace back in this case starts from the cell with the maximum score in the last row (M) or column (N) and stops when row or column 0 is reached.

a-iii) Give an algorithm to derive the number of all possible alignments for sequences of lengths M and N.

Solution: When we fill out the $M \times N$ matrix in the Needleman-Wunsch algorithm, for cell ij we account for $1 + (j-1) + (i-1)$ possible ways of one-step extensions of shorter alignments.

To derive the number of all possible alignments, we can recursively fill out another $M \times N$ matrix with entries N_{ij} that represent the number of all possible alignments between the subsequences $a_1 \ldots a_i$ and $b_1 \ldots b_j$. Then $N_{ij} = N_{i-1,j-1} + \sum_{k=0}^{j-2} N_{i-1,k} + \sum_{k=0}^{i-2} N_{k,j-1}$, where $N_{i,0}$ and $N_{0,j}$ are set equal to 1 for $0 \leq i \leq M$, $0 \leq j \leq N$. N_{MN} is the total number of all possible alignments. Confirm the validity of the following partially filled table by enumerating the alignments for small M and N.

j

		0	1	2	3	4	5	6	7	8	
	0	1	1	1	1	1					
	1	1	1	2	3						
	2	1	2	3	5						
	3	1	3	5	9						
i	4	1									
	5										
	6										
	7										
	8										

b) Derive the algorithm to calculate the optimal score as in (a) but without the restriction of avoidance of double gaps.

Solution:
$S_{00} = 0; S_{0j} = w(j); S_{i0} = w(i)$

$$S_{ij} = \max \begin{cases} S_{i-1,j-1} + \sigma(a_i, b_j) & \\ S_{i,j-p} + w(p) & p = 1, 2, ..., j \\ S_{i-q,j} + w(q) & q = 1, 2, ..., i \end{cases}$$

b-i) Determine the complexity of the algorithm: how many operations are required to calculate the optimal score?

Solution:
The number of additions is seen to be $\sum_{i=1}^{M} \sum_{j=1}^{N} [1 + j + i] = MN + M\frac{N(N+1)}{2} + N\frac{M(M+1)}{2}$. Thus, for M=N, the algorithm is of O(N^3).