Introduction to Bioinformatics (autumn 2005)

Excercise 4

Group	time	place
Riikka Kaven	Tuesday 8.11 at 12.15–14.00	BK106

1. (Problem 6.18 in J & P) What is the optimal global alignment for `MOAT` and `BOAST`? Show all optimal alignments and the corresponding paths under the scoring matrix below and indel penalty -1.

	A	B	M	O	S	T
A	1	-1	-1	-2	-2	-3
B		1	-1	-1	-2	-2
M			2	-1	-1	-2
O				1	-1	-1
S					1	-1
T						2

2. (Problem 6.20) Consider the sequences $v =$ `TACGGGTAT` and $w =$ `GGACGTACG`. Assume that the match premium is $+1$ and that the mismatch and indel penalties are -1.

   - Fill out the dynamic programming table for a local alignment between v and w. Draw arrows in the cells to store the backtrack information. What is the score of the optimal *local* alignment and what alignment achieves this score.

3. (Problem 6.22) Define an *overlap alignment* between two sequences $v = v_1 \cdots v_m$ and $w = w_1 \cdots w_n$ to be an alignment between a suffix of v and a prefix of w. For example, if $v =$ `TATATA` and $w =$ `AAATTT`, then a (not necessary optimal) overlap alignment between v and w is

   ```
   ATA
   AAA
   ```

   Optimal overlap alignment is an alignment that maximizes the global alignment score between $v_i \cdots v_m$ and $w_1 \cdots w_j$, where the maximum is taken over all suffixes $v_i \cdots v_n$ of v and all prefixes $w_1 \cdots w_j$ of w.

   Give an algorithm which computes the optimal overlap alignment, and runs in time $O(mn)$.

4. (Problem 7.5) Develop a linear-space version of the local alignment algorithm.

5. The *Exact Gene Finding* problem is defined as follows. Given a DNA sequence D and a protein sequence P, find a subsequence (gene) S of D with minimum number of *blocks* (exons) such that S *codes* for P. A block in S is its maximal substring that also is a substring of D. DNA sequence S codes for protein sequence P if

(s_1, s_2, s_3) is codon producing amino acid p_1, (s_4, s_5, s_6) is codon producing amino acid p_2, and so on. For example, given $D =$ `GATAAAAGAGTGGTT` and $P =$ `LFHQ`, the solution is $S =$ `GATAAAGTGGTT` with two blocks `GATAAA` and `GTGGTT`: `GAT` codes for Leucine `L`, `AAA` codes for Phenylalanine `F`, `GTG` codes for Histidine `H`, and `GTT` codes for Glutamine `Q`.

Give a dynamic programming algorithm to solve the Exact Gene Finding problem. Simulate your algorithm on the above example.

Hint: Find a recurrence for values b_{ij} that give the minimum number of blocks to align $d_1 \cdots d_i$ with $p_1 \cdots p_j$.