Albert-Ludwigs-Universität Freiburg
Physikalisches Institut

Johannes Berg

Information Theory: From Statistical Physics to Quantitative Biology

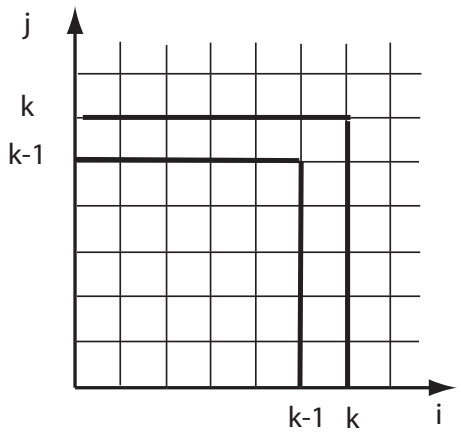
1. exercise class – 4. February 2009

1. Global alignments

Consider the global alignment of two sequences $(a_1, \ldots, a_N)$ and $(b_1, \ldots, b_N)$, given a scoring matrix $\mathbf{s}$ and a gap score γ.

a) Formulate an algorithm which finds the optimal global alignment of the two sequences. How does the number of computational steps scale with N ?

Hint: Draw the alignment lattice and show how to calculate the score of an optimal alignment ending at each of the points along the lines $i = k, j \leq k$ and $j = k, i \leq k$, given the optimal score at the lines given by $i = k-1, j \leq k-1$ and $j = k-1, i \leq k-1$ (see sketch). Show how, by iterating this step, the optimal score can be computed. Explain how the alignment path corresponding to the optimal score can be found by backtracking from the endpoint with optimal score.



b) Formulate an algorithm which finds the optimal alignment which starts at a *given* pair of elements (a_i, b_k) and ends at a *given* pair (a_j, b_l) $(i < j, k < l)$.

c) Formulate an algorithm which finds the optimal alignment which starts at a *given* element a_i and ends at a *given* element a_j $(i < j$, the partners b_k and b_l may be chosen freely).

Hint for **b)**: Draw the two cases on the alignment lattice and find suitable boundary conditions at $t = 0$. In other words, how does the condition that the path does not originate from any point but (i, j) translate into a boundary condition for the score?

Alternatively, consider an algorithm in the 'forward light cone' of (i,j). Analogously for **c)**.
(50 points)

2. Scores with affine gaps

The standard alignment score assigns a score of $-\gamma\, l$ to a gap of length l, e.g. for a gap

$$\begin{array}{cccccccccc} A & A & B & A & B & A & A & B & A & A \\ A & B & B & - & - & - & - & B & A & B \end{array} \tag{1}$$

with $l = 4$ the penalty is -4γ. In practice it is useful to assign a higher penalty $-\gamma_1$ to the first element of a series of gaps, and a lower penalty $-\gamma_2$ to all subsequent gaps,i.e. $-\gamma_1 - 3\gamma_2$ in the case above. Find an algorithm to find the optimal alignment with this new scoring function.

Hint: The recursion relation $S(r,t) = \max(S(r-1,t-1) - \gamma, S(r,t-2) + s(r,t), S(r+1,t-1) - \gamma)$ must be modified to distinguish between single gaps $l = 1$ and sequences of gaps ($l > 1$); to be able to do this the recursion must 'reach further back' in time...
(30 points)