

# Pairwise Alignment

## The dynamic programming approach

1. Inspired by examples of alignments of biological sequences (with the basic operations match, insert and delete), define a suitable computational notion of pairwise string alignment.
2. Design a statistics set-up that is suitable to further define the notion of optimal alignment (score) of two sequences.

### *Hints:*

- a. Define two statistics models: a *random* model (R), and a *match* model (M).
  - b. State certain independence assumptions you can make in order to conveniently compute the probability of a pairwise alignment  $(x, y)$  in each of the two models.
  - c. Given an alignment  $(x, y)$ , compare  $P_R(x, y)$  to  $P_M(x, y)$ .
3. Given two sequences  $x = x_1x_2 \dots x_n$  and  $y = y_1y_2 \dots y_m$ , how large is the number of all their possible alignments?  
What implication arise in what concerns the (further) design of alignment algorithms?
  4. Would dynamic programming (DP) be suitable for computing pairwise alignments? Yes/No? Why?
  5. Define the initialization and the recurrence relations for a DP algorithm for global alignment of two sequences.  
What is the biological/computational significance of the  $(i, j)$  element in the DP matrix?
  6. Propose other versions of pairwise alignment: local, local repeat, overlap, overlap repeat, suboptimal alignments.
  7. What is the complexity of the DP algorithms for pairwise alignment?  
Could we perform (e.g. global) pairwise alignment using linear space?  
What about sub-quadratic time?
  8. Could we define a FSA (finite state automaton) alternative to the DP algorithm for global alignment?
  9. What about designing heuristic algorithms for fast (not necessarily optimal) pairwise alignments?

### ◦ Recommended readings:

Richard Durbin et al. "Biological sequence analysis", 1998, Ch. 2  
Serafim Batzoglou, "The many faces of sequence alignment", 2005