Universität Bielefeld
Technische Fakultät

AG Genominformatik
Dipl.-Inform. Klaus-Bernd Schürmann

Lecture: Spezielle Algorithmen der Sequenzanalyse
Summer semester 2006

Exercises

Exercise 7, Discussion: 05/24/2006.

1. **Multiple Sequence Alignment.**

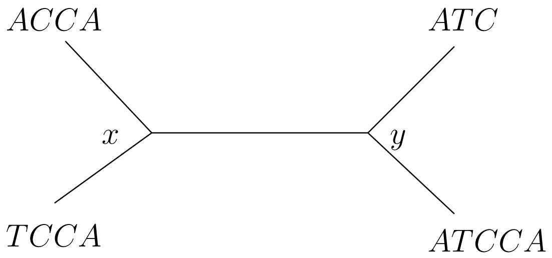
   Given the 4 sequences $s_1 = ACCA$, $s_2 = TCCA$, $s_3 = ATC$, $s_4 = ATCCCA$. The homogeneous *gapcost*= -2 and the following substitution score function:

σ	A	C	G	T
A	3	-2	-1	-2
C	-2	5	-2	-1
G	-1	-2	3	-2
T	-2	-1	-2	5

   (a) Compute the sum of pairs score for the following multiple alignment:

$$\mathcal{A} = \begin{pmatrix} A & - & C & C & A \\ - & T & C & C & A \\ A & T & C & - & - \\ A & T & C & C & A \end{pmatrix}$$

   (b) Given the tree T below. Compute x and y such that the tree alignment score is maximal and give the maximal score.



2. **Tree alignment.**

   Using the PAM250 similarity matrix during all steps of an alignment along a tree is not recommended. Why?

3. **Carrillo-Lipman heuristics.**

   (a) Characterize sequences for which the Carrillo-Lipman heuristic works good, respectively bad.

   (b) How many Carrillo-Lipman bounds $L_{x,y}$ are calculated for k sequences?

   (c) Given the three sequences $s_1 = AGATC$, $s_2 = GAGAT$, $s_3 = TACATA$ and the multiple alignmentscore 10 for a heuristic alignment of the three sequences. Calculate the matrices $M_{i,j}$, $1 \leq i < j \leq 3$ for unit costs. Highlight the regions for which the back-projection into the 3-dimensional edit matrix has not to be computed.