

Algorithmic Bioinformatics II: Exercises

Assignment 7

Deadline: Tuesday, 11.12.2007, 10am (before lecture)

Exercise 1 (T-Coffee):

Read the article “*T-Coffee: A novel method for fast and accurate multiple sequence alignment.*” by C. Notredame, D.G. Higgins and J. Heringa (available on the website).

Describe the algorithm in your own words.

Exercise 2 (Lifted alignment I):

Calculate the assignments for the internal nodes of an optimal lifted alignment of sequences $S_1 \dots S_5$ and the tree $((S_1, S_2), S_3), (S_4, S_5)$. The alignment distances are given by the following matrix:

	S_1	S_2	S_3	S_4	S_5
S_1	0	2	4	5	6
S_2		0	3	7	6
S_3			0	5	5
S_4				0	2
S_5					0

Exercise 3 (Lifted alignment II):

For k sequences of length l a lifted alignment has to be calculated. Answer the following questions (with an explanation of your answer):

- What is the complexity of the distance matrix calculation?
- What runtime is necessary for labeling the internal nodes if at each node the sequences are divided into two equally sized groups.

Hint: Derive the recurrence equation for the runtime. Use the master theorem or series expansion.

- What is the worst-case runtime? What kind of tree can lead to a worst-case runtime?

Exercise 4 (Randomized multiple alignment):

A multiple sequence alignment (MSA) has to be calculated for k given sequences of length l . Let us assume that there is a single (unique) optimal multiple alignment of this sequences. To create a MSA, we combine pairwise sequence alignments created by an imaginary method. This method randomly aligns a given pair of sequences. With probability p such a pairwise sequence alignment corresponds to the alignment induced by the unique optimal MSA. The alignments calculated by the method are statistically independent of each other.

- (a) How many pairwise alignments are necessary to uniquely determine a multiple alignment of all k sequences?
- (b) n multiple alignments can be calculated by generating this necessary number of pairwise alignments for each MSA. What is the probability that at least one of the resulting MSAs is optimal?
- (c) If this probability should be at least q , how many of those 'random' MSAs must be created?