Exercises - Week 9

Genomics and bioinformatics

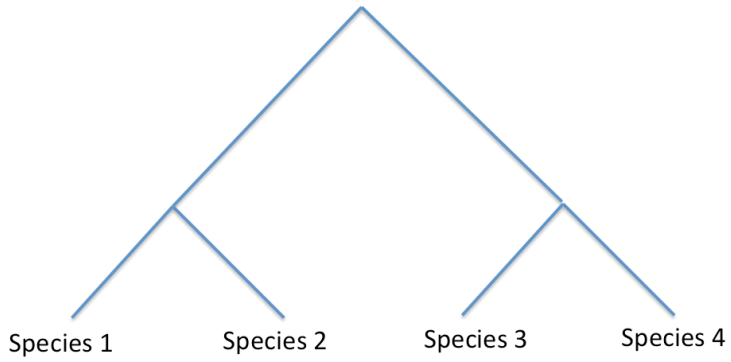
This series is about Phylogenic Tree Reconstruction from Multiple Sequence Alignment (MSA).

1 The Fitch's algorithm

Consider the following MSA:

	1	2	3	4
species 1	A	C	G	T
species 2	C	C	G	T
species 3	A	C	C	G
species 4	A	C	C	T

Given a valid phylogeny tree T for the above MSA (i.e. every leaf of T is labeled by a unique taxa in the MSA), the parsimony length $L(T)$ is the minimum number of mutations required to explain the tree T. The aim of the parsimony problem is to compute the tree T which minimises $L(T)$. The tree T in the following figure is the most parsimonious tree for the MSA (this tree is also found by ClustalW):



Assuming that the columns are independent of each other, use the Fitch's algorithm to compute the parsimony length $L(T)$ and the corresponding assignment of the states to the internal nodes. *Hint:* Go from the leaves to the root (bottum-up) when applying the Fitch's algorithm and in the opposite direction (top-down) when generating the label for each internal node.

2 The Sankoff's algorithm

Consider the substitution matrix M presented in the lectures and apply the Sankoff's algorithm to the first column of the MSA in question 1.

3 The UPGMA algorithm

Questions 1 and 2 used the parsimony methods. Here we shall consider the distance-based approach using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm. Consider the following distance matrix:

M	a	b	c	d	e
a	0	8	8	14	14
b	8	0	2	14	14
c	8	2	0	14	14
d	14	14	14	0	10
e	14	14	14	10	0

Use the UPGMA algorithm to build the rooted tree T corresponding to the distance matrix M.