Assignment #3: Maximum Parsimony

Part I. Sequence alignment and parsimony reconstruction using paper and pencil

1. Sequence alignment

a. Use the Needleman-Wunsch algorithm and the Blosum 62 matrix shown below to align the following two proteins sequences:

```
Seq1: ALIGNME
Seq2: AILMENT
```

```
Partial Blosum 62 matrix.
    A   N   E   G   I   L   M   T   *
A   4  -2  -1   0  -1  -1  -1   0  -4
N  -2   6   0   0  -3  -3  -2   0  -4
E  -1   0   5  -2  -3  -3  -2  -1  -4
G   0   0  -2   6  -4  -4  -3  -2  -4
I  -1  -3  -3  -4   4   2   1  -1  -4
L  -1  -3  -3  -4   2   4   2  -1  -4
M  -1  -2  -2  -3   1   2   5  -1  -4
T   0   0  -1  -2  -1  -1  -1   5  -4
*  -4  -4  -4  -4  -4  -4  -4  -4   1
```

b. Is this the best alignment for these two sequences? Are the sequences homologous? Explain!

2. Fitch-Hartigan parsimony

2) Complete problem 1 on p. 40 in the Mathematics of Phylogenetics textbook.

3. Sankoff parsimony

3) Consider the following alignment with variable sites marked by *:

```
                          **         *        *    *        *  *  *
Dolphin         ATG ACG AAC ATC CGA AAT TCA CAC CCT CTT
Hippopotamus    ATG ACA AAC ATC CGA AAA TCT CAC CCA CTA
Camel           ATG ACA AAC ATC CGA AAA TCA CAC CCA CTA
Cow             ATG ACA AAC ATT CGA AAG TCC CAC CCA CTA
Giraffe         ATG ATA AAC ATC CGA AAG TCC CAC CCA CTA
Sperm_whale     ATG ACA AAC ATC CGA AAA TCA CAC CCA CTA
Blue_whale      ATG ACA AAC ATC CGA AAA TCA CAC CCA CTA
Pig             ATG ACA AAC ATC CGA AAA TCA CAC CCA CTA
Sheep           ATG ACA AAC ATC CGA AAA TCC CAC CCA CTG
Goat            ATG ACA AAC ATC CGA AAG TCC CAC CCA TTA
```

a. Which of the variable sites are parsimony informative? What does it mean?**

b. Use Sankoff's algorithm on parsimony informative sites to find the parsimony score for the following tree: (((Dolphin,Sperm_whale),Blue_whale),((Hippo,Pig),(((Cow, (Sheep,Goat)),Giraffe),Camel))). Weight the cost of transversions 3 times the cost of transitions.

Part II. Maximum Parsimony analysis in PAUP.

For this part of assignment use molecular sequences in the accompanying file prestin_nt.fa. You need to align them and convert your file to NEXUS format before conducting the analyses below. **.5 extra points** for aligning them with [pal2nal](pal2nal) (hint: these sequences correspond to aa sequences we used in class).

Heuristic searches

4) Perform four heuristic searches with one simple-addition replicate but four different branch-swapping options (NO, NNI, SPR, TBR) and another heuristic search with 100 random-addition replicates and TBR branch-swapping option. How, if at all, did your results (topologies of the trees; lengths of trees found, numbers of trees found) differ among the searches? Discuss.

Bootstrap analysis

5) Perform a 200-replicate bootstrap search with 10 random additions TBR replicates per bootstrap rep. Load the bootstrap trees into memory and calculate the strict and the majority rule consensus trees. Are these two trees identical? Explain. Print out the consensus trees (only) and turn them in along with your answers.

Part III. Extra point

What is the position of Dolphin in aa-based and nt-based analyses? What could be the reasons for the difference (if any)?

Good luck!