Biological Computation 20.181

Homework 1

Review of Recursion. In class we looked at a recursive definition for the factorial function f(n) = n! = n*(n-1)*(n-2)*...*1. Write a recrusive function that takes three numbers -- k,first,second -- as input, and returns a list including the first k terms of the Fibonacci series, using the recursive definition F(n+1) = F(n) + F(n-1), and the initial conditions F(1) = first and F(2) = second.

Your solution should be of the following form:

```
def fib(k,first,second):
  #return fibonacci series up to k terms
  #insert code here
```

Introduction to phylogenies. Consider the following five sequences:

1. AAGGCCCACTA
2. GATGTCCGATA
3. AAGGCCCACTT
4. AATGGCCCCTA
5. GATGTCCGATA

Compute all pairwise distances between these sequences in terms of number of mismatched base pairs (hint: sequences 1 and 2 have 5 mismatches, so their distance is 5).

The UPGMA algorithm (which stands for unweighted pair group method using arithmetic averages) is an intuitive clustering method that uses pairwise distances to generate trees. The algorithm goes as follows:

INITIALIZATION
Put each sequence in its own 'subtree' with one leaf.

ITERATION
Find the two subtrees i,j with the smallest pairwise distance d(i,j) between them (if several pairs are equidistant, choose one randomly).
Add a new subtree k by merging these two: add a root node and two branches (one to each subtree).
Add the new subtree k to the current list and remove the two subtrees i,j.

TERMINATION
If only two subtrees remain, join these two and stop.

We know how to compute distances between sequences... but how do we compute distances between 'subtrees'? That's where the 'arithmetic averages' come in. Take the average distance between any pair of sequences in subtree i and j:

d(i,j) = 1/(Size(i)*Size(j)) * [SUM(p in i, q in j) d(p,q)]

Use the UPGMA algorithm to infer a tree showing the most likely evolutionary relationship between these five sequences.

Inferring ancestral states. Make a copy of the tree you inferred above, and at each of the leaves write the first letter of the DNA sequence associated with that leaf. How many internal branchpoints (ancestors) are shown in this tree? What kind of event might lead to a branchpoint?

Try to guess the most likely (parsimonious) sequence of each internal node (ancestor) in the tree. (Hint: some nodes may not be completely defined, *i.e.*, [A/G] or [A/G/C]).

Repeat for each DNA basepair in the sequences from above, and report the likely sequence at each ancestral node.