Problem Set #4

Due Thursday, February 25, 2015, at the beginning of class.

Be sure to read each question carefully and answer all the parts.

1. (20 points) Apply Fitch’s algorithm to the following tree (with the shown nucleotide assignment for the tips). In your answer show the results of both the bottom-up phase (i.e., the set of possible states in each internal node) and the top-down phase (i.e., an actual assignment for each of the internal nodes). What is the parsimony score for this tree?

![Tree Diagram](image)

2. (10 points) If you were now given an additional species (species_15) with some nucleotide assignment and were told that this species is known to be an outgroup (first diverging species) to the above 14 species, by how much could the parsimony score change if you will add this species to the tree? Can the parsimony score go down? Would you have to apply Fitch’s algorithm from scratch to get the new parsimony score?

3. (10 points) Recall the hill climbing algorithm we described in class. Such an algorithm is often referred to as a “local” search. Explain, in your own words, what’s local about this search.

4. (15 points) Figure out how many possible Nearest-Neighbor Interchanges there are on a specific unrooted tree with 8 leaves (that is, the number of competing trees that would be considered in one step of the hill-climbing method using NNIs). Hint: a subtree can be any part of the tree, including a single leaf. Justify your answer.

5. (15 points) Write a function that gets two lists as arguments and returns their intersection. The function should work both for lists of numbers and for lists of strings (or for mixed lists). See the code below. Hint: for a variable x and a list L, the expression “x in L” returns true if x appears in the list L and false otherwise (and can therefore be used in an “if” command).

```python
def my_intersection(list1, list2):
    # YOUR CODE COMES HERE

list_num = my_intersection([1,3,5,21],[5,4,19,21])
print list_num # should print [5,21]

list_mix = my_intersection([1,"r","hello",7],[4,"hello"])
print list_mix # should print ["hello"]
```
6. (30 points) Write a program that finds the most common word (or words) with N or more letters and the number of times they appear in a text document, where N and the document name are command-line arguments. Ignore case (Natasha and natasha are the same word) and be sure to get rid of the common punctuation marks (.,"?;:-()!). Use a function as part of your program. In addition to showing your program, give the answer for N from 4 through 9 in Tolstoy’s War and Peace (the text can be downloaded from the course website). By the way, you will learn a better way to remove a set of characters from a string in a couple of weeks.

7. Challenge problem. A perfect binary tree (i.e., a full bifurcating tree in which all leaves are at the same depth or same level), can be conceptually mapped into a list or a dictionary, such that the root is mapped into index 0, and the two children of a node that is mapped into index i, are mapped into indices (i*2)+1 and (i*2)+2; (see illustration below). With this representation, you can calculate the index of a parent given the index of the child and vice-versa. Using this idea, write a program that gets the number of species, N, in a perfect binary tree (hint: which indices in the list/dictionary would these species occupy?), and N characters that represent the nucleotide assignment of these species from left to right (see example). The program should use Fitch’s algorithm (hint: you can use the intersection function you wrote in problem 5 and additionally write a union function) and print the parsimony score.

---

```python
>>> python Fitch.py 8 C T C T G G A A
Parsimony score = 4
```

8. Challenge problem. Improve your program from problem 8 so that it will also print the minimal assignment of all internal nodes, using Fitch’s algorithm second (top-down) phase.

For example:

```python
>>> python Fitch.py 8 C T C T G G A A
Parsimony score = 4
Assignment: T T G T T G A C T C T G G A A
```

This means that the proposed internal assignment is: