CSCI3220 Algorithms for Bioinformatics 2018-2019 First Term
The Chinese University of Hong Kong

Assignment 4

Due date: 9 December 2018 (Sun) 23:59 Full mark: 100
Expected time spent: 3-6 hours

Aims:
1. Get familiar with concepts related to phylogenetic tree reconstruction and genetic data inference.
2. Practice applying the learned concepts to solve some new non-trivial new problems.
3. Conduct a small amount of research on specific topics.

Description:
In this assignment, you will reconstruct the phylogenetic trees among some sequences using the UPGMA and Neighbor-Joining methods. You will then implement the maximum parsimony algorithm, and your program needs to handle phylogenetic trees in the Newick format . Finally, you will answer several questions about genetic data inference, which require you to make use of some knowledge acquired in previous lectures. You may also need to conduct a small research when answering the last part of the question.

Questions:

1. This question is about phylogenetic tree reconstruction. You are given the following distance matrix among 5 sequences:

	s_1	s_2	s_3	s_4	s_5
s_1	0	8	6	8	2
s_2	8	0	2	6	8
s_3	6	2	0	10	6
s_4	8	6	10	0	6
s_5	2	8	6	6	0

   (a) Is it possible to reconstruct an additive tree for these five sequences? Explain why or why not. (3%)

   (b) Use the UPGMA algorithm to reconstruct a phylogenetic tree among the five sequences. You need to show the distance matrix and tree topology after every step, but do not need to show branch lengths.

   During the reconstruction process, if multiple pairs of clusters have the same smallest distance, use the following rule to break ties. First, get the indices of all the sequences in the two clusters and sort them in ascending order. Then treat the resulting string as a signature of the pair and merge the one ordered first lexicographically. For example, if the pairs ($\{s_1, s_3\}$ and $\{s_2, s_5\}$), ($\{s_1, s_3\}$ and $\{s_4\}$) and ($\{s_2, s_5\}$ and $\{s_4\}$) all have the same smallest distance, the pair ($\{s_1, s_3\}$ and $\{s_2, s_5\}$) is merged because its signature is 1235, which is lexicographically smaller than both 134 and 245. (17%)

   (c) Use the Neighbor-Joining algorithm to reconstruct the phylogenetic tree among the five sequences. Show the tree, distance matrix, u vector and Q matrix after every step. Break ties using the same rules as in Part b. (20%)

2. Write a computer program called **SmallParsimony** in C, C++, Java or Python that finds the ancestral DNA sequences based on maximum parsimony. The program only needs to implement the simple version of the algorithm, i.e., it only needs to output **one** optimal solution.

It should take a single line as input, namely a tree in Newick format. You can assume the followings:

- The tree is binary
- Each leaf node is labeled by a DNA sequence of the same length
- Each internal node is labeled with an ID
- No branch lengths are provided in the input tree
- The input tree is in correct Newick format

The program should print the inferred sequences of the internal nodes to standard output (stdout), where each line reports the inferred sequence of one internal node in the following format:

<Node label>:<Node sequence><newline>

The nodes in the output can follow any order.

For example, if your implementation is Java, below are several typical runs of the program (based on examples from the lecture notes):

```
>java SmallParsimony
((A,C)x1,((G,T)x2,A)x3)x4;
x1:A
x2:T
x3:A
x4:A

>java SmallParsimony
((A,C)x1,(((A,A)x2,G)x3,G)x4)x5;
x1:A
x2:A
x3:G
x4:G
x5:A

>java SmallParsimony
((AC,GC)x1,GT)x2;
x1:GC
x2:GC
```

The files Newick.java and TreeNode.java are provided as a reference of parsing a Newick string and storing the result in an internal data structure. You may use it (if you use Java) or any external library for parsing Newick strings. If you do, state clearly what library you use and how your program can be compiled and run with this library. Of course, you may also implement your own parser.

Your program will be graded based on i) whether it can be compiled/run successfully, ii) whether it follows the input/output formats as specified above, iii) its accuracy on a number of test cases and iv) whether the program is well-documented with appropriate comments added to explain the meaning of the code. (40%)

3. This question is about genetic data inference. In the whole question, we consider diseases caused by a single gene and assume there are no somatic mutations. The disease considered in each part of the question can have different properties.
   Hint: You may use the Excel file to help you deduce/verify some answers.

(a) In a family quartet, if the father, the mother and one of the children all do not have a disease, is it possible for the other child to have the disease? If so, explain how this can happen; If not, explain why it cannot happen. (5%)

(b) Consider a three-generation pedigree with a child, the two parents and the four grandparents. If the disease status of the parents is already known, can the disease status of the grandparents be useful for inferring the disease status of the child? If so, describe one situation in which it is useful; If not, explain why it is not useful. (5%)

(c) Suppose for a certain disease, we know where the causal gene is and whether the disease allele is dominant or recessive. Suppose we also know the disease status and the genotype of the disease gene of a daughter. Is the above information sufficient to infer for sure the genotype and the disease status of the father? If so, explain all the situations in which this inference can be done; If not, explain why it is impossible. (5%)

(d) Suppose in a family quartet, both parents and one of the children (but we do not know which one) are heterozygous, what is the probability that the other child is also heterozygous? (5%)

(e) [Optional] Suppose among all the descendants of a couple, all the males are affected by a disease but none of the females is affected. If this disease is not Y-linked, give three possible explanations for this observation. (bonus 5%)

Submission:

For the programming question, you should first submit your program to our **online judge system**. Please see tutorial notes set 1 for explanations.

For the non-programming question, give all your answers in a single file named <ID>_asmt4.<ext>, where <ID> is your student ID and <ext> is either doc, docx or pdf. We prefer pdf files because it has better portability. Then put your files for both the programming and non-programming questions in a zip file named <ID>_asmt4.zip and submit it to Blackboard.

Both your written and source code files should contain the following header. Contact Kevin before submitting the assignment if you have anything unclear about the guidelines on academic honesty.

```
CSCI3220 2018-19 First Term Assignment 4

I declare that the assignment here submitted is original except for source
material explicitly acknowledged, and that the same or closely related material
has not been previously submitted for another course. I also acknowledge that I
am aware of University policy and regulations on honesty in academic work, and
of the disciplinary guidelines and procedures applicable to breaches of such
policy and regulations, as contained in the following websites.

University Guideline on Academic Honesty:
http://www.cuhk.edu.hk/policy/academichonesty/

Student Name: <fill in your name>
Student ID  : <fill in your ID>
```

Marking Scheme and Notes:

1. Remember to submit your assignment by 23:59pm of the due date. We may not accept late submissions.

2. For the written part, if you submit multiple times, **ONLY** the content and time-stamp of the **latest** one before the submission deadline will be considered. For the program, the version submitted to the online judge system with the highest score will be graded.

University Guideline for Plagiarism

Please pay attention to the university policy and regulations on honesty in academic work, and the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details can be found at http://www.cuhk.edu.hk/policy/academichonesty/. With each assignment, students will be required to submit a statement that they are aware of these policies, regulations, guidelines and procedures.