Algorithms in Bioinformatics

Spring 2018

Project 1, week 8

The goal of this project is to use the k^{th}-order markov method to classify a set of short genomic sequences. Each short sequence will be classified to one group, which is represented by a genome in a reference genome data set. The file containing short genomic sequences can be found at /share/home/ccwei/courses/2018/pab/proj1/reads.fa. The 10 genomes can be found under the directory /share/home/ccwei/courses/2018/pab/proj1/genomes/. A test set has been created for you: 20,000 short reads are given in /share/home/ccwei/courses/2018/pab/proj1/test/test.fa. The mapping of aligned genomes and the short sequences listed in test.fa is given in a file called seq_id.map file. You can use these files to test your own script.

You are expected to include in your report the implementation of your k^{th}-order markov method together with the following results. Your scripts needs to be handed in as supplementary materials.

1. the total number of short sequences in the file reads.fa;
2. (*) the number of reads that can be assigned to a genome in the genome dataset (with whatever criteria you use);
3. (*) the number of reads that can't be assigned to any genome in the genome dataset ;
4. the number of groups with at least j short sequences assigned, where j = 1, 5, 10 or 50. You can describe the results in a table similar to the one below.

Minimum number of short sequences in a group	Number of groups
1	
5	
10	
50	

5. for those groups with at least 10 short sequences assigned, list the total numbers of short sequences assigned to those groups. Then you can draw a pie chart to show the relative frequency of each group.

This is a small-scale research project using the knowledge learned from your course. You need to give detailed information about each step in your analysis so that your result can be reproduced by others if it is needed.

Students are encouraged to form a team of two or three (at most three), and submit a report for the whole team instead of a report for each individual. However, you have to describe the contribution of each individual in the report if you choose to work as a team. The lecturer may pick some of you to give a short presentation about your project.

(*Methods other than the kth-order markov model are also allowed to finish this project. However, you have to answer these questions started with * if you decide to use a method other than the kth-order markov method.)

Turning in your project work

Submit your result files (all in one file and name it as Yourname_yourID_proj1.doc or pdf, please) to TA (Huimin Lu at linuslu6@outlook.com) before 10:00AM May 10, 2018. The result file(s) should include your project report and your source codes for the project. Please put your source codes separately so that we can run it directly. Also, please give the command lines to generate your results in the project report so that others can reproduce your results.

--cut-----here---

独立作业承诺：（请选择一个， 并签名）

1. 本人，__________，保证本项目由自己独立完成。

   签名

   时间　　　年　月　日

或者

2. 本人，__________，保证本次作业为和_______________同学合作完成，并由自己独立完成报告。

   本人在项目中的的贡献包括____

   ____

   签名

   时间　　　年　月　日