Programming languages for Bioinformatics
Spring 2019

Project, week 11

(All files mentioned below can be found under directory /share/home/ccwei/courses/2019/plb/proj1/ in the course server).

1. Write a program to find differences between two files containing bioinformatics data.
   Synopsis:

   ```
   Biodiff [options] from-file to-file
   ```

   If you have two files A (from-file) and B (to-file), you are expected to generate all lines in A-B, A & B, and B-A in terms of the criteria you set. The file format of file A and B can be different. There will be two styles for comparison: one is coordinate based (option `-c`) and the other is name based (option `-n`). You can set one of these two options as the default style. The two styles were described as follows.

   1). Coordinate-based diff. Two or more columns from file A and B will be selected and compared to check if the two regions overlap. If two regions from the two files overlap, then these two regions will be put into to A&B_A and A&B_B; those regions in A but not in A&B will be put into A-B; and those in B but not in A&B will be put into B-A. Note, the comparison is based on the coordinates specified by two columns set by the user, but the output result contains whole lines in the original files.

   For example, we have two example files *A_ucsc_genes.txt* and *B_ucsc_gene.gtf.* If you run

   ```
   Biodiff -c -a 3,4 -b 3,4 A_ucsc_genes.txt B_ucsc_gene.gtf
   ```

   Column 3 and 4 from A_uscs_genes.txt file will be selected to represent a region and column 3 and 4 from B_ucsc_gene.gtf file will be selected to represent a region, then they are compared. If these two regions overlap, it should generate 4 result files corresponding to A&B_A, A&B_B, A-B, and B-A, where A&B_A contains those lines from file A and overlap with some entries in file B; A&B_B contains lines from file B and overlap with entries in file A; A-B contains those lines from file A and have no overlapping entries in B; and B-A stands for those lines from file B but have no overlapping entries in A.

   2) Name-based diff. Two columns from file A and B will be selected and compared in terms of string comparison. Users need to specify the column numbers in two files to be compared. For example, two example track files (A_ucsc_genes.txt and B_ucsc_gene.gtf) were downloaded from the WashU Genome Browser website (http://genomebrowser.wustl.edu). Both Files A_ucsc_genes.txt and B_ucsc_gene.gtf contain some UCSC genes with different file formats. If you run

   ```
   Biodiff -n -a 0 -b 8 A_ucsc_genes.txt B_ucsc_gene.gtf
   ```

   The first column from A_uscs_genes.txt file and the 9th column from B_ucsc_gene.gtf file will be selected and compared. If their names “overlap”, it should generate 4 result files corresponding to A&B_A, A&B_B, A-B, and B-A, where A&B_A contains those lines from file

A and overlapping with some entries in file B; A&B_B contains lines from file B and overlapping with entries in file A; A-B contains those lines from file A and with no overlapping entries in B; and B-A stands for those lines from file B but with no overlapping entries in A. Here, we call a string s "overlaps" with another string t, if s contains the whole string t or t contains the whole string s.

Please write your program in C and test it thoroughly. Your program is expected to deal with very large size files (the test files may be of hundred MBs). Both the accuracy and speed will be evaluated for your program. (Hint: when you compare two files, first sort the entries in each file based on the column of your pick; then compare them.)

In addition, please write your code as pretty as you can and put as much explanation as you can.

You report should include at least 4 parts: 1). Design of the program; 2) implementation of the program; 3). Usage of your program and test examples together with results. 4) Conclusions and discussions. Part 2 should include the source code as the appendix.

Turning in your project report

Please hand in a hard copy and an electric copy of your homework report, which includes the source code, how you compile it, how you test your program and the result of the test run of you program. You are strongly suggested to test your code in a local machine or in the teaching server first before you submit your homework to the submission website. Please submit your electric copy to TA's email box. The project report should be handed in before the class start on June 13th, 2019.

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

独立作业承诺：（请选择一个， 并签名）

1. 本人，_____，保证本次作业由自己独立完成。

   签名

   时间　　年　月　日

或者

2. 本人，_____，保证本次作为和_____同学讨论后，由自己独立完成。
   讨论内容包括_____

   签名　　，

时间　　年　月　日