Project I
Advanced Applications of Machine Learning (2016 Fall)
(Due October 28, 2016)

Overview: In this homework, you are asked to implement a machine learning algorithm to predict whether an RNA sequence can form a circular RNA (or circRNA).

For final submission, you need to write a report in *English* to describe your algorithm, evaluate the performance of your algorithm through numerical results, and make a discussion.

Policy:

- You are allowed to use any reference from papers, books or materials from Internet. If so, please cite these references in your report. If you have discussed with other people (e.g., professors, friends or classmates), please mention these discussions in the acknowledgement section of your report.
- Please provide experimental results through tables or figures in your report, instead of asking TAs to run your codes.
- We have optional questions. We will give bonus points to those who have addressed these questions.
- You are allowed to form a team (with ≤ 2 persons). In that case, please clearly describe the exact contribution and provide the percentage of contribution of each team member.

Questions: If you have any question about this homework, please contact TAs (Fangping, wfp15@mails.tsinghua.edu.cn; Zhang Sai, zs11235@gmail.com).

Academic Honor Code: This homework must be done independently. We have zero tolerance on cheating and plagiarism.

1 Problem Description

In this problem, you are given reference genome of *Homo sapiens* (i.e., the DNA sequences) and the genomic loci of all circRNAs. Given a certain genomic locus, your goal is to develop a classifier to predict whether a pair of two loci will form a circRNA or not.

2 Getting started

2.1 Data format

FASTA format is used to store reference genome. Details of this format can be found in `https://en.wikipedia.org/wiki/FASTA_format`. If you are a python user, we suggest you to use tools like biopython to parse fasta files (`https://www.biostars.org/p/710/`).

All other data are stored in BED format. *chrom* is the name of the chromosome. *chromStart* is the starting position of the circRNA in the chromosome. Note that the first base in a chromosome is numbered 0; *chromEnd* is the ending position of the circRNA in the chromosome. Note that *chromEnd* base should not be included in the display of circRNA. For example, the first 200 bases of a chromosome are defined as *chromStart*=0, *chromEnd*=200, and span the bases numbered 0-199. *strand* defines the strand that circRNA lies in. Note that the FASTA file of the reference genome only contains + strand, you should covert + strand to − strand (i.e., reverse the complement strand of + strand) by yourself. A complete description of BED format can be found in `http://genome.ucsc.edu/FAQ/FAQformat.html#format1`.

hsa_hg19_Rybak2015.bed stores the genomic loci of circRNAs (positions where DNA sequence can be transcribed and form circRNAs) of *Homo sapiens*. These data are considered as positive examples.

all_exons.bed store thes genomic loci of exons. You should use this file to construct negative examples. All exons that are not overlapped with circRNAs can be considered as negative examples.

hg19_Alu.bed stores the genomic loci of Alu, which is for bonus 1 (see Section 2.5).

2.2 Helpful materials

[1] proposed a method to predict circRNAs. [2] [3] [4] [5] provided some sequence-based machine learning methods in computational biology.

2.3 Implementation

Implement your algorithm in any programming language that you are familiar with, such as Java, C/C++, Matlab, Python, etc. You are allowed to call any other available public package in your program. If so, please include the library in your final submission.

2.4 Measure your algorithm

Use 10-fold cross validation and report the average AUC (area under ROC curve), AUPR (area under precision-recall curve) and F1 score.

2.5 Bonus 1

As mentioned in [6], Alu elements are commonly seen in circRNAs. We also provide the Alu fasta file. Try to incorporate this feature into your algorithm and compare the difference of performance of your algorithm with or without this feature.

2.6 Bonus 2

In Section 2.1, we simply use exons as negative examples. Based on the statistics of circRNAs (like the GC content and the start-end distance of a circRNA), can you design a new way to construct negative examples? If so, does such a choice of negative examples lead to better performance?

3 Requirement of Report

In your final report, you should address the following points:

(1) Details of your algorithm, such as overview, pseudo-code (or flow chart), etc.

(2) Performance evaluation of your algorithm.

(3) Discussion about strength and limitation of your algorithm.

4 Final Submission

For final submission, you need to provide: (1) report; (2) source code and binary executable file of your program, and a short readme file that describes how to compile and run your program.

References

[1] Pan X, Xiong K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features.[J]. Molecular Biosystems, 2015, 11(8):2219-2226.

[2] Zhou J, Troyanskaya O G. Predicting effects of noncoding variants with deep learning-based sequence model.[J]. Nature Methods, 2015, 12(10).

[3] Ghandi M, Lee D, Mohammad-Noori M, et al. Enhanced regulatory sequence prediction using gapped k-mer features[J]. PLoS Comput Biol, 2014, 10(7): e1003711.

[4] Alipanahi B, Delong A, Weirauch M T, et al. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning[J]. Nature biotechnology, 2015.

[5] Zhou Y, Zeng P, Li Y H, et al. SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features[J]. Nucleic acids research, 2016, 44(10): e91-e91.

[6] Jeck W R, Sorrentino J A, Wang K, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats[J]. Rna, 2013, 19(2): 141-157.