Spring 2018 STAT115 Homework 5

HMM, Epigenetics, GWAS Interpretation

(your name)

2018-03-29

Part I: Hidden Markov Models

CpG islands are stretches of CG-rich sequences in the genome. They are often of functional importance, as 50% of the human genes have a CpG island around 500bp upstream of the transcription start site. Of course, CpG island sequences are not only CG's, and non-CpG island sequences could still contain some CG's. Therefore, we could use HMMs to predict CpG islands by looking at a long stretch of DNA. Now as a HMM practice, we just have a short sequence AGGCGT.

The parameters of the HMM model are: Initial probability: 0.4 of CpG (abbreviated as C) and 0.6 of non-CpG (abbreviated as N). Transition probability: P(CpG to CpG) = 0.7, P(non-CpG to non-CpG) = 0.6. Emission probability: P(A, C, G, T | CpG) = (0.1, 0.4, 0.4, 0.1), P(A, C, G, T | non-CpG) = (0.3, 0.2, 0.2, 0.3).

1. Use the forward-backward procedure to calculate the probability of CpG at every observation position and Viterbi algorithm to calculate the Hidden state path. Output the following probabilities:

Algorithm	t_1	t_2	t_3	t_4	t_5
	α_1(CpG)	α_2(CpG)	α_3(CpG)	α_4(CpG)	α_5(CpG)
Forward	α_1(N)	α_2(N)	α_3(N)	α_4(N)	α_5(N)
	β_1(CpG)	β_2(CpG)	β_3(CpG)	β_4(CpG)	β_5(CpG)
Backward	β_1(N)	β_2(N)	β_3(N)	β_4(N)	β_5(N)
	γ_1(CpG)	γ_2(CpG)	γ_3(CpG)	γ_4(CpG)	γ_5(CpG)
Forward-backward	γ_1(N)	γ_2(N)	γ_3(N)	γ_4(N)	γ_5(N)

2. After finishing calculating, you can compare your results with the R package "HMM".

3. For Graduate Students: Adjust the initial probabilities to a range of values and evaluate its effect on the outcome of the forward-backward inference.

4. For Graduate Students: Describe the similarity and differences, in terms of the calculation steps, the R command, and the final paths, if you were to use the Viterbi algorithm instead of forward-backward.

If we use the HMM to solve the following bioinformatics problems, what is the observations (e.g. coin flip), and what is the hidden states (e.g. the coins used)? What's the transition probability (e.g. the transition from fair to biased coin) and what is the emission probability (e.g. P(H | fair coin).

5. Predict protein secondary structure from amino acid sequence.

6. Predict genome-wide chromatin states from histone mark ChIP-seq.

7. Predict TAD domain boundaries based on HiC interaction map.

8. Predict copy number variations from whole genome sequencing data.

Part II: Python programming

From UCSC download page http://hgdownload.soe.ucsc.edu/downloads.html, download the human RefSeq annotation table (find the file refGene.txt.gz for Hg38). To understand the columns in this file, check the query annotation at http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/refGene.sql.

9. Write a python program that can calculate the length of each gene (txStart to txEnd), length of full transcript (concatenated exons) and length of all introns concatenated. Write the output as a tab-delimited table, with RefSeq ID on first column, and the three numbers (length of gene, transcript, introns) after.

Hint: The TSS is different for genes on positive or negative strand. That is, the TSS is "txStart" for genes on the positive strand, "txEnd" for genes in negative strand. When testing your python code, just parse out the first 100 transcripts to check whether the first number is the sum of second and third number for each RefSeq.

Part III: Feature selection and regression, epigenetic gene regulation

Different histone marks are enriched in different elements of the genome and have different effect on gene expression. In this homework, we want to look at the K562 cell line with gene expression data and ChIP-seq profiles of 10 different histone marks: H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K9ac, H3K27me3, H3K27ac, H3K79me2, H3K36me3, H4K20me1. The gene expression data for each RefSeq transcript is summarized in a file called `data/k562expr.txt`.

For each histone mark ChIP-seq data, we already parsed out the following read counts for each RefSeq sequence (in the file `data/histone_marks_read_count_table.txt`): distal promoter [-5KB, -1KB] from transcription start site (TSS), proximal promoter [-1kb, +1kb] from TSS, gene body (from transcription start to end, including all exons and introns), transcript (concatenate all the exons), first 1/3 of transcript (concatenate all the exons, length-wise), middle 1/3 of transcript, last 1/3 of transcript, all the introns (concatenate all the introns). The table has one line for each RefSeq, and 81 columns (RefSeq ID, 10 histone marks, each with 8 features, so 1 + 10 * 8), the value is log read count for each feature.

10. Write a quick python script to filter out RefSeq that have only histone mark or expression data available but not both.

11. For Graduate Students: based on the histone mark count table, do you see enrichment of some histone marks on the different parts of genes?

12. Run linear regression of each of the histone mark features (one of 81 columns) with gene expression. List the histone mark features statistically correlated with gene expression. Which feature is the most positively correlated with gene expression? Which is the most negatively correlated?

Hint: Do you need multiple hypothesis correction?

13. Draw a PCA plot of the different columns in the count table. How do features in the PCA correspond to the significant correlated features in Q12?

14. For Graduate Students: since different transcripts have different length, the read count table might need to be normalized by the promoter/gene/transcript/exon/intron length. Try to normalize each read count column by region length. How does that change the results in Q12 and Q13?

Install glmnet from: http://cran.r-project.org/web/packages/glmnet/index.html We want to select a small subset of histone mark features that best recapitulates gene expression.

```r
install.packages("glmnet")
```

15. Run LASSO regression to select the most informative histone mark features of gene expression. How many features are selected and what are those by LASSO? List the strongest 5 factors.

16. Do a 3-fold cross validation to see how good are the histone mark features selected by LASSO are at predicting gene expression in the training and testing data?

17. For Graduate students: Based on your LASSO model, how well does histone mark predict gene expression? Plot the residual between the predicted gene expression (from histone marks) and the actual gene expression (Y axis) along different gene expression level (X axis)? Does the residual look normally distributed along the gene expression level? Can you guess why?

Hints: http://liulab.dfci.harvard.edu/publications/NucleicAcidsRes12_6414.pdf

Part IV: GWAS and epigenetics

Visit the ENCODE Encyclopedia page at: https://www.encodeproject.org/data/annotations/

You will see candidate Regulatory Elements (cREs) and an entry to SCREEN. Watch the tutorials, especially on GWAS. Browse the list of genome-wide association studies (GWAS) in SCREEN to see how many significant SNPs each study reported and whether based on these hits, the ENCODE project is able to find an epigenetic profile with peaks that significantly matched to these SNPs.

18. Find the study on QT interval by Arking et al (PMID 24952745). This GWAS study identified 73 SNPs significantly associated with QT interval, which allowed SCREEN to identify cells whose epigenetic profile significantly overlap with the 73 GWAS SNPs. What is the cell type with enhancer-like cREs that most significantly overlaps SNPs identified in this GWAS? For SNP rs6843082, comment on the gene is it likely to regulate and why this might be related to QT interval.