Final Project

Transcription Factor-DNA binding prediction

Introduction

The goal of this project is both to have some hands-on experience on real-world machine learning problems, and to potentially solve a very challenging problem in biology.

This project is team-based. Each team can have at most four members, and at least two members. Projects of larger teams will be graded with higher expectations.

Background

Transcription factors (TFs) control the expression of genes through sequence-specific interactions with genomic DNA. Different TFs bind preferentially to different sequences, with the majority recognizing short (6-12 base), degenerate ‘motifs’. Because many types of genomic analyses involve scanning for potential TF binding sites, modeling the sequence specificities of TFs is a central problem in understanding the function and evolution of genomes.

Ideally, models of TF sequence binding specificity should predict the relative affinity (e.g. dissociation constant) to different individual sequences, and/or the probability of occupancy at any position in the genome. Currently, the major paradigm in modeling TF sequence specificity is the Position Weight Matrix (PWM) model. However, it is increasingly recognized that shortcomings of PWMs, such as their inability to model gaps, to capture dependencies between the residues in the binding site, or to account for the fact that TFs can have more than one DNA-binding interface, can make them inaccurate. Alternative models that address some of the shortcomings of PWMs have been developed, but their relative efficacies have not been directly compared.

A major difficulty in studying TF DNA-binding specificity has been scarcity of data. The process of training and testing models benefits from a large number of unbiased data points. In the case of TF binding models, the required data is the relative preference of a TF to a large number of individual sequences. Recently, Protein Binding Microarrays (PBMs) have been developed for the purpose of determining TF sequence preferences (Berger et al. 2006). In a nutshell, each array consists of ~40,000 unique probe nucleic acid sequences (each containing 35 bases). The array is designed so that all possible 10-mers, and 32 copies of every non-palindromic 8-mer are contained on each array, offering an unbiased survey of TF binding preferences. The PBM data provide a quantitative score representing the relative binding affinity of a given TF to the sequence of each probe contained on the array.

The Challenge

The dataset for this challenge is derived from the PBM data for 20 mouse TFs. Two arrays (with completely different probe sequences) are used in the experiment. For each TF and each array, I selected the 1000 probes with the highest binding signals (positive sequences, ‘p’) and 3000 probes with the lowest binding signals (negative sequences, ‘n’). The challenge is to predict the class labels for the sequences presented on one array, given the sequences and their class labels on the other array.
Dataset

For 10 TFs, data is provided from both arrays, for “practice” and method calibration. These are available under the “practice” folder. For example, TF_5_data_1.txt contains the positive and negative sequences for TF_5 from array 1. You can build classification models for each TF (using, for example, data from array 1) and evaluate their performance using both cross-validation and data from the other array.

The challenge consists of predicting the class labels for the remaining 10 TFs. Under folder “challenge”, you will find files for TF_11 to TF_20. For each TF, the sequences from both arrays are given, but class labels are given only for one array. You should use the one with class labels for model training, and submit your prediction results for the one without class labels for evaluation.

Data format:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGGCCGTACGAGTAACGGACTGGCTGTCTTCTCGT</td>
<td>n</td>
</tr>
<tr>
<td>CCGATACCCCCCACCGAABACCTACACATCAAAT</td>
<td>p</td>
</tr>
<tr>
<td>AGCTAAGTAGTCCTCTTCTGGATAGTGAGCGT</td>
<td>n</td>
</tr>
<tr>
<td>AGACAGAAATGCTCGGCCGCCTGCTGACTGAAT</td>
<td>p</td>
</tr>
<tr>
<td>GTAGGACAACAAATATTGCCGTGTGTACCAGTGAC</td>
<td>n</td>
</tr>
<tr>
<td>ACGCGGTGGCGCATGGTGCTCCGAAAGTGTTGT</td>
<td>n</td>
</tr>
<tr>
<td>CTATATCTACGCGGCCACATATTAGCTGCTAG</td>
<td>p</td>
</tr>
<tr>
<td>TGCTCCTTTTCGCGGTCCCAGCAGCAGCAGCAGAC</td>
<td>n</td>
</tr>
</tbody>
</table>

Project requirement

Each team needs to attempt both a consensus-based model and a PWM-based model.

Submission

Submissions will be handled by email. Please email submissions to cs6243finalproject@gmail.com. For regular inquiries, you can use my cs email: jruan@cs.utsa.edu.

1. A zip file that contains all predictions.
   a. For each TF and a particular model, you predictions should be placed in separate files in the same folder, with intuitive folder and file names, for example: kmermodel/TF_11_data_2_pred.txt, or pwmmodel/TF_20_data_1_pred.txt The file format is as follows.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Label</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGCTCCTTTTCGCGGTCCCAGCAGCAGCAGCAGAC</td>
<td>n</td>
<td>0.005</td>
</tr>
<tr>
<td>AGCTAAGTAGTCCTCCTTGGATAGTGAGCGT</td>
<td>n</td>
<td>0.345</td>
</tr>
<tr>
<td>AGACAGAAATGCTCGGCCGCCCGGTACTGAAT</td>
<td>p</td>
<td>0.985</td>
</tr>
<tr>
<td>GTAGGACAACAAATATTGCCGTGTGTACCAGTGAC</td>
<td>n</td>
<td>0.489</td>
</tr>
</tbody>
</table>
2. A short write-up (1~2 pages) with a description of your methods and some analysis. The analysis is extremely important in case your prediction went wrong for some simple / forgivable mistakes.

Milestones and grading

1. April 25\textsuperscript{th}, 7pm – send group names of group members to instructors. (5 points)

2. April 30\textsuperscript{th}, 7pm – send preliminary predictions to instructor. (20 points) The instructor will use a script to check the performance of your prediction and return to you. This will not only help ensure that your output has the right format, but also give you a chance to know how your results compare with your peers. Points will be given mainly for having the right format as long as there is evidence of reasonable effort.

3. May 8\textsuperscript{th}, 11:59pm – final prediction and writeup are due. (65 points)

4. May 9\textsuperscript{th}, 5-7pm – final presentation and analysis. (10 points)