CS229 Predicting Transcription Factor binding

Authors:
Jon Tatum (jdtatum)
Daniel Williams (danielhw)
Zoey Zhou (cuizy)

Advisor: Sofia Kyriazopoulou-Panagiotopoulou (sofiakp@stanford.edu)

Introduction

Since scientists have successfully determined the DNA sequence of the human genome, the next step is to understand the roles of the different regions, especially of gene regulation. The transcription activity of a gene can be controlled by transcription factors (TF), proteins that bind to specific DNA sequences and either promotes or represses the transcription of the nearby DNA to mRNA. (1) The DNA sequences these transcription factors are likely to bind are called motifs, and the ENCODE consortium has listed some known motifs derived from experimental datasets.

With just the knowledge of the DNA sequence, it is not enough to predict which regulatory sites will be bound and the effects it has in overall gene regulation. Indeed our bodies are composed of many different cell types that have different gene expressions profiles according to cell type, but all stem from the same copy of our DNA. What biological factors can predict TF binding can allow us to better understand the basic biological mechanisms and perhaps make new discoveries. In addition, understanding how gene regulation changes in different cell types and especially in diseased cell states is essential.

Chromatin immunoprecipitation with massively parallel sequencing (ChIP-seq) is the standard assay for genome-wide detection of TF binding locations for a specific TF. However the limitations are that ChIP-seq can only detect one TF a time over one cell type when the reality is that hundreds of TFs are binding and regulating DNA transcription. Using machine learning methods to predict TF binding can potentially allow us to compare across numerous different TFs as well as across cell type. In addition to using ChIP-seq data, we incorporate other biologically relevant data such as DNase 1 sensitivity. Chromatin usually exists in the cell as complicated structures tightly bound with histones that prohibit transcription. DNase activity indicates whether it is an open region of chromatin that TFs can bind to. (2) We use machine learning with supervised learning and features extracted from ChIP-seq and DNase activity data to make predictions about TF activity in different cell types.

Methods

<u>Text extraction</u>

Specific base-pairs that are referenced in ChiP-seq peaks can be extracted using the hg19r eference genome. Our ChiP-seq peak data provides the location of the peak and a window of the 1001bp surrounding the peak. With the chromosome number and position, we can access the DNA sequence in FASTA format using the hg19 genome. (5) This format accounts for

possible variations in the DNA such as SNPs. Given the size of the reference genome, this must be done efficiently. Initially, we attempted loading the genome into memory. Because the genome is so large (> 3x109 base-pairs) we found that our development machines couldn't handle this. This led us to use random disk accesses to extract the sequences. This proves very expensive, but only needs to happen once for each unique peak location.

Motif finding

First, we extract the DNA sequence corresponding to high CTCF binding levels using the reference genome and ChiP-seq peak coordinates. Initially, we used all 532 ENCODE motifs, and their associated PSSMs (position-specific-scoring matrices) to find motif matches on the DNA sequence for each peak. These matches are only dependent on the DNA sequence. The PSSMs relate a notion of how likely a sequence is compatible with a molecule that binds to DNA. We used the log of this likelihood as a feature. On advice from our advisor, we used only the 239 features that have been experimentally determined to affect TF binding (the remaining 293 have been statistically determined) . This is also expensive: for each peak and motif, we need to calculate the PSSM score for all subsequences[ie s(1,6),s(2,7) ...].

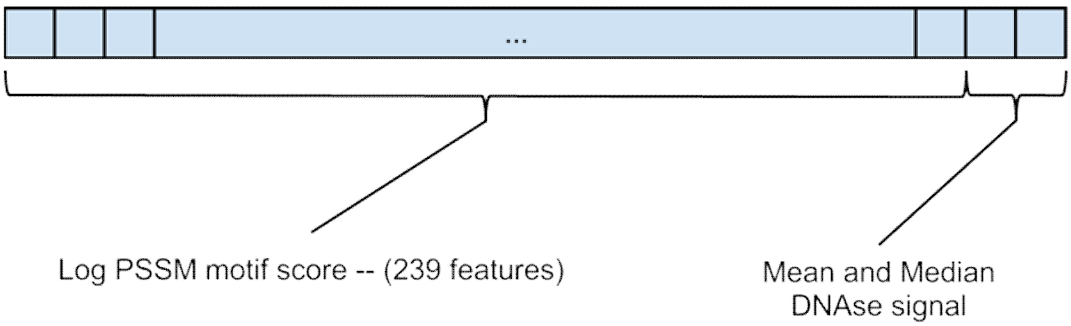
DNase and Histone modifications

Since DNase activity indicates if the chromatin is open (ie. transcription factors can bound in that region), high cumulative DNase levels should correlate with TF binding. In addition, the acetylation and methylation state of various histones are also an indicator of the shape of the DNA chromatin complex and if transcription factor binding is promoted or repressed. We used histones H3k27, H3k4, H3k9 and H3k36 in various states as features because there are prior studies indicating related activity and also because it is available. (7) DNase and histone data is also available in the form of subsamples over 10bp regions in the 1001bp region around peak CTCF TF binding. From this, we extract the maximum and median signal of each as features.

Machine Learning

Using 1000 positive sets of high ChIP-seq scoring peaks and 1000 negative sets of low ChIP-seq scoring peaks of the Gm12878 lymphoblastoid cell line from a donor, we trained both a linear SVM and a logistic binary classifier. The feature vector we used includes the maximum log probability for a set of motifs matching a given subsequence in the peak, the maximum and the median of the DNAse signal and chromatin protein signal. The linear SVM was implemented with the liblinear SVM software and logistic regression was implemented using the matlab native command glmval and glmfit.

We tested our model on 100 negative and 100 positive data points from the K562 cell line, from a leukemia patient.



Results

We found that our models performed relatively well. We first included only the motif likelihood and DNase signals as features. With Linear SVM, we obtained 81% test set accuracy. With the logistic regression model the test set accuracy was increased to 90%. This provides evidence that the model fits the data relatively well. The highest weighted features were the max and median DNAse score. This indicates that openness of the chromosome is one of the most important factors determining whether a transcription factor may bind.

We then implemented the same machine learning methods with the additional data from the various histone states that are known to be associated with transcription. With the inclusion of the new features, the training set accuracy did not differ much from the previous learned models, however the test set accuracy was dramatically improved with the linear SVM model producing a test set accuracy of 92% and the logistic regression model with a test set accuracy of 97%. It could be that the model learned worked particularly well with the K562 cell line we used as the test data.

In general the logistic regression model seems to be a better fit for predicting transcription factor binding. This is most likely due to the features contributing non-linearly to the binary classifier. We also implemented a radial basis kernel SVM method, however the accuracy was poor, only 67% even after cross-validation to find the best parameters, thus this model was excluded for our final evaluation.

We found that the motif scores with the greatest influence were ets_known8 and SP1. The presence of ets_known8 indicates that CTCF binding is unlikely whereas SP1 indicates that CTCF binding is likely.

% Accuracy	training data	test data
linear SVM	88%	88%
logistic regression	90%	81%

	precision	recall	F1 score
linear SVM	74.8%	94%	83.3
logistic regression	83.9%	94%	88.7

Including data from other histone factors:

% Accuracy	training data	test data
linear SVM	85%	92%
logistic regression	90%	97%

	precision	recall	F1 score
linear SVM	90.4%	94%	92.2
logistic regression	83.9%	100%	91.2

Conclusion

Chromatin immunoprecipitation with massively parallel sequencing (ChIP-seq) is the standard assay for genome-wide detection of TF binding locations for a specific TF. However the limitations are that ChIP-seq can only detect one TF a time over one cell type when the reality is that hundreds of TFs are binding and regulating DNA transcription. In addition to using ChIP-seq data, we incorporated other biologically relevant data such as DNase 1 sensitivity and histone modification, which indicates the availability of chromatin structures for transcription. Using linear SVM and logistic regression machine learning methods, we can predict TF binding in a different cell type with good accuracy. In particular, logistic regression was a good model for this biological application.

Future Research

In addition to the set of features outlined above, it may be worthwhile to consider more rich features. For example: we could include the relative position of Motifs with respect to the peak. Additionally, it would be worthwhile to consider different models for this problem. Furthermore we can check for the biological significance of our prediction algorithm, by comparing if the predicted positive and negative TF binding regions existed in the training cell data, and the relation between transcription of benign and cancerous cells.

Acknowledgments

Our advisor, Sofia Kyriazopoulou-Panagiotopoulou suggested the project. She was an immense help throughout the process: she provided general guidance throughout the process, supplied the datasets that were used in this project, and provided many references for background knowledge.

Works Cited

1. 1. Cheng C, Alexander R, Min R, et al 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Research 22:1658-1667
2. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U 2012. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Research 22:1711-1722
3. Thurman RE, et al 2012. The accessible chromatin landscape of the human genome. Nature 489: 75-82
4. Pique-Regi R, Degner JF et al 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21(3): 447–455.
5. single letter codes for nucleotides. http://www.ncbi.nlm.nih.gov/staff/tao/tools/tool_lettercode.html
6. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. Histone code. http://en.wikipedia.org/wiki/Histone_code