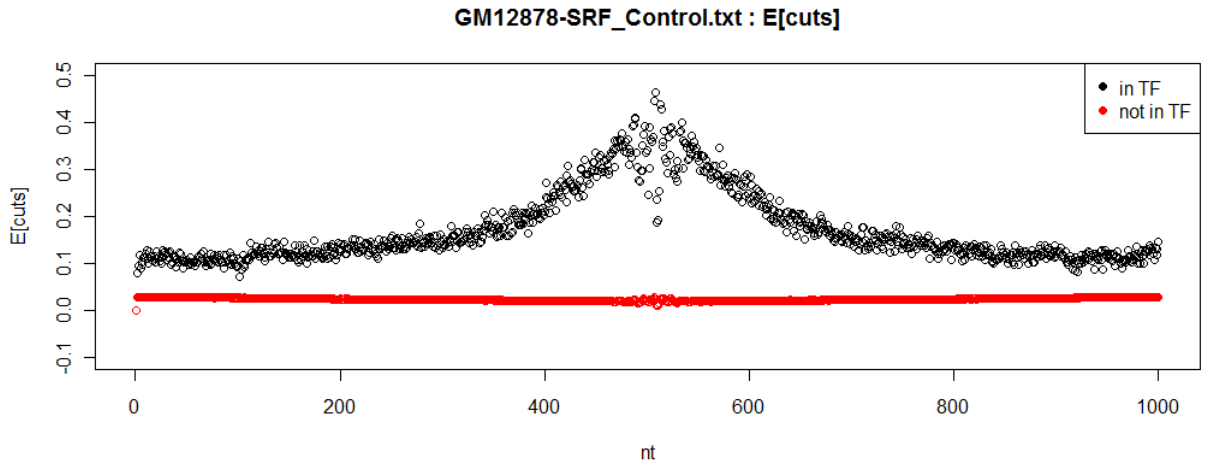
Shin Lin
CS229 Final Project
Identifying Transcription Factor Binding by the DNase Hypersensitivity Assay

BACKGROUND

In the DNA of cell nuclei, transcription factors (TF) bind regulatory regions throughout the genome in tissue specific patterns. The binding occurs for three reasons: 1) the TF is present, 2) a section of DNA is in an open conformation (so-called euchromatin), and 3) the sequence contains a TF-specific cognate sequence. To date, the most reliable method to assess if and where a TF is binding is to perform the chromatin immunoprecipitation (ChIP) assay. Briefly, TFs are cross-linked to their bound DNA, and the protein-nucleic acid complex is isolated by a particular TF-specific antibody. The DNA is then freed and sequenced (so-called ChIP-seq). The sequences may be mapped back to loci in the genome. In this way, the binding locations of one specific TF can be determined globally.

The DNase hypersensitivity assay (DHS) presents a potential method of querying all locations of all TFs in a single assay. Briefly, chromatin is treated with DNase I. Euchromatin regions are cut, and these cuts are then sequenced. Looking more closely at these spans of high cut regions, one sees short stretches notable for fewer cuts (viz. the footprint), which correspond to ~8-20 nucleotide (nt) sequences where TF binding protects the DNA from the DNase activity and usually harbor TF-specific cognate sequences.

Figure 1. Cut profiles using lymphoblastoid cell line GM12878 for TF SRF. The average cut number is plotted against nucleotide position with the motif in the middle. The profile corresponding to bound TF is in black; not bound TF, red. Note the divot or footprint in the middle of the bound TF profile.



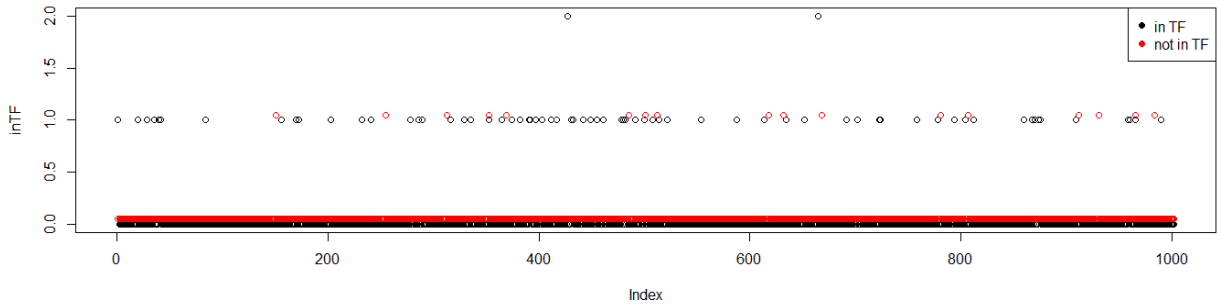
Concretely, the features are vectors X corresponding to sequence lengths several hundred nt long. Each position can be a non-negative integer, representing the number of cuts produced by the DNase I hypersensitivity assay. The target variable is whether a sequence harbors a bound TF (Y = 1) or not (Y = 0). An estimate of E[X|Y=1] (black) and E[X|Y=0] (red) is shown in Figure 1. Simply, the work described herein applies machine learning models to distinguish regions resembling the black profile from the red. The training and test cases are created using ChIP-seq results for specific TFs as the gold standard. The objective is to be able to recognize areas bound by TF with DHS results only, so the biologist may dispense with undertaking multiple ChIP-seqs .

To date, there have been two major approaches to handling this problem. Pritchard's group used what they termed a hierarchical mixture model[1]. It is an unsupervised approach analogous to the mixtures of Gaussians computed with the EM algorithm in class. The model distributions employed are different, however. Instead of multinomial priors for the latent variable, the prior is modeled with a Bernoulli

distribution whose parameter is derived from a logistic regression using the features TF motif similarity score, distance to transcription start site, and sequence conservation. The probability of the data given the latent variable is modeled by a negative binomial (for the number of cuts) multiplied by a multinomial (for the position of the cuts), instead of, as in the class example, Gaussian distributions.

Boyle and colleagues took another approach[2]. They used a hidden Markov model (HMM) to assess whether a region was likely to harbor a binding TF simply from the cut profile. The advantages of this method was that it requires only DNase cut sites and did not rely on other information such as TF motif similarity score, distance to transcription start site, and sequence conservation, which is extra information that must be inferred and is not without its own underlying uncertainty.

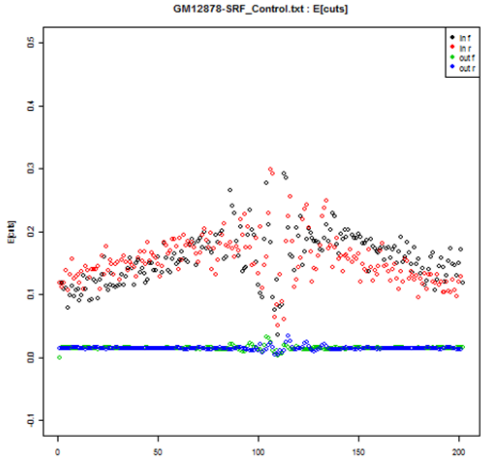
Figure 2. Representative $X^{(i)}$s. Number of cuts is plotted against nucleotide position with the cognate TF sequence centered. The $X^{(i)}$ with $Y^{(i)} = 1$ is in black; the other with $Y^{(i)} = 0$, red.



RESULTS

At first blush, the DNase cut profiles between TF-bound and not TF-bound seem distinct enough, and one may presume the problem to be easily solved. This unfortunately did not prove to be true. We first employed linear support vector machine (SVM) model[3] with rather poor results (precision 1.61%, recall 31%). L2-regularized linear logistic regression was no better (precision 1.11%, recall 12.5%). Finally, we tried another SVM model using the radial basis function as kernal[4], which was a better fit but still poor (precision 83%, recall 23%). The data used were DHS results for human embryonic stem cell H1 with TF Atf3 ChIP-seq data as the gold standard, all generated from the ENCODE Project (http://www.genome.gov/10005107).

Figure 3. Average Cut Profile after Features Modified. E[X|Y=1] for the forward strand is represented in black; E[X|Y=1] for the reverse strand, red; E[X|Y=0] for the forward strand, blue; E[X|Y=0] for the reverse strand, green.



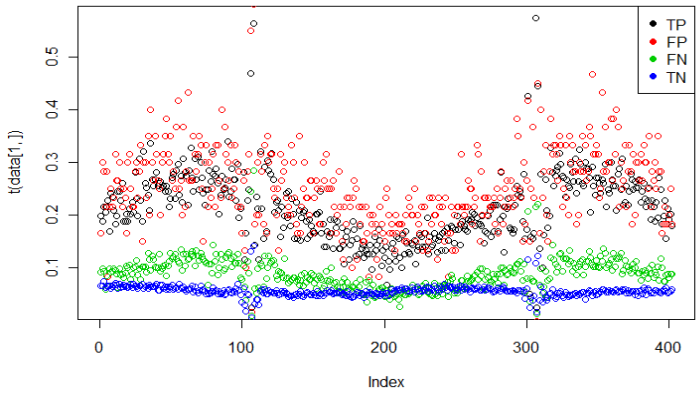
In the aforementioned computations, the error was unacceptably high in the context of using the training set as the test set. This result implied a problem of bias. We therefore sought to reformulate the features. A careful look at the features of individual data points suggested this was the right path to take. Although on average, the X(i)s between those with Y(i) = 1 and Y(i) = 0 are very different (Figure 1), on an individual case by case basis, the difference is not so easy to distinguish (Figure 2).

One of the steps Pique-Regi and colleagues performed in

processing their data was to distinguish between cuts of the forward strand from those of the reverse [1]. Also, it seemed that using cuts approximately 500nt to the left and right of the cognate TF motif was excessive, since the peripheral positions seemed to harbor few cuts. The range was reduced to 100nt to the left and right of the motif. The subsequent SVM fit with the radial basis function as kernal was improved with a better recall (precision 79%, recall 31%).

We next sought to understand the reason for the mistakes from this SVM model. Figure 4 shows a plot of the average cut profiles of the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In this representation, the values of the forward and reverse strands are placed end to end (i.e. the forward strand runs from positions 1 to 201; the reverse strand , 202 to 402). As can be seen, the SVM model does separate out the different cut profiles well in the sense that the TP and FP as well as the FN and TN appear similar. This plot suggests that the there may be issues with the training data, rather than with the model.

Figure 4. Average Cut Profiles from SVM Fit. Points are stratified as true positives (black), false positives (red), false negatives (green), and true negatives (blue).



Indeed, the gold standard may not be, in fact, perfect. The presence of a bound TF was inferred probabilistically from ChIP-seq experiments. Operationally, the TF peaks were called with QuEST[5], which has a fairly stringent threshold. Heretofore, a region with a specific motif was considered not to harbor a TF if QuEST did not call a ChIP-seq peak in the region. However, there could have been regions for which there was mild-moderate evidence for TF binding that QuEST discounted. These same regions could have had very strong evidence for TF binding by the DHS assay. Our method would leave these potentially mislabeled points in the training set.

To address this issue, the training set was reconstituted with stricter criteria for designating a region not TF-bound. Briefly, the area had to have had significantly more control sequences over TF sequences to be considered a negative data point (for more on how ChIP-seq is analyzed, please refer to Valouev et al.[5]). After fitting the SVM model, we achieved precision of 100% and recall of 77%. These numbers were calculated using 10-fold cross validation.

A natural question is how these methods compare with the work of Pique-Regi et al.[1] and Boyle et al.[2]. The latter framed their validation in a subtly more nuanced manner and will not be considered in this work. The former used other information in addition the DHS cuts (viz. TF motif similarity score, distance to transcription start site, and sequence conservation), and moreover the group's data sets were not available to us.

To directly compare our SVM model with the hierarchical mixture model of Pique-Regi et al. [1], we analyzed the DHS data of lymphoblastoid line GM12878 and the TF CTCF ChIP-seq data. With the SVM model with the radial basis function kernal fitted by 10 fold cross validation, we achieved a precision of 96% and recall of 89%. This compares favorably with Pique-Regi and colleagues' reported results of 98.47% and 87.56% for the same measures, respectively. Of note, our results are no worse even

though the latter used other biological information in making final calls. The comparison must be viewed with a jaundiced eye, however, because we did not use precisely the same validation set as they did. Even though we started with the same raw sequence files, the actual data points abstracted from these sources were different.

To make an even more direct comparison, we ran Pique-Regi et al.'s hierarchical model implemented in R[6] package CENTIPEDE[1] on our CTCF data. We used only the GM12878 DHS cuts as input. The model achieved a precision of 92% and recall of 77%. The decrement in accuracy may be attributed to not using extra-DHS information, increased difficulty of our data set, and/or the lesser discriminatory power of their hierarchical model vis-a-vis an SVM one.

DISCUSSION

We have shown that SVMs can be a powerful tool to distinguish between regions with TF binding and non-binding with the DHS assay. Further work must be done to compare the method with others for many more transcription factors beyond CTCF. Another avenue is to see if the same SVM fit can be used across multiple (or maybe all TFs). The concept of Boyle et al.'s HMM sought to achieve this objective; Pique-Regi et al.'s model required a different model to be fit for every TF.

This last point has important implications for how DHS data can be used to supplant an exhaustive set of ChIP-seq TF experiments to elucidate the cellular regulome. The promise of DHS data was not only to define the positions of known TFs, but to allow identification of new ones, which would only be known by their sequence motifs. Since the hierarchical model must be fit for every new TF, the identification of new TF positions is not as direct.

It is by no means impossible. Pique-Regi et al. were able to achieve this goal by looking for sequence motifs not previously described across DHS cut enriched areas. Having identified common motifs, they fit their hierarchical model for each set of regions with a common, novel motif.

This is in contradistinction to what Boyle et al.'s approach can possibly do (and what the SVM can potentially do better). A single model may be run across the DHS cuts across the genome. Regions marked highly probability of bound TF are marked. The search for common motifs can then be relegated to a much smaller search space. The material difference of these two approaches is that this latter method can identify motifs of low abundance across the genome. Further work will be required to realize this goal.

References

1. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*21(3):447-455.

2. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*21(3):456-464.

3. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research.* 2008;9:1871-1874.

4. Chang C, Lin C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.* 2011;2(3):1-27.

5. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods.* 2008;5(9):829-834.
6. *R: A Language and Environment for Statistical Computing* [computer program]. Version; 2011.