Title: Genome-Wide Predictions of Transcription Factor Binding Events using Multi-Dimensional Genomic and Epigenomic Features

Team members: David Moskowitz and Emily Tsang

Background

Transcription factors (TFs) regulate gene expression by binding to specific sequences of DNA. This drives disease progression and differences between cell types. Biologists can elucidate TF binding sites using a technique called ChIP-seq. Although sensitive and specific, ChIP-seq can only provide information for a single TF per experiment. Because there are hundreds of TFs expressed in each cell, compiling an exhaustive set of TF binding sites for even a single tissue is prohibitively expensive and time-consuming. Here, we discuss applying machine learning to an expansive set of genomic features to predict genome-wide binding, simultaneously for all TF of interest, in a cell type for which ChIP-seq data may be unavailable.

Many of our features are from sequencing-based assays that probe different characteristics of the cell. Sequencing of DNA or RNA gives hundreds of millions of short sequence strings. A typical workflow involves aligning these strings to their original location in the genome and further steps specific to the particular assay. For example, DNase-seq probes DNA openness by only sequencing DNA that is accessible and can be cleaved into short reads. Most DNA is tightly coiled, preventing proteins like TFs from binding there. We are therefore restricting our analysis to open sites, which constitute ~1-5% of the genome in our cell lines of interest.

We used publicly available data to build our feature set and training examples for several different cell lines. As our use case, we assume that a given cell type has DNA accessibility information, but no ChIP-seq data. We want to use the data compiled for other cell lines to predict binding events in the tissue of interest. To simulate this situation, we evaluated different multi-class classification methods on this data set via leave-one-out cross-validation with respect to cell line. With small adjustments, we get average cross validation accuracies of ~90%. We then explored using a smaller or less comprehensive training set, and smaller feature sets. Finally, we simulated prediction for a factor for which there are no training data to see whether our approach can learn general features of TF binding.

Labels and Features

Most experimental data we used were from the ENCODE project[1]. This included assays of several cell-specific features such as RNA-seq, ChIP-seq of both activating and repressing histone modifications, ChIP-seq for several TFs, and DNase-seq. We also built features from cell-type-independent data such as genomic conservation, HT-SELEX scores, and distance to the nearest gene. We chose five cell lines for which all the data listed above were available: GM12878, H1 hESC, HeLa-S3, HepG2, K562. We decided it would be best to restrict the TFs for which we were making predictions to ones that had HT-SELEX scores as well as ChIP-seq data for all five cell lines. This left us with five transcription factors: CTCF, MAFK, MAX, NRF1, and RFX5.

We combined information from ChIP-seq, DNase-seq, and HT-SELEX to define the class labels. ChIP-seq reveals where specific proteins, such as TFs, are bound by selectively sequencing the DNA with which they interact. The assayed protein is putatively bound within regions enriched for ChIP-seq reads. These regions are broad (~200 base pairs each), so we restricted them in two ways. This is needed because the majority of sites in those regions are actually unbound and thus the features therein would not be characteristic of TF binding. First, we postulated that TFs cannot interact with highly condensed DNA, so we intersected the ChIP-seq with the DNase-seq peaks, which mark accessible DNA regions. We then use HT-SELEX scores to pinpoint the most likely binding site. HT-SELEX[2] is a technique used to find motifs to which TFs bind. We devised a means

by which this can be translated into the likelihood with which a TF will bind at a given locus: $\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{4}\log\left(C_{ij}\right)1\{b_j=g_i\}\right)-\sum_{j=1}^{4}\log\left(C_{ij}\right)\right)$, where n is the motif length, C_{ij} is the count (in the HT-SELEX data) of base j (in {A,C,G,T}) at the i^{th} position in the motif, and the indicator function is 1 if the genomic base g_i is equal to the current base. The site with the highest HT-SELEX score within each restricted ChIP-seq peak is labeled as bound by the TF.

The number of training examples for each TF was 102,986; 68,469; 136,817; 20,873; and 72,632 for CTCF, MAFK, MAX, NRF1, and RFX5, respectively. The number of bound instances per cell line was 67,467; 59,010; 117,460; 72,399; and 85,441 for GM12878, H1 hESC, HeLa-S3, HepG2, and K562, respectively. For each cell line, we randomly selected an equal number of unbound sites within DNase regions as negative examples, maintaining a balanced class ratio to diminish classification biases.

When building our feature vector, we used the idea that a small number of bases on either side of a TF binding site are informative. Therefore, we created a 101-base-pair (bp) window centered on the current prediction site. For each site within that window, we include the phyloP score for that site as well as indicators for activating and repressing histone marks. phyloP[3] is a metric of genetic conservation between species. The idea is that more highly conserved sites are likely functional. On the other hand, histone modifications hint at the activity level of a site. Histones are proteins around which the DNA wraps and the location of those with particular modifications can be identified by ChIP-seq. We intersected the ChIP-seq peaks for six marks of active sites to get an indicator function for activating histone marks at each site. We use the peaks of the single repressing histone mark assayed to define the second indicator function.

In addition to the features in the window mentioned above, we incorporate the distance to the nearest gene, the expression level of that gene and the HT-SELEX scores of the five TFs of interest at the prediction site. We determine the location of the nearest gene using the GENCODE gene annotations[4] and calculate its expression level based on the RNA-seq data. To get the expression level of each gene, we first counted the number of RNA-seq reads that overlap gene annotations using HTSeq[5], then normalized these raw counts with DESeq[6], which accounts for variability in the number of reads in the different cell types and individual gene lengths.

With conservation, and two indicators for histone modifications at each position in a window of 101 base pairs, five HT-SELEX scores, and two gene-related features, we created a total of 310 features per prediction site.

Methods and Results

Classification using entire training set

After the creation of the feature vector and the attribution of class labels, we investigated several methods for predicting the TF binding events. Using liblinear[7], we applied an SVM with a linear kernel using L^2 regularization, the same type of SVM on scaled data, multinomial logistic regression, and multinomial logistic regression using L^1 regularization. Additionally, we evaluated the robustness of our model by down-sampling our training set and by performing forward selection on our features. Finally, we quantified cell-specific and transcription-factor-specific effects by repeating the classification on different training and test sets.

In our first attempt at classification, we applied liblinear to train a linear-kernel SVM with L^2 regularization. With five-fold cross-validation, leaving out one cell line in each iteration, we obtained an average accuracy of just over 50% (Table 1, column 1). When we visualize the accuracy, distinguishing between classes, we find that the SVM is performing reasonably on unbound cases, but only modestly for each TF (Fig. 1A). However, we still viewed this as promising, since, in six-class classification, this performance is significantly above random.

In evaluating our method, we realized that our features were drawn from widely varying distributions. We became concerned that those with greater ranges, such as gene expression, might be overpowering the signals from features with more confined ranges, such as genetic conservation. Thus, we decided to scale each feature independently to a range of [-1, 1]. We then reran the SVM and found our accuracy had improved to ~90% for all cell lines (Table 1, column 2). Visualizing these results, we find that most errors arose from misclassifying bound versus unbound, and that bound classes were rarely confused (Fig. 1B). Because the scaled data set produced such a marked improvement over the unscaled, all subsequent analyses were also performed on the scaled data.

We attempted to further increase our accuracy by empirically searching different penalty parameters in the error term of the SVM optimization problem. However, we found that this increased our accuracy by only ~0.1%, while increasing our run-time by ~30x, and thus decided to abandon this approach.

However, we were still lacking a quantitative assessment of our confidence that each locus belonged to a particular class. By instead running multinomial logistic regression, which offered a slight increase in accuracy (Table 1, column 3), we found that the distributions of probabilities varied greatly between correct and incorrect classifications (Fig. 1C). Specifically, when an instance is correctly classified, the associated probability is generally high (Fig. 1C, black curve). When an instance is misclassified, the confidence is substantially lower (Fig. 1C, blue curve). Indeed, for misclassified instances, the probability given to the correct class is often nearly as high as that given to the assigned class (Fig. 1C, red curve). This implies that with minor improvements we could further boost our accuracy, since the classifier, even when mistaken, is typically already determining that the actual class is almost as likely as the predicted class.

Finally, we reapplied multinomial logistic regression, this time with L^1 regularization, to test whether any of our features were superfluous. Overall, this produced a slight dip in accuracy (Table 2, column 4). In none of the five cross-validation folds was more than a single feature given zero weight, indicating that the features we used are informative with respect to TF binding.

Model and feature evaluation

Having verified via L^1 regularization that our features were informative, we evaluated their combinatoric performance, as measured through forward selection. Because we are performing multinomial classification, and the HT-SELEX scores are the only TF-specific features, we first included the HT-SELEX scores as the baseline. With the HT-SELEX scores alone, we achieved an average accuracy of 83% (Fig. 2B). We then tried running the model with each feature in combination with the HT-SELEX scores and found that adding the expression of the nearest gene improved the mean cross-validation accuracy the most, bringing it to over 91%. Subsequently adding activating histone marks improved the average accuracy slightly to 92%, but further feature addition after this point resulted in very little gain in prediction accuracy. Interestingly, using only HT-SELEX scores with distance to the nearest gene gave an average accuracy of 91%. This is exciting because it means that with a single experiment per cell line, DNase-seq, we can get reasonable predictions. Therefore, we can apply our method to the multitude of cell lines which may have DNase data, but lack the other experimental data types we also used as features.

After establishing the minimal set of features required for accurate classification, we investigated the lower bound on the number of training examples needed, through down-sampling. On average, a TF will have tens of thousands of bound sites per cell type, offering a large pool of examples on which to train. However, lowly expressed TFs or those with unusual motifs may only have hundreds of bound sites in a given tissue. We found that we reached saturation after including only ~5,000 training examples per TF, far fewer than is typically available (Fig. 2A). To our pleasant surprise, we observed that, even with only a few hundred examples for each TF, accuracy

suffered a drop of only 5-10%. This demonstrates that, even when a ChIP-seq experiment produces only a small number of peaks for a particular TF, computational inference is still effective.

Extension of classification to under-sampled TFs

Following our determination that TF binding can still be effectively predicted with orders of magnitude fewer training examples than present within our data set, we sought to answer whether we could also decrease the number of cell lines in the training set without substantially diminishing performance. Specifically, we reassessed the performance of the SVM after training on a single cell line. We found that this impacted accuracy, on average, by only a few percent (Table 2). Because there are many TFs that have not had ChIP-seq experiments performed on them in multiple tissues, this represents a common use case for TF binding prediction. Our results indicate that the predictions generated are valuable even for under-sampled TFs.

Next, we evaluated the predictions on TFs for which no ChIP-seq data exist. This, unfortunately, is an extremely frequent circumstance. As a proxy, we removed from our training set all examples of CTCF. Then, for the remaining bound instances, we removed the HT-SELEX scores for the four TFs that were not bound at that position; at unbound positions, we set this score to 0. We next composed a test set solely from CTCF-bound and unbound loci, and included HT-SELEX scores for only CTCF. On this training and test set, we applied standard logistic regression. This was intended to replicate the scenario where, lacking training data for a particular TF, we instead use a general set of training examples comprised of other TFs' bound loci. If the features that enable binding prediction are relatively consistent across TFs, this workflow would be expected to be effective. The ROCs for the novel-factor classification (Figure 3) had AUCs of 0.99832, 0.99751, 0.99796, 0.99812, 0.99820, for GM12878, H1 hESC, HepG2, HeLa-S3, and K562, respectively. Interestingly, in all five cases, every instance misclassified by the algorithm is a false positive; the regression never assigns greater than 50% probability of being unbound to a bound position. Overall, this suggests that existing data might be sufficient for general TF binding prediction, even for TFs that have not yet been characterized directly through ChIP-seq.

Conclusions and Future Directions

Our goal was to improve on the state-of-the-art in TF binding prediction by leveraging publicly available data from the ENCODE project and other sources. Using multi-class SVM and logistic regression on scaled features, we offer a solution that can generate predictions genome-wide simultaneously for all TFs. We incorporated 310 features, and were able to obtain classification accuracies of ~90% for 5 TFs, in 5 cell lines, across 803,554 sites. We further established that having only one cell line for training does not significantly reduce the power of our inference, and that binding sites for novel TFs can also be predicted extremely effectively.

In the immediate future, we plan to test the predictive power of other features, including methylation (which we excluded due to large amounts of missing data) and DNase-seq read distributions. We also aim to expand the number of classifiers applied so we can assess the relative advantages and disadvantages of each. We attempted to run both the glmnet and randomForest R packages, but terminated them after they had run for several hours without finishing. Because we have determined that down-sampling the number of training examples by two orders of magnitude still produces comparable results, we will re-evaluate these methods on data sets of reduced size. We are particularly interested to see whether the feature weights produced by lasso via glmnet are in concordance with the weights given by liblinear or our results from forward selection.

It is our hope that, with minor adjustments to our feature vector and algorithm, we can achieve near-perfect classification of TF binding; we can then apply our method to unexplored tissue types to offer novel insight into the cell's underlying regulatory processes.

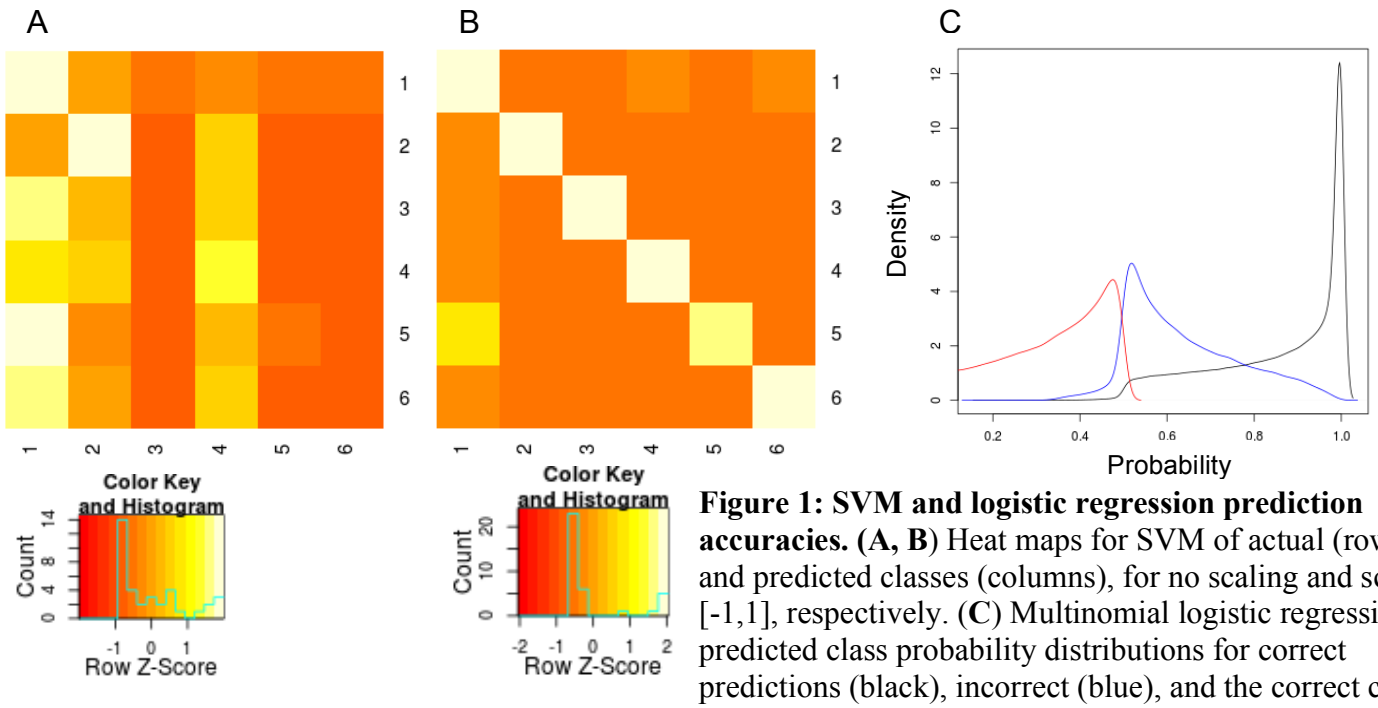

Figure 1: SVM and logistic regression prediction accuracies. (A, B) Heat maps for SVM of actual (rows) and predicted classes (columns), for no scaling and scaling [-1,1], respectively. (**C**) Multinomial logistic regression predicted class probability distributions for correct predictions (black), incorrect (blue), and the correct class when an incorrect prediction was made (red).

	No scaling	Scaling [-1,1]	Logistic regression	Logistic, L^1 regularization
GM12878	52.853	89.064	**90.439**	89.648
H1 hESC	55.194	92.443	**93.085**	91.741
HeLa-S3	53.245	92.422	**93.124**	92.291
HepG2	56.510	92.753	**93.364**	92.123
K562	56.940	92.312	**93.274**	91.664

Table 1: Performance of various classifiers broken down by cell line. Five-fold cross-validation accuracies for: SVM with no feature scaling, SVM with features scaled [-1,1], multinomial logistic regression with L^2 regularization, and multinomial logistic regression with L^1 regularization.

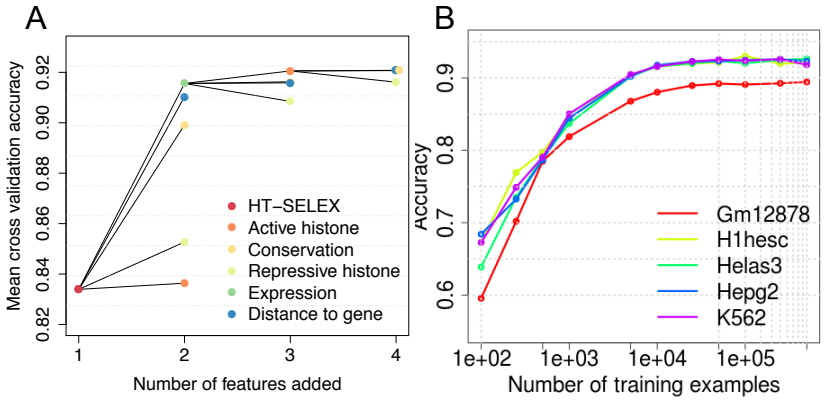

Figure 2: SVM performance with fewer training examples or features. (A) Depiction of forward selection process starting with inclusion of HT-SELEX scores. (**B**) Cross validation accuracies given increasing training set size.

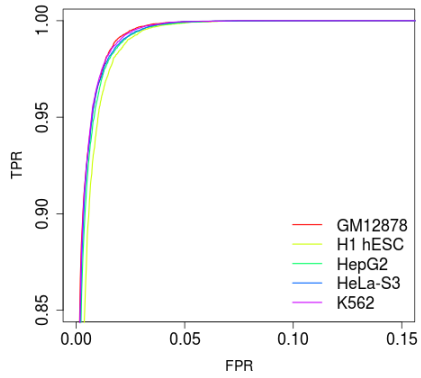

Figure 3: Predictions for novel TF. ROC curves for prediction of CTCF binding sites using logistic regression, without training on CTCF examples. Details of procedure in text.

	GM12878	H1 hESC	HeLa-S3	HepG2	K562	Overall
GM12878		88.866	85.198	85.088	81.849	84.966
H1 hESC	89.236		88.681	89.511	91.986	89.789
HeLa-S3	87.066	89.905		92.350	91.325	90.281
HepG2	87.638	91.066	93.130		92.375	91.439
K562	87.924	92.303	90.536	91.022		90.420

Table 2: Performance of SVM trained on a single cell line. Accuracies for training on one cell line (rows) and testing on others (columns).

[1] ENCODE Project Consortium *et al.* (2012) Nature. **[2]** Jolma, A., *et al.* (2013) Cell. **[3]** Siepel A. *et al.* (2006) RECOMB. **[4]** Harrow, J. *et al.* (2012) Genome research. **[5]** http://www-huber.embl.de/users/anders/HTSeq/doc /overview.html **[6]** Simon Anders and Wolfgang Huber (2010) Genome Biology. **[7]** Fan, R.-E. *et al.* (2008) Journal of Machine Learning Research.