TF-Finder: Predicting the Underlying Transcription Factors in Genomic Sequencing Data

Diwakar Ganesan
Stanford University
cganesan@stanford.edu

David Wu
Stanford University
dwwu16@stanford.edu

Abstract

In this paper, we try to predict the transcription factors (TFs) that are associated with a given set of genomic regions. Becuase TF misregulation is a major cause of disease, this work has serious implications for disease research and drug development. We used a well-established algorithm, PRISM, to predict binding sites for 974 human TFs. Then, we ran two sets of experiments on this computationally derived binding site data. Our first set of experiments mapped binding sites to ontologies using an algorithm called GREAT. There were roughly 9000 ontology terms in the ontologies we used, with each term linked to a TF with a p-value; thus yielding a high dimensional matrix of summary statistics. Second, we binned the human genome into 1 kilobase windows, and mapped our PRISM predicted binding sites to these bins to yield a matrix of binding site counts. Then, we ran PCA and NMF on these large matrices to predict what TFs are associated with unseen sets of binding sites. To test our low rank approximations, we created a test set of 49 TF binding site sets empirically determined to belong to a single TF. We projected these binding sites into a lower dimension using PCA/NMF, and used an L2 norm metric to determine which TFs in our training matrix the new lower-dimensional approximation was closest to. This yielded a rankings of likely TFs. Our best performing experiment with the ontology terms representation was using NMF, yielding a Top 10 accuracy rate of 22%. The binned genomic data performed much better with NMF, with a Top 10 accuracy rate of 65%.

1 Introduction and Motivation

The human genome encodes 20,000 proteins involved in many biological processes, such as cell proliferation and metabolism. However, not all of these 20,000 proteins are present in each cell of the body. Muscle cells only express proteins that build up muscle fibers, and brain cells only express proteins that help transmit electrical impulses from neuron to neuron. Transcription factors (TFs) are a 2,000 member subclass of proteins that are critical in the regulation of how proteins are expressed throughout the body. They are spatial-temporal switches, binding to specific DNA sequences to either turn on/off transcription of the protein coded at that particular genomic location. Misregulation of transcription factors can lead to devastating diseases such as cancer and heart disease (1). Specifically, each TF contributes to the expression of one or more normal phenotypes, but when misregulated, these phenotypes can be altered to cause disease. For example, the TF TP53 normally detects and repairs DNA damage. However, misregulation of TP53 can inactivate its protective function, leading to aberrant cell division, which is one of the phenotypes that precedes cancer (2). Because transcription factors play such an important role in disease, it is critical to develop tools that can help understand transcription factor function.

Recent developments in sequencing technology, specifically, ATAC-seq and ChIP-seq, have allowed scientists to quickly and cheaply determine the overall transcription factor binding profile of

.

a sample. For example, a tissue sample from a liver cancer patient can be analyzed by ATAC-seq to determine each location in the genome that is being stimulated by transcription factors. Additionally, with this set of genomic regions, a tool called Genomic Regions Enrichment of Annotations Tool (GREAT) can be used to infer the phenotypes that are most likely associated with TF binding to the regions identified by the aforementioned ATAC-seq or ChIP-seq (3). For example, ATAC-seq data from a diabetic patient fed into GREAT outputs enriched terms such as "Abnormal Carbohydrate Metabolism" and "Vascular Damage", phenotypes associated with diabetes.

Currently, one missing link in the field of transcription factor biology is that there is no way to efficiently determine which specific TFs are binding to the set of genomic regions given the set of genomic regions only. While wet-lab techniques can potentially be used for this purpose, they are often extremely expensive and time consuming, which make them impractical in a large number of cases. Our goal is to bridge this gap using techniques from this course. We seek to utilize the large amount of publicly available TF binding data to predict the specific TFs that are linked to a set of genomic regions. This way, scientists studying a particular disease can focus development of potential treatments on targeting the TFs that are most likely involved in the disease.

2 Methods

2.1 Inputs and Outputs

Our computational pipeline will take in two inputs:

(1) A set of genomic regions. Our pipeline will calculate the TFs that are most likely involved in this set of genomic regions.

(2) Computationally predicted sets of genomic regions that are bound by individual TFs, to use as a reference set we derive our rankings from.

(1) is user provided. To generate (2), we scraped TF binding preference data for 974 TFs from several major TF databases, namely, JASPAR, HOCOMOCO, TRANSFAC, and UniProbe (4)(5)(6)(7). These binding preferences are represented as position weight matricies (PWMs), which are 4 by n matricies, with rows representing each of the 4 nucleotides that make up DNA, and n columns, representing the length of DNA that the TF binds to. Each entry A_{ij} represents the probability that the TF will bind to the ith nucleotide at the jth position (Figure 1). We then generate the set of predicted TF binding sites in the genome by feeding our PWMs into PRISM (Predicting Regulatory Information from Single Motifs), a computational tool that predicts TF binding sites from PWMs (8).

$$\begin{array}{cc}
\begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{matrix} & \\
\begin{bmatrix}
0.97 & 0.10 & 0.02 & 0.03 & 0.10 & 0.01 & 0.05 & 0.85 & 0.03 \\
0.01 & 0.40 & 0.01 & 0.04 & 0.05 & 0.01 & 0.05 & 0.05 & 0.03 \\
0.01 & 0.40 & 0.95 & 0.03 & 0.40 & 0.01 & 0.3 & 0.05 & 0.03 \\
0.01 & 0.10 & 0.02 & 0.90 & 0.45 & 0.97 & 0.6 & 0.05 & 0.91
\end{bmatrix} & \begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix}
\end{array}$$

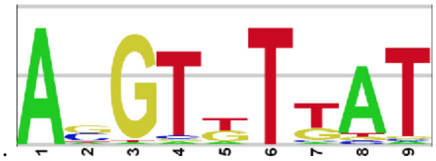

Figure 1: **Top** A PWM **Bottom** A visualization of the PWM, where the height of each base represents the preference of the TF for that base at the specified position in the binding site. Note that the y-axis is information content (a common metric to visualize PWMs), not probability.

Given the inputs, our pipeline will output a ranked set of TFs, ranked by the "likelihood" of a particular TF being associated with the set of genomic regions provided by the user.

2.2 TF-Finder Pipeline

One technique in computational biology that shows promise in discovering knowledge is the development of clever representations of raw experimental data. Raw data is first "encoded" into the representation, and analysis is done on the transformed data. For example, Tanigawa et al. used a GREAT summary statistic matrix derived from a dataset of genomic variants from the UK Biobank and principal components analysis (PCA) to discover new determinants of obesity (9). We will follow a similar procedure as Tanigawa et al. for our project, transforming our input data into one of two representations.

(1) The first encoding maps genomic regions to the space of ontology terms. We feed PRISM predicted genomic binding sites for our 974 TFs into GREAT to get summary statistics. GREAT currently supports 10 different ontologies, and for each ontology term, calculates 8 different statistics (such as p-value) on the term. With these calculations we create a GREAT summary statistic matrix with columns representing a particular TF in our PRISM dataset and rows representing an ontology term for the ontology we use. An entry of the matrix is the GREAT statistic for the corresponding TF-ontology term pair. We filter the summary statistics to encourage sparsity. Namely, any statistic corresponding to an ontology term that has a p-value greater than $.05$ is discarded, since a high p-value indicates it is unlikely the term is associated with the given region.

(2) The second encoding method involves binning the genome into windows of 1000 base pairs (a common length used in computational biology), resulting in 958,681 windows total. For each of our 974 PRISM-predicted TF binding sites, we record the number of predicted binding sites that falls within each window. We then create a matrix with columns representing a particular TF and rows representing genomic bins. Entries in the matrix will represent the number of binding sites in a bin for the corresponding combination of PRISM TF and genomic bin.

Our goal is to find a special structure in these matricies to predict potential TFs associated with a new set of genomic regions. To do this, we use two dimensionality reduction techniques, principal components analysis (PCA) and nonnegative matrix factorization (NMF). PCA was studied at length in CS229 and needs no introduction. On the other hand, NMF attempts to decompose a nonnegative matrix into two lower-rank matrices, each of which are also nonnegative. As such, the goal of NMF is to minimize the following loss:

$$\min_{W \in \mathbb{R}^{n \times r}, H \in \mathbb{R}^{r \times p}} ||A - WH||_F$$

subject to the constraint that $W \geq 0$ and $H \geq 0$. In practice NMF is implemented as a coordinate descent, since the objective is convex w.r.t W and H separately, and the coordinate descent at each step reduces to solving a nonnegative least-squares problem (a well known convex program).

We apply either PCA or NMF to both of our matrices to get a mapping from the original matrix space to that of a lower-dimensional space. To get the final output TF rankings, we project both our set of encoded 974 PRISM TFs and the encoded input set of genomic regions into the aforementioned low-dimensional space, and take the normalized inner product (otherwise known as the cosine similarity) between each of our projected 974 TFs and the input set. Mathematically, given two vectors $A, B \in \mathbb{R}^n$, the normalized inner product is defined as

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

and is a measure of how similar two vectors are in a space. Finally, TFs from our set of 974 are ranked based on this normalized inner product.

For evaluation, we scraped the National Center for Biotechnology Information (NCBI) GEO database, which contains empirically determined genomic binding sites for a collection of TFs (10). We then manually curated the data to include at most one set of genomic regions per TF, and only sets that had at least 1 significant GREAT-enriched phenotype term. Lastly, we threw out data from samples that were artificially altered (for example, from transgenic cells, or cells grown in abnormal conditions that were not representative of physiological conditions). Manual curation of GEO data resulted in 49 sets of TF binding site predictions. To measure our accuracy, we generated a ranked list of TFs for each of the 49 sets, and analyzed the rank of the ground truth TF for each of our 49 files.

3 Results

Our first set of experiments used GREAT summary statistics to represent PRISM binding site data. We ran our dimensionality reduction algorithm on all the different summary statistic matrices across all possible ontologies. The two results we collected were top-k accuracy and median rank, for $k = 10, 25, 50$. We could not create a sensible confusion matrix, because we didn't have enough pieces of test data (49 testing TFs vs 974 possible TFs). Other metrics like top-k precision and top-k recall were difficult to assess, since there was no way to determine the relevance of the other rankings in our predicted top-k ordering. Median rank is useful to judge how effective our ranking is at bubbling the correct TF to the top of the list. We achieved the best accuracy on the GOBiologicalProcess ontology using NMF. A table containing our interesting results is below. We reported the summary statistics that gave the best overall performance. The Binomial p-value statistic computes the likelihood a ontology term is associated with a particular set of genomic regions assuming that occurrences of genomic regions comes from a binomial distribution. The hypergeometric p-value arises from assuming that occurrences of genomic regions comes from a hypergeometric distribution. The binomial fold (BFold) statistic represents the ratio of the observed correlation between a TF-ontology term pair and the expected correlation. Our baseline experiments ranked TFs using the same normalized inner product metric, but with no dimensionality reduction.

GOBiologicalProcess Ontology Results					
Statistic used	Dim. Reduction	Top 10	Top 25	Top 50	Median rank
Binomial P-value	Baseline PCA NMF	.204 0.102 0.224	0.244 0.143 0.346	0.367 0.204 0.346	110 207 131
Hypergeometric P-value	Baseline PCA NMF	0.142 0.0612 0.082	0.224 0.122 0.184	0.285 0.184 0.29	238 301 375
BFold Statistic	Baseline PCA NMF	0.102 0.081 0.081	0.183 0.122 0.081	0.368 0.143 0.122	112 270 223

To better visualize our system's performance, the histogram below displays the predicted rank of the true TF for each element in our test set. Ideally, the predicted rank is always 1, so histograms that are clustered near the left side of the chart indicate better performance.

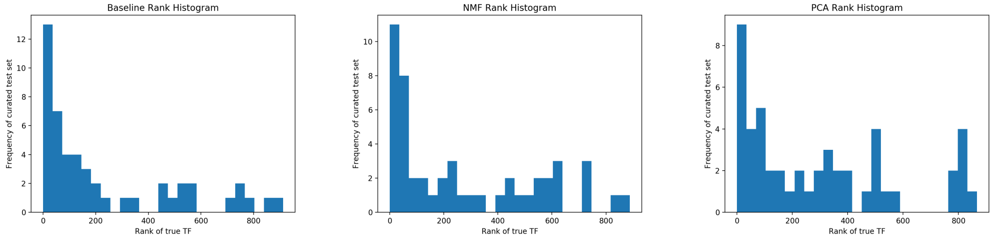

Figure 2: Rank prediction using GREAT summary statistics

The second set of experiments we ran used the genomic binning encoding matrix. A table that summarizes the results from these experiments, along with the corresponding rank histograms is presented below:

1kb Genomic Bin Results				
Dim. Reduction	Top 10	Top 25	Top 50	Median rank
Baseline	.632	0.710	0.816	3
PCA	0.18	0.51	0.69	24
NMF	0.652	0.70	0.81	7

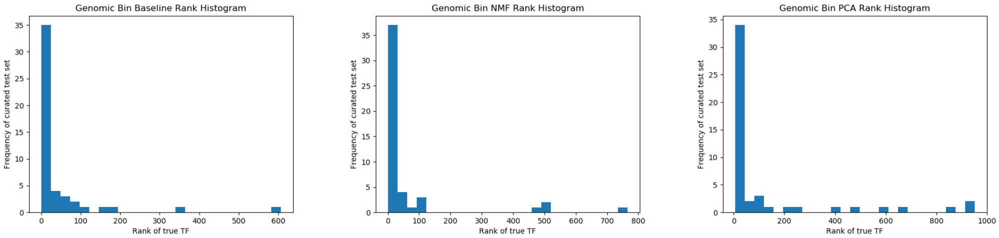

Figure 3: Mean rank using binned genomic data

4 Discussion

Our results give rise to several important points of discussion. First, we see that the genomic bin encoding yielded much better results than that of the ontology term encoding. The binned genomic data had better median-rank scores, higher accuracy, and the histograms are clearly more clustered towards the lower rankings. This implies that GREAT summary statistics are not a good representation of genomic regions, and that binning is a better way to represent the human genome than through a set of ontology terms. We also notice that PCA was consistently worse than the baseline experiments with no dimensionality reduction. This implies that the principle components of the high dimensional matrix we were analyzing do not tell us much about the information at hand. Nonnegative matrix factorization yielded a much better representation of the matrices than PCA, which is to be expected. Creating low dimensional representations of data works best when one makes use of all the prior knowledge about the structure of the data. Nonnegativity is one of the simplest invariants in our data – both GREAT summary statistics and counts of binding frequency are never negative. Exploiting this simple structure gives us a significant performance boost.

We'd also like to point out how good the baseline results were relative to the PCA/NMF results. Baseline experiments outperformed PCA on almost all counts, and was better than NMF in median rank. NMF yielded better accuracy in the most important category, top 10, but only by a small margin. This was an unexpected result. Usually, notions of distance are distorted in high dimensions, which is why problems like face verification often use algorithms like PCA as a preprocessing step. This likely means that representing genomes as ontologies or bins of binding sites actually does preserve notions of distance in some sense, even in the higher dimensional space.

5 Future Work

There are several avenues of future work our project suggests. First, we'd like to experiment more with GREAT summary statistics. We created statistic matrices individually, but we'd like to devise some way to incorporate all statistics at once into the matrix, since each GREAT statistic captures a different concept of association between genomic regions and ontology terms. Finding special structure in this higher-order matrix may yield more useful information about transcription factor binding sites.

Additionally, the two encoding models used in this paper are quite simple, and do not take into account full knowledge of TF dynamics. For example, TF binding is not equally distributed across the genome. Furthermore, it is known that TFs form multi-member complexes with other TFs, and it is these complexes that are responsible for binding to the genome. Therefore, we hope to develop more complex models that are motivated by our biological understanding of how TFs work. We expect such models to have increased performance compared to our current models.

Lastly, wet-lab data is the gold-standard in biological research. Therefore, we propose to collaborate with experimentalists to take samples of diseased cells from patients. We will use ATAC-seq to determine the regions bound by TFs in these samples, run our algorithm to determine the top TFs predicted to be involved in the samples, treat laboratory models of the disease with drugs that affect our top predicted TFs, and observe the results. We hope to observe decreased disease severity, which would demonstrate TF-Finder's usefulness as a computational biology tool.

6 Contributions

David Wu performed the dataset collection and curation for both the TF binding preference data and the test set from GEO and ran the data through PRISM and GREAT to obtain input data. Diwakar Ganesan implemented the NMF and PCA classifers, and created a data pipeline to convert David's curated dataset into a form palatable to numpy/scikit. The two worked together to put together the figures and cowrote the paper.

Our code is here: https://drive.google.com/open?id=1qky2icbS5U3RkZUNXDKd-zYgTzfoDBCP. Note that code for running PRISM and GREAT, and downloading/processing ChIP-Atlas files is not included, as they are run via command line tools.

References

[1] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[2] A. C. Joerger and A. R. Fersht, "The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches," *Annual review of biochemistry*, vol. 85, pp. 375–404, 2016.

[3] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "Great improves functional interpretation of cis-regulatory regions," *Nature biotechnology*, vol. 28, no. 5, p. 495, 2010.

[4] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "Jaspar: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D91–D94, 2004.

[5] D. E. Newburger and M. L. Bulyk, "Uniprobe: an online database of protein binding microarray data on protein–dna interactions," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D77–D82, 2008.

[6] I. V. Kulakovskiy, Y. A. Medvedeva, U. Schaefer, A. S. Kasianov, I. E. Vorontsov, V. B. Bajic, and V. J. Makeev, "Hocomoco: a comprehensive collection of human transcription factor binding sites models," *Nucleic acids research*, vol. 41, no. D1, pp. D195–D202, 2012.

[7] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, *et al.*, "Transfac®: transcriptional regulation, from patterns to profiles," *Nucleic acids research*, vol. 31, no. 1, pp. 374–378, 2003.

[8] A. M. Wenger, S. L. Clarke, H. Guturu, J. Chen, B. T. Schaar, C. Y. McLean, and G. Bejerano, "Prism offers a comprehensive genomic approach to transcription factor function prediction," *Genome research*, vol. 23, no. 5, pp. 889–904, 2013.

[9] Y. Tanigawa, J. Li, J. M. Justesen, H. Horn, M. Aguirre, C. DeBoever, C. Chang, B. Narasimhan, K. Lage, T. Hastie, *et al.*, "Components of genetic associations across 2,138 phenotypes in the uk biobank highlight adipocyte biology," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[10] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, *et al.*, "Ncbi geo: archive for functional genomics data sets—update," *Nucleic acids research*, vol. 41, no. D1, pp. D991–D995, 2012.