Prediction of Transcription Factors that Regulate Common Binding Motifs

Dana Wyman and Emily Alsentzer
CS 229, Fall 2014

Introduction

A. Background

Proper regulation of mRNA levels is essential to nearly all cellular processes. Transcription factors are responsible for regulating levels of mRNA. By binding to a specific DNA sequence via a DNA-binding domain, transcription factors can promote or block the recruitment of RNA polymerase, an enzyme responsible for transcription, effectively controlling when and how much of a gene is expressed [1]. Predicting the binding sites of transcription factors is an important area of research, as identifying these regulatory sites can shed light on the mechanisms that regulate specific genes. A variety of algorithms have been developed in order to predict transcription factor binding sites (TFBSs) a priori [2]. These techniques have the potential to identify novel transcription factor binding motifs. However, in some cases, the transcription factor responsible for binding a given motif is unknown. Here, we address this issue using gene set composition to predict which transcription factors bind common binding motifs. This approach differs from other work in the field, as previous efforts have mainly focused on predicting targets of known transcription factors.

B. Data

I. Gene Sets

A common way of determining which transcription factor regulates a given gene is to look for characteristic DNA binding motifs within the promoter of the gene [3]. In general, each transcription factor has a motif around 10 base pairs in length to which it preferentially binds [1]. Therefore, it is possible to sequence genes and group them based on which transcription factor site(s) their promoters contain. One of our datasets is a list of transcription factor-gene sets compiled from the ChIP Enrichment Analysis system (CHEA), the Broad Institute MsigDB database, and the Encyclopedia of DNA Elements (ENCODE) Project [4]. It includes 757 transcription factors, along with the set of genes that each is expected to bind to and regulate, and some transcription factors appear with multiple gene sets.

II. Expression Data

RNA-seq is a recent technique used to quantify gene expression by directly sequencing RNA molecules from a cell sample. The sequenced RNA is converted to a library of cDNA fragments and mapped to specific genes in the genome, and the number of reads for a gene can be used to quantify gene expression [5]. Fragments per kilobase of exon per million reads mapped (FPKM) measurements are often used to quantify gene expression because they allow for normalization of reads by gene length. FPKM is defined as the following:

$$FPKM_i = \frac{x_i}{\tilde{l}_i N} * 10^9$$

where x_i= number of reads that align to a gene, $\tilde{l}_i$ = effective length of a gene, and N = number of reads sequenced [5].

Our dataset contains RNA-seq FPKM measurements from 95 human individuals across 27 different tissue types [6]. Each tissue sample contains F
PKM measurements for 20,050 genes. The data was retrieved from ArrayExpress under accession number E-MTAB-1733 [7].

Methods

I. Data Preprocessing

A. Expression Data

In the original table, genes were labeled using the Ensembl gene ID convention. These identifiers are not directly compatible with other bioinformatics resources, so Biomart was used to generate a mapping of Ensembl IDs to official HGNC gene symbols [8]. When a gene symbol appeared more than once in the table, gene expression was summed across the rows to produce a single entry. Finally, FPKM values were $\log_2$-

transformed, which is a standard RNA-seq processing step.

B. Gene Set Filtering

All gene sets for species other than *Homo sapiens* were removed. After filtering, the data contained 757 gene sets.

II. Assessing the Relationship Between Transcription Factor and Gene Set Expression

In order to determine whether gene expression data could be used to predict transcription factors that regulate common binding motifs, we first tested the assumption that transcription factors and their gene targets have correlated expression profiles. For each gene set, we computed the sum of the Euclidean distance of the TF from each gene in the gene set as shown:

$$D_{geneset} = \sum_{gene} \|TF - gene\|$$

Then, we repeated the above calculation using a random gene set selected from a list of the genes across all gene sets, and calculated the fraction of times that value obtained with a random gene set was less than the original $\mathrm{D_{geneset}}$ out of N repetitions. Here N = 100.

Level	Name	Definition
1	Superclass	General topology of the DBD*
2	Class	Structural blueprint of the DBD
3	Family	Sequence and functional similarities
4	Subfamily	Sequence-based subgroups
5	Genus	TF gene
6	Factor 'species'	TF polypeptide

Table 1. TFClass Hierarchy Definitions. *DBD = DNA binding domain [9].

Category Number	Description	Training Examples
1	Basic Domains	81
2	Zinc-coordinating DNA-binding domain	113
3	Helix-turn-helix domain	105
6	Immunoglobulin folds	33
1.1	Basic leucine zipper factors (bZIP)	44
1.2	Basic helix-loop-helix factors (bHLH)	37
2.1	Nuclear receptors with C4 zinc fingers	42
2.3	C2H2 zinc finger factors	71
3.1	Homeo domain factors	41
3.3	Fork head / winged helix factors	31
3.5	Tryptophan cluster factors	17
6.2	STAT domain factors	19

Table 2. Superclasses and classes used

$$P_{bootstrap} = \frac{\sum_b 1\{ D^{b}_{geneset} \le D_{geneset}\}}{N}$$

II. Classification of Transcription Factors Using the TFClass Hierarchy

To make the number of response classes in the data more tractable for machine learning, the transcription factors were organized separately into superclasses and classes using the TFClass hierarchy (Table 1). This framework orders transcription factors based on their mode of interaction with DNA [9]. The TFClass Hierarchy (Sept. 2014 version) was downloaded in the .obo ontology format [9]. The superclass and class of each transcription factor from the gene set data was obtained by matching its name against the records in this ontology. When a factor failed to match, its synonyms were used to search the hierarchy as well. At both the class and superclass level, categories with fewer than 15 training examples were omitted. After this filtering, four superclasses (1, 2, 3, and 6) remained, as well as eight classes (1.1, 1.2, 2.1, 2.3, 3.1, 3.3, 3.5, and 6.2) (Table 2).

III. Multinomial Elastic Net Regression

A. Features

Each of the 12,089 features is a gene that appears at least once in any transcription factor gene set. The presence of a given gene in the gene set of a transcription factor is represented by a 1 at that index, and its absence is represented by 0.

B. GLMnet Model

We decided that multinomial logistic regression was the most appropriate machine learning method for this problem because the data contains more than two classes and has binary rather than continuous numerical features. Because there are many fewer gene set examples than individual genes, the dataset is prone to overfitting. To best address this, we chose multinomial elastic net regression as implemented in the GLMnet R package because it allows for variable selection via regularization with combined L1 and L2 norm penalization [10]. After the model was trained on all of the data, 10-fold cross validation was performed to obtain the test error.Equations for the multinomial model are shown here as described in the GLMnet Documentation [10]:

Suppose the response variable has K levels G = {1,2,…,K}. Here we model:

$$\Pr(G = k|X = x) = \frac{e^{\beta_{0k}+\beta_k^T x}}{\sum_{\ell=1}^{k} e^{\beta_{0\ell}+\beta_\ell^T x}}\ .$$

Let Y be the N x K indicator response matrix, with elements $y_{i\ell} = I(g_i = \ell$. Then the elastic-net penalized negative log-likelihood function becomes

$$\ell(\{\beta_{0k},\beta_k\}_1^k) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{k} y_{il}(\beta_{0k} + x_i^T\beta_k) - log\left(\sum_{k=1}^{k} e^{\beta_{0k}+x_i^T\beta_k}\right)\right)\right]$$

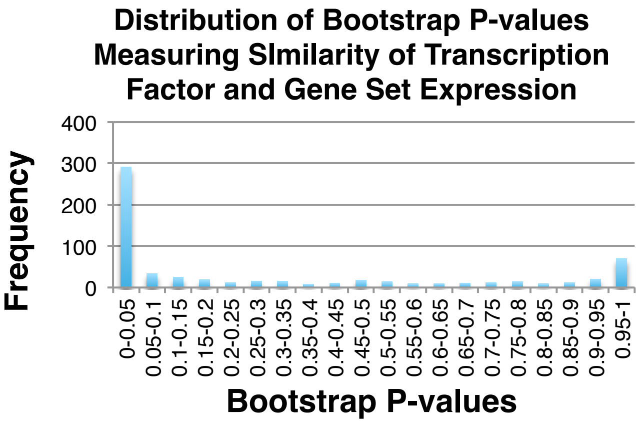

Figure 1: Distribution of Bootstrap P-values Measuring Similarity of Transcription Factor and Gene Set Expression. Out of 631 gene sets tested, 291 had a significant p-value (threshold of 0.05). This indicates that some transcription factors show similar expression to that of their gene sets across tissues, but it is not a valid assumption for the entire data set.

$$+\lambda\left[\frac{(1-\alpha)\|\beta\|_F^2}{2} + \alpha\sum_{j=1}^{p}\|\beta_j\|_q\right].$$

where β is a $p\ x\ K$matrix of coefficients. β_k refers to the kth column (for outcome category k), and β_j the jth row (vector of K coefficients for variable j).

Results

It has been proposed previously that transcription factors exhibit similar expression to that of the targets they regulate [3]. After an unsuccessful attempt to use elastic net regression to predict transcription factors using gene expression, we decided to test the validity of this statement in our data. To do this, we calculated the Euclidean distance of the transcription factor to every gene in

Superclass	Training Error	Test Error
1	11%	48%
2	0.88%	28%
3	12%	43%
6	15%	45%

Table 3. GLMNet Error for Superclasses

Class	Training Error	Test Error
1.1	6.8%	39%
1.2	11%	78%
2.1	7.1%	67%
2.3	2.8%	35%
3.1	12%	39%
3.3	19%	45%
3.5	24%	76%

Table 4. GLMNet Error for Classes

the gene set. Then, we computed a bootstrap p-value for each gene set to determine whether the distance was greater than that of a randomly selected set. We found that only 46% of the gene sets had significant bootstrap p-values (Figure 1), which suggested that predicting transcription factors using gene expression could have limitations for our dataset.

In light of these findings, we took an alternative approach to predicting transcription factors by using gene set composition as features for a multinomial elastic net regression model. We performed two regression analyses using the transcription factor superclass and the transcription factor class as the response variables (Figure 2). Cross validation was used to determine the optimal lambda value for each regression. The regression predicting transcription factor superclass was 60% accurate with an overall training error of 8% and an overall test error of 40% whereas the regression predicting transcription factor class was 49% accurate with an overall training error of 9% and an overall test error of 51% (Table 3 and 4).

Discussion and Conclusions

Identification of transcription factors that bind genes with common binding motifs can provide insight into the regulatory mechanisms of these genes. Our results show that gene set composition is a better predictor of transcription factor classes compared to expression data. Furthermore, our model is better able to predict transcription factor superclass than class, which is unsurprising because there are fewer training examples per class compared to superclass.

Although gene set composition shows some promise for transcription factor prediction, there is still a need for improvement of our model. The 32% and 42% differences between test and training errors for superclass and classes respectively suggests that our model is overfitting the data despite the use of L1 and L2 norm penalized logistic regression. Other alternatives for feature selection may better eliminate overfitting. In the future, we could use L1 norm penalization (lasso regression) or L2 norm penalization (ridge regression) instead of elastic net, which has both L1 and L2 norm penalization. Feature selection via forward search, backwards search, or filter feature selection with mutual

Figure 2: Superclass and Class Confusion Matrices.
A)Confusion matrix for superclasses illustrating how training examples were classified during the testing run of multinomial GLMnet. Entries on the diagonal were classified correctly.
B) Confusion matrix for classes illustrating how training examples were classified during the testing run of multinomial GLMnet.

A	Predicted Superclasses			
Actual	42	33	6	0
	7	81	23	2
	8	36	59	2
	1	7	7	18

B	Predicted Classes								
		1.1	1.2	2.1	2.3	3.1	3.3	3.5	6.2
Actual	1.1	27	4	4	8	0	1	0	0
	1.2	5	8	5	17	1	0	1	0
	2.1	1	2	14	22	2	0	0	1
	2.3	2	7	5	46	5	4	1	1
	3.1	0	1	0	11	25	3	0	1
	3.3	1	0	3	8	2	17	0	0
	3.5	0	3	0	9	0	1	4	0
	6.2	0	0	2	8	1	1	0	7

information scoring may also yield better test errors.

In addition to correcting for overfitting in the model, there is also a need for further granularity of response variables. Although it is interesting to predict transcription factor class for a set of genes with a common binding motif, in order for these results to be biologically useful, prediction at the family, subfamily, or even genus level is needed. Additional data describing the transcription factors that regulate genes with common binding motifs is necessary in order to predict transcription factors at a finer granularity, and it is also needed to ensure that there are training examples of transcription factors in each superclass and class so that all x superclasses and all x classes can be predicted. With additional data and improved measures to reduce overfitting, our approach could be a useful tool for identifying unknown transcription factors with a set of known gene targets.

Future

Previous research has indicated that transcription factors may regulate genes that have similar functional roles [3]. Therefore, another approach to predicting transcription factors using gene sets would be to use gene ontology (GO) terms as features in a multinomial logistic regression model. GO terms classify genes by cellular component, molecular function, and biological process, and are organized in a hierarchy with increasing specificity [11]. This hierarchical structure is useful in this context because it would allow for different levels of detail to be tested in the features. In the future, an approach along these lines could further improve our ability to predict which transcription factor binds to a given gene set.

Acknowledgements

We would like to thank Andrew Gentles for helping us develop our project idea and for providing guidance and technical support along the way.

References

1. Weingarten-Gabbay, Shira, and Eran Segal. "The Grammar of Transcriptional Regulation." Hum Genet 133 (2014): 701-11.
2. Mathelier, Anthony, and Wyeth W. Wasserman. "The Next Generation of Transcription Factor Binding Site Prediction." Ed. Ilya Ioshikhes. PLoS Computational Biology 9.9 (2013)
3. Ernst, Jason, Heather L. Plasterer, Itamar Simon, and Ziv Bar-Joseph. "Integrating Multiple Evidence Sources to Predict Transcription Factor Binding in the Human Genome." Genome Res. 20 (2010): 526-36.
4. Gentles, A. J., Alizadeh, A. A., Lee, S.-I., Myklebust, J. H., Shachaf, C. M., Shahbaba, B., … Plevritis, S. K. A pluripotency signature predicts histologic transformation and influences survival in follicular lymphoma patients. Blood, 114(15), (2009) 3158–66.
5. Mortazavi, Ali, Brian A. Williams, Kenneth Mccue, Lorian Schaeffer, and Barbara Wold. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." Nature Methods 5.7 (2008): 621-28.
6. Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol. Cell. Proteomics 13 (2014): 397–406.
7. Rustici.G et al. 2013 ArrayExpress update - trends in database growth and links to data analysis tools. Nucleic Acids Res, doi: 10.1093/nar/gks1174.
8. Kasprzyk, A. "BioMart: Driving a Paradigm Change in Biological Data Management." Database 2011.0 (2011).
9. Wingender, E., Schoeps, T. and Dönitz, J.: TFClass: An expandable hierarchical classification of human transcription factors. Nucleic Acids Res. 41, D165-D170 (2013).
10. Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
11. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.