33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Prioritizing Predicted *cis*-regulatory Elements for Co-expressed Gene Sets based on Lasso Regression Models

Hong Hu, Damian Roqueiro *Member, IEEE* and Yang Dai, *Member, IEEE*

***Abstract* — Computational prediction of *cis*-regulatory elements for a set of co-expressed genes based on sequence analysis provides an overwhelming volume of potential transcription factor binding sites. It presents a challenge to prioritize transcription factors for regulatory functional studies. A novel approach based on the use of Lasso regression models is proposed to address this problem. We examine the ability of the Lasso model using time-course microarray data obtained from a comprehensive study of gene expression profiles in skin and mucosal wounds in mouse over all stages of wound healing.**

I. Introduction

TIME-COURSE transcriptome studies of biological processes have provided valuable insights into the dynamic regulatory mechanisms of these processes [1-3]. The analysis of temporal profiles of gene expression reveals that genes can be divided into groups with distinct co-expression patterns over a time period. The identification of specific transcription factors (TFs) that regulate a specific set of co-expressed genes is an important step towards the understanding of underlying biological mechanisms. Numerous computational methods have been developed for the prediction of over-represented transcription factor binding sites (TFBSs) in the promoter regions of a set of genes [4-6]. However, the information obtained about the total number of the predicted TFBSs, their locations, binding scores and statistical significance can be overwhelming. This presents challenges when trying to accurately determine the set of potential TFs for a regulatory functional study in a biological investigation.

We propose a novel approach based on the use of Lasso regression models to address the issue of selecting the most likely related TFBSs for multiple sets of co-expressed genes obtained from a typical analysis of time-course microarray data. The lasso regression models [7][8] are very well suited as selection and shrinkage estimation methods for classification problems using microarray gene expression data. As it is widely known, in these data the number of genes is much larger than the number of arrays involved in the experiment [9].

For our proposed lasso models, we utilize a unique set of features that describe a TFBSs' binding strength, its location on the promoter region and its statistical enrichment p-value, all of which were obtained from promoter analysis tools. Additionally, we consider extra features representing the joint contribution of two TFBSs. Herein, we use the term "cluster" to represent a set of co-expressed genes that were obtained from clustering analysis of mRNA expression profiling data under certain condition. Our models take multiple clusters of genes and attempt to produce a set of individual TFBSs, in addition to pairs of TFBSs, for each cluster simultaneously. These TFBSs are considered regulators of the gene expression in each cluster. We examine the ability to discern the representative TFs for each cluster in the proposed model using time-course data from a comprehensive study of gene expression profiles in skin and mucosal wounds in mouse over all stages of wound healing [3]. We demonstrate the potential utility of this approach in identifying the regulation strength of individual TFs and the joint effect of two TFs.

Manuscript received April 15, 2011, revised June 20, 2011.

Hong Hu, Damian Roqueiro and Yang Dai are with the Department of Bioengineering (M/C 063), University of Illinois at Chicago, 851 S Morgan St, SEO 218, Chicago, IL 60607 USA (e-mail: {hhu4, droque1, yangdai}@uic.edu).

II. Methods

We assume that a set of genes, with a similar gene expression profile across several time points, are regulated by a common set of TFs. By analyzing the promoter regions of these genes, a set of over-represented TFBSs can be obtained. We use the lasso penalized logistic regression model to select a set of TFBSs that are mostly associated to a set of co-expressed genes. This model allows the selection of a set of TFBSs for each individual cluster when multiple clusters are involved. The goal of our model is not only to identify individual TFBSs that regulate a set of genes, but also to detect pairs of TFBSs that concurrently affect the expression of those same genes.

A. Lasso-penalized multinomial logistic regression

The lasso regression is a penalized regression model that uses $\ell 1$ penalty to achieve a sparse solution, especially for problems where the number of predictors p far exceeds the number of observations n [7][8]. Suppose a dataset consists of a set of n points X_i in R^p, each with a response $y_i \in R$. Let $\theta = (\beta_0, \beta) \in R^{p+1}$. The lasso linear regression is defined as a minimization problem with the following objective function:

$$f(\theta) = \sum_{i=1}^{n}(y_i - \beta_0 - {X_i}^T\beta)^2 + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}}|\beta_{\mathrm{j}}|,$$

where λ is a penalty parameter. With the proper choice of value for this parameter in the regression model, the over-fitting can be avoided and the prediction is more reliable. The lasso model has been extended to a logistic regression in which we maximize the following objective function:

$$f(\theta) = l(\theta) - \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $l(\theta)$ is the logistic log likelihood function. This model can be further generalized to a multinomial logistic regression when the categorical response variable has more than two levels [8].

We utilize the multinomial lasso logistic regression to identify groups of TFBSs, each of which is unique to a cluster of genes. The dataset in consideration consists of clusters of genes that were obtained from a genome-wide transcriptome study. We hope to simultaneously identify the TFBSs that act on the individual clusters.

The recently released database cREMaG (*cis*-Regulatory Elements in the Mammalian Genome) [6], is designed for *in silico* studies of the promoter properties of co-expressed mammalian genes. This tool can identify over-represented TFBSs in the promoter regions of a set of specific co-expressed genes. For each of our genes' clusters, we obtained information about the most over-represented binding sites of TFs predicted by cREMaG. We utilize 3 data fields to build variables for our regression models: a) The similarity score of a TFBS based on a Position Weight Matrix (PWM) [10-12] of the TF; b) The position of a TFBS on the promoter; c) The TF's fold p-value, which is a statistical significance measure of the TF fold enrichment in the input gene set against a random fold distribution. For details please refer to [6].

B. Preparation of model variables

Two types of independent variables are considered in our lasso models. The type I variables relate to individual TFBSs. The type II variables relate to the joint effect of two TFBSs in the regulation of genes. We also incorporate the information of the distance between a TFBS and the transcription start site (TSS) of a gene. To reduce the number of variables, we divide a promoter sequence into several regions depending on the distance to the TSS. Each region is represented by a bin.

Here we define the **type I** variable β_{jk} associated to the vector $X_{jk} \in R^N$. Each element of X_{jk} is a binding score x_{ijk} (i=1...N; j=1...J; k=1...K) for gene i, TF j, and bin k.

x_{ijk} represents the strength of the TFBSs of TF j identified in bin k in the promoter region of gene i. Therefore, if TF j has no TFBSs identified in bin k, then x_{ijk} is set to 0. On the other hand, if TF j has one or more TFBSs identified in bin k, then the value of x_{ijk} is calculated based on the output from cREMaG as follows. We first compute average similarity scores $\bar{s}_{ijk}$ for all identified TFBSs of TF j in bin k, then we multiply $\bar{s}_{ijk}$ by the negative logarithm of the fold p-value of TF j in the current cluster. The obtained value is then assigned to x_{ijk}.

$$x_{ijk} = \begin{cases} -\bar{s}_{ijk} \cdot \ln p_j, & \text{if bin } k \text{ includes one ore more TFBSs of TF } j \text{ in the promoter of gene } i \\ 0, & \text{Otherwise} \end{cases}$$

where:

$i = 1 \ldots N$, N is the total number of genes

$j = 1 \ldots J$, J is the total number of TFs

$k = 1 \ldots K$, K is the total number of position bins in the promoter

We also considered different weights for $\bar{s}_{ijk}$ based on a threshold p_0 for the fold p-value in each score x_{ijk}:

$$x_{ijk} = \begin{cases} \bar{s}_{ijk} \cdot (\ln p_j)^2, & \text{if } p_j \le \frac{1}{e} \text{ or } p_j \ge p_0, \text{and bin } k \text{ includes one or more TFBSs of TF } j \text{ in the promoter of gene } i \\ \bar{s}_{ijk} \cdot \sqrt{-\ln p_j}, & \text{if } p_0 < p_j < \frac{1}{e}, \text{and bin } k \text{ includes one or more TFBSs of TF } j \text{ in the promoter of gene } i \\ 0, & \text{Otherwise} \end{cases}$$

In summary, X_{jk} is defined as:

$$X_{jk} = [x_{1jk}, x_{2jk}, \ldots, x_{Njk}]^T$$

In our model, the **type II** variables have the same format as type I variables and they represent the joint effect of two TFs. As before, a type II variable $\beta_{j_2k_2}$ will be associated to a vector $X_{j_2k_2}$ defined as:

$$X_{j_2k_2} = \sqrt{X_{j_1'k_1'} \cdot X_{j_1''k_1''}}$$

where j_1', j_1'' are indices for two TFs; k_1', k_1'' are bin indices for two locations on the promoter. j_2 and k_2 are TF-pair and bin-pair indices, respectively.

The dependent variable in our model, $Y \in R^N$, indicates the cluster to which each gene belongs. Each element y_i can take a value between 1 and C where C is the total number of clusters (e.g., $y_i = m$, if gene i belongs to cluster m).

III. Results

We consider two lasso multinomial regression models in our study. The models will analyze multiple clusters of genes simultaneously to determine the TFs that are most heavily involved in each cluster of genes. The two models are provided as follows.

Model 1 - only type I variables are included. A non-zero variable obtained from the lasso model for one gene cluster is explained as the regulatory strength of the TF at the identified location on the promoter regions, within the gene cluster.

Model 2 - both type I and type II variables are included. A non-zero type II variable is explained as the regulatory strength of two TFs at (possibly different) locations on the promoter regions, within the gene cluster.

A. Dataset

We selected a temporal microarray gene expression dataset from the published literature to demonstrate the utility of this approach. This dataset is the first systematic, comprehensive and dynamic study of gene expression profiles in skin and mucosal wounds in mouse over all stages of wound healing [3]. The significantly similar and differentially expressed genes in skin and mucosal wounds were successfully identified and grouped using well-established mouse models and microarray technology. The

identification of 5 clusters of genes shows similar, but not identical, patterns of expression in wounds in skin and tongue. As described in [3], for both tissues, the differentially expressed genes are grouped into 5 clusters as: early down, early up high, early up low, early up medium and late up. The number of genes in each cluster is provided in Table I. For details of this gene expression analysis please refer to [3]. The critical step to develop an understanding of the transcription regulation related to the wound healing process is the discovery of the common transcription regulators for each cluster.

TABLE I
NUMBER OF GENES IN EACH CLUSTER

Cluster	Skin	Tongue
Early down	236	47
Early up high	54	18
Early up low	376	119
Early up medium	158	72
Late up	130	88

The number of genes in each gene cluster in skin and tongue identified by wound healing study in Chen et al. [3].

We queried the cREMaG database using the genes in each cluster in skin and tongue respectively. The parameter settings of the queries are listed in Table II.

TABLE II
PARAMETER SETTINGS OF CREMAG DATABASE QUERY

Parameter	Selection
Conservation threshold	70%
Top percent of conserved region	100%
Max number of most conserved TFBSs	Top 100
Coding/Non-coding sequence	Non-coding
Length of upstream segment	10,000bps
Length of downstream segment	1,000bps
Random TFBS occurrence	2TFBS/10,000bps
Precompiled background	Conserved promoter

The query results from the cREMaG database were reorganized into matrices to be the inputs for the lasso multinomial logistic regression. The position bin size was set to 2,000bps. Therefore, the promoter region upstream of the TSS was partitioned into 5 position bins covering a length of 10,000bps. The promoter region downstream of the TSS was represented by 1 position bin of 1,000bps. We calculated the value for type I and type II variables as described before.

B. Evaluation procedure

We used the R package *glmnet* [9] to solve the multinomial lasso-penalized logistic regression models. Using its internal 5-fold cross-validation procedure we obtained, for each tissue, the optimal λ values of both models with the minimum cross-validated error rate. Non-zero values for type I and type II variables for optimal λ and cross-validated error rate were retained in each individual gene cluster for further analysis. Table III summarizes the information on the optimal λ and the corresponding minimum cross-validated error rate for each model and tissue.

To further control the number of variables in the final solution, we restricted the maximum number of variables (Dfmax) to be included in each model. From our experiment we observed that when the upper bound on the number of variables allowed for the model (Dfmax) is reduced to 100, there is no significant increase in the Min error rate (Table IV) compared to the models where there is no Dfmax restriction. Therefore, we provide our final results based on the models determined with Dfmax=100 in Table V.

TABLE III
OPTIMAL λ AND MINIMUM ERROR RATE

		Skin		Tongue	
Fold p threshold(p_0)		λ	Min error rate	λ	Min error rate
Model 1	0.01	0.018	0.464	0.018	0.455
Model 1	0.05	0.009	0.302	0.014	0.425
Model 2	0.01	0.037	0.538	0.067	0.616
Model 2	0.05	0.027	0.404	0.078	0.501

TABLE IV
MINIMUM ERROR RATE WITH MAXIMUM NUMBER OF VARIABLES

		Skin		**Tongue**	
		Min error rate		**Min error rate**	
	p	No Dfmax	Dfmax = 100	No Dfmax	Dfmax = 100
Model 1	0.01	0.464	0.467	0.455	0.484
Model 1	0.05	0.302	0.359	0.425	0.480
Model 2	0.01	0.538	0.544	0.616	0.627
Model 2	0.05	0.404	0.420	0.501	0.513

TFs corresponding to type I and type II variables are selected from the final lasso models using the following criteria: for each of model 1 or model 2, we selected variables with values ≥ 0.001 and commonly found in both regression models of different p_0. We can observe the consistency of common TFs selected in the same groups for the two tissues as well as unique TFs for different clusters and tissues. For instance, one of the TFs, NF-kappaB [13], which plays a key role in regulating the immune response to infection, is identified in both models in both tissues in the *early up high* group. Lhx3 [14], which is a TF required for pituitary development and motor neuron specification, is only identified in skin. Several joint effects are also identified, including the effects between RELA and NF-kappaB. Researchers have revealed that NF-kappaB will bind to RELA to form a complex which will be activated, then translocated into the nucleus and will later bind to DNA [15].

IV. CONCLUSION

Studies have shown that *cis*-regulatory elements are bound by transcription factors, in a sequence-specific manner, to determine biological processes in cells and tissues. However, the overwhelming number of possible TFs obtained from sequence analysis presents a difficulty to

determine the regulatory relationship between TFs and their target genes. Further, the identification of combinatorial patterns of two TFs is not trivial from the predicted results. In this work we proposed multinomial logistic regression models to learn a set of strongly associated individual TFBSs in addition to pairs of TFs and their joint effect in a set of co-expressed genes.

Our models integrated the TFBSs' binding similarity, the statistical measure of over-representation and the TFBSs' positions. In this way, we were able to determine the importance of multiple TFBSs belonging to the same TF, the combination of two TFBSs of the same TF, or the combination of different TFs observed at different positions in promoter regions. The preliminary evaluation of the proposed lasso multinomial logistic regression based on the dataset of wound healing in skin and tongue demonstrated the potential ability of this method to identify functional *cis*-regulatory elements.

TABLE V
SOME IDENTIFIED EFFECTS IN DIFFERENT MODELS

	cluster	TF	TF Class
Skin Model 1	Early down	Lhx3	HOMEO
	Early up high	RELA	REL
		NF_kappaB	REL
	Early up low	CREB1	bZIP
		MZF1_5_13	ZN-FINGER
		ELK1	ETS
		ELK4	ETS
	Early up medium	ELF5	ETS
		SPIB	ETS
	Late up	Evi1	ZN-FINGER
		Nobox	HOMEO
Skin Model 2	Early down	Lhx3	HOMEO
		*CREB1	bZIP
	Early up high	RELA	REL
		NF_kappaB	REL
	Early up low	ELK1	ETS
	Early up medium	RELA	REL
		*RELA	REL
		FOS	bZIP
		*SPIB	ETS
	Late up	Evi1	ZN-FINGER
Tongue Model 1	Early down	MZF_5-13	ZN_FINGER
		Pax4	PAIRED-HOMEO
		NR1H2-RXRA	NUCLEAR RECEPTOR
		ELK1	ETS
	Early up high	Cebpa	bZIP
		NF_kappaB	REL
	Early up low	Foxq1	FORKHEAD
		IRF2	TRP-CLUSTER
	Early up medium	IRF1	TRP-CLUSTER
		Foxd3	FORKHEAD
	Late up	MEF2A	MADS
Tongue Model 2	Early down	Myb	TRP_CLUSTER
	Early up high	NF_kappaB	REL
	Early up low	Foxq1	FORKHEAD
		IRF2	TRP-CLUSTER
		Arnt-Ahr	bHLH
		*ELK1	ETS
	Early up medium	RELA	REL
		*HLF	bZIP
	Late up	RELA	REL
		*Arnt	bHLH

* This TF has joint effect with the TF in the previous row

To further expand our model, several possible directions can be explored. First, our model can be modified for further dissection of associated TFs in condition-specific regulatory networks. Second, in a multinomial lasso regression model we will normally incorporate one explanatory variable for each class. In our case, the algorithm will have a strong tendency to choose a vector X_{jk} for just one cluster. This may not be biologically significant because one TF may have similar binding patterns in different sets of genes. Further development of our mathematical model is required to capture this behavior. Third, in our model, independent variables represent the binding strength of different TFs in different promoter regions, which may not fully meet the orthogonal assumption in lasso regression. To address this issue, i.e. when strong correlations exist between the independent variables, additional modifications can be done to identify explanatory TFs in a group manner. This can be achieved through group lasso regression or other novel methods.

REFERENCES

[1] Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424: 147-151

[2] Juskeviciute E, Vadigepalli R and Hoek, JB (2008) Temporal and function profile of the transcriptional regulatory network in the early regenerative response to partial hepatectomy in the rat. BMC Genomics 9: 527

[3] Chen L, Arbieva ZH, Guo S, Marucha PT, Mustoe TA and DiPitetro LA, (2010) Positional differences in the wound transcriptome of skin and oral mucosa. BMC Genomics 11:471

[4] Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287

[5] Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW. (2007) oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Res 35: W245-252

[6] Piechota M, Korostynsk M, Przewlocki R (2010) Identification of cis-Regulatory Elements in the Mammalian Genome: The cREMaG Database. PLoS ONE 5(8): e 12465.

[7] Tibshirani R, (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58:267-288

[8] Krishnapuram B, Carin L, Figueiredo M A, Hartemink A J, (2005) Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattn Anal Mach Intell,27,957-968

[9] Frideman J, Hastie T, Tibshirani R, (2010) Regularization Paths for Generalized Linear Models via Coordinated Descent. Journal of Statistical Software, 33(1), 1-22

[10] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32:D91-4.

[11] Gribskov,M., Liithy.R. and Eisenberg.D. (1990) Profile analysis. Methods Enzymoi., 183, 146-159.

[12] Bucher,P. (1990) Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 5023 unrelated promoter sequences. J. Mol. Biol., 212, 563-578.

[13] Gilmore TD (2006). "Introduction to NF-κB: players, pathways, perspectives". Oncogene 25 (51): 6680–4.

[14] Sloop KW, Showalter AD, Von Kap-Herr C, Pettenati MJ, Rhodes SJ (May 2000). "Analysis of the human LHX3 neuroendocrine transcription factor gene and mapping to the subtelomeric region of chromosome 9". Gene 245 (2): 237–43.

[15] Nolan GP, Ghosh S, Liou HC, Tempst P, Baltimore D (Apr 1991). "DNA binding and I kappa B inhibition of the cloned p65 subunit of NF-kappa B, a rel-related polypeptide". Cell 64 (5): 961–9