Statistical Methods for Joint Data Mining of Gene Expression and DNA Sequence Database

Marla D. Curran
Aventis Pharmaceuticals
Biometrics & Data Mgmt
Bridgewater, NJ 08807
Marla.Curran@aventis.com

Hong Liu
Aventis Pharmaceuticals
Molecular Immunology
Bridgewater, NJ 08807
Hong.Liu@aventis.com

Fan Long
Aventis Pharmaceuticals
Molecular Immunology
Bridgewater, NJ 08807
Fan.Long@aventis.com

Nanxiang Ge
Aventis Pharmaceuticals
Biometrics & Data Mgmt
Bridgewater, NJ 08807
Nanxiang.Ge@aventis.com

ABSTRACT

One of the purposes of microarray gene expression experiments is to identify genes regulated under specific cellular conditions. With the availability of putative transcription factor binding motifs, it is now possible to relate gene expression pattern to the pattern of transcription factor binding sites (TFBS), as well as study how TFBS interact with each other to control gene expression. The objectives of this study are to develop a systematic approach for combining data from microarray gene expression experiments and the corresponding regulatory motif patterns in order to delineate gene regulation mechanisms. A secondary goal is to develop a predictive model for finding similarly regulated genes. Three consecutive procedures are proposed for such data mining activities. First, a linear mixed-effect model is fit to data from microarray gene expression experiments and potential regulated (positive) genes are identified based on a specific biological hypothesis. Putative TFBS are then retrieved for the identified positive genes and randomly selected controls. Second, a cluster analysis is conducted to reduce collinearity among the binding sites. In the third step, logistic regression is applied to choose the best model to predict gene type (positive, control) based on the numerous TFBS predictors. The above approach was applied to an internal example and a model was developed to predict up-regulated genes in activated *T*-helper (*Th*) cells. Using a leave-one-out cross- validation scheme, the model has an 18.9% false positive rate and a 41.7% false negative rate.

Keywords

Microarray, transcription factor binding site (TFBS), cluster analysis, logistic regression, modeling, regulatory motifs, *T*-helper cells.

1. INTRODUCTION

The activation or repression of gene transcription in a eukaryotic genome involves binding of DNA sequences by transcription factors (regulatory proteins). These binding sites are arrayed within several hundred base pairs predominantly upstream from the Transcription Start Site (TSS) in the promoter [1]. Identifying regulatory regions with transcription characteristics and then finding potential TFBS, which may interact or directly contribute to the specific gene expression, continues to be a challenge. It is important to find a powerful method to accomplish these goals.

Research has been conducted to propose several methods for identifying regulatory genes important to cell expression, for cluster analysis of expression data and TFBS data, and for identifying putative TFBS. Analyzing the expression data first to identify particular genes before the TFBS data analysis has been proposed, as well as using TFBS data to predict expression levels. We now proceed to describe a few of these statistical methods.

Pilpel, et al. screened genomic sequences against a database of known and putative motifs (or significant TFBS) to identify genes that contain the motifs [2]. This information is then modeled to calculate the effect of the motifs on gene expression. For each motif, an expression coherence score is calculated to measure how similar the genes are with that motif using expression profiles. Caselle, et al. presented a novel computational method to identify regulatory elements in the upstream region of eukaryotic genes [3]. The genes are clustered based on an overrepresented motif, the average expression level is determined, and then if the level is higher or lower than the whole genome average, the motif is likely responsible.

Most recently, Conlon et al. proposed MOTIF REGRESSOR for discovering sequence motifs upstream of genes that undergo expression changes in a given condition [4]. The genes are initially ranked by expression and the DNA sequence upstream from these genes is selected. They used Motif Discovery scan or MDscan [5] to independently identify candidate motifs and then scored the given sequences by the number of motif matches to the MDscan output. Insignificant motifs were removed using a simple linear regression between the score and gene expression and significant motifs were grouped based on a stepwise regression analysis.

In this study, we propose to use microarray gene expression profiling experiments to identify potential regulated genes, based on similar expression profiles in distinct subsets of human *T*-helper (*Th*) cells. We will develop a model to predict up-regulated genes in activated *Th* cells based on the pattern of appearance of TFBS from expression data. *Th* cells are a subset of *T*-cells that carry the *T4* marker and are essential for turning on antibody production, activating cytotoxic *T*-cells and initiating other immune responses [6]. They can be divided into 3 subsets based on their cytokine secretion profiles. Type 1 (*Th1*) cells predominantly secrete IL-1, IL-2, TNF-beta, and IFN-gamma. The secretion of IL-4, IL-5, IL-6, IL-10 and IL-13 characterizes type 2 (*Th2*) cell responses. The last set, referred to as *Th0*, can produce both a *Th1*- and *Th2*-type cytokine, so they are not clearly differentiated from the others.

DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide, provide a practical and economical tool for studying gene expression on a very large scale [7]. Expression analysis using

microarrays is conducted by identifying changes in mRNA levels. The mRNA from cell samples is used to conduct gene-profiling experiments with chip analysis, most often with Affymetrix GeneChip® arrays. The experiments or scans represent a combination of different factors with results collected for a large number of genes. From cord blood samples, these factors are cell types (*Th0, Th1, Th2*) from different donors (A, B, C), maintained for specified time points (4 days, 8 days) and classified as either resting or activated (cell action).

Linear mixed effect modeling will be used to analyze the microarray data and detect genes that are positive for a particular cell expression. Coupled with a set of genes determined to be unexpressed in these cells, we have a set of positive training samples and negative (control) training samples. Using the information of putative TFBS in the promoter region of these genes, we propose to use a logistic regression model to identify important binding sites to differentiate positive genes from negative genes. Furthermore, the developed model can be used to screen the whole genome to predict regulatory genes for a particular expression profile.

2. METHODS

2.1 Expression Data and Normalization

Typically, expression data is normalized due to a number of reasons, including unequal quantities of starting RNA, differences in labeling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels [8]. Scaling or normalization will allow informative comparisons of the expression levels to be made.

Although median scaling is a popular method used in previous research [9,10,11], we choose to normalize the data using a 'median-regression' or piece-wise linear regression (PLR) approach [12]. First, the median scan is found by calculating the median of all scans for every gene. Each scan is then regressed to the median scan and a slope is estimated. Dividing each scan value by its corresponding slope normalizes the data.

2.2 Expression Analysis

In order to identify genes[1] showing higher expression, we fit a mixed model to the normalized data for each individual gene. A mixed model analysis can be used in place of the standard linear model in this case, since we are assuming the sample donor is a random effect impacting the variability of the data while all other effects are fixed. If we assume there are two fixed effects, one random effect, and one interaction effect of interest, the model can be written as the following:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad (1)$$

where Y_{ijk} represents the observation for the i^{th} factor, j^{th} factor, and k^{th} donor and μ represents the overall mean (unknown fixed parameter). The fixed effect of the i^{th} and j^{th} factors are represented by parameters α_i, *and* β_j, respectively. The random effect of the k^{th} donor, γ_k, has the distribution $N(0, \sigma_\gamma^2)$. The interaction between the fixed effects are represented by parameter $(\alpha\beta)_{ij}$ and the experimental error (ε_{ijk}) is distributed as $N(0, \sigma^2)$. This model can be easily expanded to accommodate more than two fixed effects and more than one interaction effect.

Fitting this model to the data allows statistical inferences to be made. To determine which genes have the most effect on expression levels, criteria for the significance of model parameters are set *a priori*. For our purposes, genes are selected as "positive" if the main effect or interaction of chosen factor(s) are highly significant (p<0.0001) and selected as "negative" (or control) if the p-value is the least significant (0.6<p<1.0).

2.3 TFBS Data and Filtering

The positive genes, showing higher expression, and the control genes are mapped to Celera Human Transcripts (hCTs). For each hCT, we retrieved putative TFBS, which are located within –1.5 kilobases (kb) and +0.5 kb relative to the TSS and within a human-mouse conserved segment from Celera human TFBS database. Each hCT is mapped using The Transcription Factor Database, TRANSFAC [13], and all high scoring (p-value < 2×10^{-4}) TFBS are retrieved.

Since the goal of the final stage is to perform a logistic regression analysis with the binary response (1=positive gene, 0=negative gene) being dependent on the presence of the binding sites, filter criteria need to be defined to reduce the large number of predictors. Due to the sparseness of the data (i.e. large number of 0's), it is common for two predictors to display a linear relationship, when in fact the relationship is unknown. Filtering will reduce this occurrence.

The first step in filtering the predictors is to choose binding sites appearing on a minimum number of genes. The criterion chosen for the minimum number in this study is 1. If each binding site did not appear on at least 1 gene, the site is deleted. If a TFBS is found in either the control data and not in the positive data or vice versa, the missing values are converted to zeros as the two data sets are combined.

2.4 Cluster Analysis of TFBS

Cluster analysis can also reduce linear relationships or collinearity among binding sites. One of the biggest problems in regression modeling is collinearity, though in some instances, collinearity can provide a great deal of useful information. This is especially true for this study since collinearity can be due to redundant motifs (less informative) or synergistic motifs (more informative).

Redundant motifs could be two binding sites belonging to the same motif family at different positions or closely located based on their positioning. In both of these cases, the motifs are likely to be the same, but with the current statistical methods, we can only assume collinearity. Synergistic or cooperative binding of factors is the simultaneous interactions of two factors with closely situated target sites that can result in a non-additively high level of transcriptional activation [14]. We are hoping to find collinear binding sites, establish the type of collinearity based on previous research, and combine the binding sites without discarding valuable information.

The TFBS are compared pairwise for similarity (either they appear on the gene (1) or do not appear on the gene (0)). For example, given N genes, if TFBS1 and TFBS2 both appear, only

[1] At this stage, the term 'genes' is used instead of 'qualifiers' for ease of understanding.

one appears, or neither appears on a given gene, the following 2 × 2 table can be constructed.

TFBS1 → TFBS2 ↓	0	1	Total
0	n_1	n_2	n_1+n_2
1	n_3	n_4	n_3+n_4
Total	n_1+n_3	n_2+n_4	N

Table 1: Counts for pairwise comparison of binding sites

In order to cluster the binding sites, similarity (S) between a pair of TFBS is defined as the percentage of genes for which either both TFBS are present at its promoter region or neither present. Using the table above, the similarity percentage would be

$$S = \left(\frac{n_1 + n_4}{N} \times 100 \right). \quad (2)$$

Since we are working with sparse data, similarity only tells half of the story. For instance, two binding sites may be 98% similar, but only appear on 1 or 2 genes with the rest zeros. These variables are highly collinear, but mostly due to sparseness. The number of times both TFBS appear on a given gene (n_4) defines the more prevalent sites. If we set our cluster criteria as a balance of S and n_4, a list of the most prevalent among the most similar (collinear) variables is formed. The three cluster criteria will be S=90-92.5% and $n_4 \geq 15$, S=92.5-97.5% and $n_4 \geq 10$, and S=97.5-100% and $n_4 \geq 5$.

A new variable is then created to represent the union of the original two binding sites, i.e. if one OR the other TFBS appear on a gene (=1), then the new variable equals 1 for that gene, otherwise the new variable equals 0.

It is important to note this cluster analysis method may find its own use for analyzing this type of data, since none of the information from similar binding sites will be discarded. The new variable incorporates the original information from the collinear binding sites, so once the analysis is complete and the new variable is found to be significant, the individual binding sites can be explored further.

2.5 Logistic Regression Analysis

Multivariate logistic regression analysis is the statistical approach we used to classify objects (TFBS) into two outcomes or binary response (1=positive gene, 0=negative gene). The objects (regression coefficients) in our study are the individual binding sites or the new variables formed by the cluster analysis. Many variable selection methods can be used to formulate a final model, such as backward and forward elimination, recursive partitioning, and stepwise regression [15].

In this study, backward elimination and forward selection procedures are chosen to begin the logistic regression analysis. All explanatory variables in the initial model are specified as main effects, since testing pairwise interactions is impractical in the presence of many predictors.

The forward selection procedure adds the variable having the highest correlation with the dependent variable to the model. At each selection step, each variable is tentatively added to the model and the χ^2-statistic is calculated. After all predictors have been evaluated, the one with the highest χ^2-statistic is added to the model, if its associated p-value is less than the pre-specified level considered to be "significant". Once there are no longer significant predictors, the process terminates.

The backward elimination procedure is used to remove variables possibly incorporated in the model in error or because independent variables are correlated (redundancy). If removing the variable with the lowest χ^2-statistic does not result in a significantly poorer model fit, the variable is permanently eliminated. This continues until all the variables left in the model are considered significant.

The significance level of the score χ^2 for variable entry and the Wald χ^2 for variable elimination for these two procedures is set at 0.10. The significance level at this early stage is less conservative to allow all possible significant effects to remain in the model. The variables exhibiting significant main effects from both analyses are kept in the model for the next step.

In order to estimate the power (sensitivity) and Type I error (false positive rate), Receiver Operating Characteristic (ROC) curves are used to evaluate the full model and the reduced model using backward elimination. A goodness-of-fit test is appropriate to determine the fit of the reduced model. See the following section for details and an example.

Since the number of predictors in the model is reduced using the procedures above, interaction terms can be added. The forward selection and backward elimination procedures are employed again to find significant interaction terms. If the two procedures do not find the same interaction terms to be significant, significant effects from both are added to the model. Higher order interactions are added when two-way interactions involving common variables are observed.

2.6 ROC Curves and Goodness of Fit

An ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) of a model for varying cut-off points or criterion. The closer the curve follows the vertical axis and the top of the graph, the better the model, since a high sensitivity coincides with a low false positive rate. The following example ROC curves (Figure 1) show the difference between different tests with the least accurate predictive model being the 45-degree diagonal.

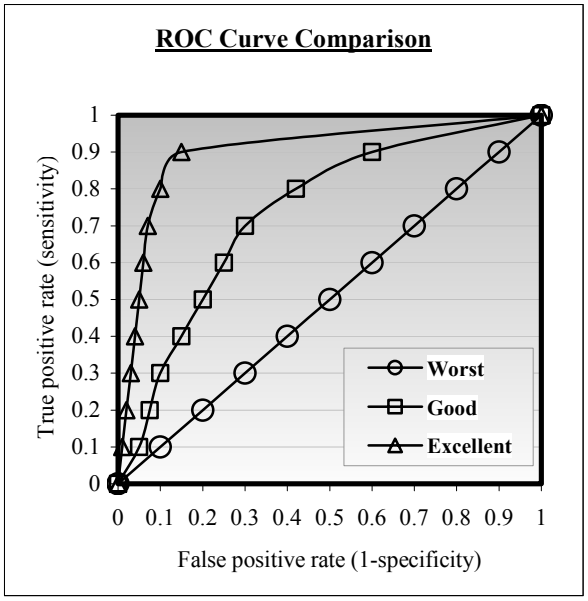

Figure 1: Example ROC curves

There are two applications of the ROC curve in the scope of this study. First with the original data, it is used to compare the full model including all main effects with a reduced model based on logistic regression analysis. Ideally, the power (sensitivity) will remain the same for both models at a chosen Type I error rate (1-specificity).

The second is to select cut-off values for validation of the model, once the final model is determined using the training sample. This final model is fit and the predicted values of the response variable ($\hat{y}$) are calculated. The ROC curve shows the range of sensitivity for an ideal range of false positive rates. Next, if we plot the error rates (false positive and false negative) against the predicted probability level, the cut-off probability for validation will be found where the upper and lower limit of the ideal false positive range intersect the two curves.

We chose to validate the final model using a leave-one-out cross-validation method. SAS® LOGISTIC calculates the cross - validated individual predicted probability of each response level. These probabilities are derived by leaving out one observation each time and recalculating the parameter estimates. Using the cutoff probability from the ROC curves, we can estimate the false positive and false negative error rates of our 'best' model.

3. RESULTS

From the example data set, the fixed effects are *time*, cell *type* and *action*, whereas *donor* is the random effect. Using (1), the mixed model is fit and we test the random and fixed main effects. The selection of positive and negative genes is based on the significance of the cell *action* main effect. With this data, 156 genes are selected as positive and 1000 genes are classified as controls. SAS® MIXED is used to analyze this data set. Once the positive and negative genes are selected with mixed model analysis, they are mapped to Celera Human Transcripts (hCTs).

Out of the 156 positive genes, 140 map to 133 hCTs. Out of the 133 positive hCTs, 72 have putative TFBS within –1.5 kb and +0.5 kb relative to the TSS in a human-mouse conserved segment. These 72 hCTs contain 209 distinctive TFBS. Using the same procedure for the 1000 negative genes, 713 map to 691 unique hCTs. Out of the 691 negative hCTs, 381 have putative TFBS within a conserved promoter region. The final negative set contains 381 hCTs and 228 distinctive TFBS. After combining the positive and negative sets, 453 hCTs contain 228 binding sites. Out of these 228 binding sites, 227 remain after the filtering, i.e. only one binding site appeared on one gene or less. Therefore, the final data set prepared for cluster and regression analyses contains 453 hCTs with 227 TFBS.

The cluster analysis was performed with three criteria combining similarity (S) percentage using (2), and the total number of appearances on the genes (n_4). The number of times these collinear motifs occurred at the same location on a gene was also counted. Table 2 shows the 47 collinear TFBS we combined to form 21 new TFBS. The percent of total genes, where they occurred at the same position, is found in the last column.

Most of the collinear motifs in Table 2 belong to the same motif family. This may be due to their very similar binding sequences/matrix, which will generate two motif binding sites for the same sequence. In fact, detailed analysis of those collinear binding sites indicates most of them do locate at the same sequence position. Using *GATA-1* and *GATA-2* as an example, Table 2 suggests they either both appear or both do not appear on 90-92.5% of all genes in the training sample, appear together on more than 15 hCTs, and when they appear together, they are found at the same location 83.8% of the time.

Alternatively, the cluster analysis has detected those collinear binding sites from different motif families, such as *MEIS* and *TGIF*, *CRE-BP1* and *vJun*, *Arnt* and *Max*, *RSRFC4* and *aMEF-2*. All of these motif pairs have been reported to cooperatively work together [16, 17, 18, 19]. You can see from Table 2 the percentage of time these pairs appear in the same location is much lower (0.0-42.9%) than the supposed motif families. This shows that our motif cluster method can be applied to identify cooperative binding motifs in a more general manner, as well as motifs belonging to the same family.

S	n_4	TFBS1	TFBS2	TFBS3	New TFBS	Occurred at same location on gene (%)[2]
90.0-92.5%	≥ 15	Egr-2	Egr-3	--	Egr_C	95.5%
		GATA-1	GATA-2	--	GATA_C	83.8%
		OCT-x	OCT1	--	OCT_C	42.1%
		STAT1	STAT3	--	STAT_C	71.8%
		NF-kappaB	NF-kappaB (p65)	c-Rel	NFkcR_C	77.9%
		CDP CR1	CDP CR3+HD	--	CDPCR_C	88.9%
		Zic1	Zic2	Zic3	Zic_C	94.4%
92.5-97.5%	≥ 10	CRE-BP1	v-Jun	--	CRvJun_C	0.0%
		Ik-1	Ik-2	Ik-3	Ik_C	63.2%
		Muscle initiator sequence-20	Muscle initiator sequence-19	--	Muscle_C	89.0%
		Arnt	Max	--	ArMx_C	0.0%
		MEIS1	TGIF	--	MsTGIF_C	5.6%
		HSF1	HSF2	--	HSF_C	71.4%
		HNF-4	HNF-4alpha1	--	HNF4_C	0.0%
		AR	AR (rat)	--	AR_C	45.5%
		AP-2alpha	AP-2gamma	--	AP2ag_C	95.6%
		FOXO1	FOXO4	Freac-2	FOXF_C	41.2%
		Tal-1alpha/E47	Tal11beta/E47	Tal-1beta/ITF-2	Tal1_C	86.7%
97.5-100%	≥ 5	IRF-1	IRF-2	--	IRF_C	87.5%
		STAT5A (homodimer)	STAT5B (homodimer)	--	STAT5_C	60.0%
		RSRFC4	aMEF-2	--	RSaMEF_C	42.9%

Table 2: Collinear binding sites combined to form a new variable

The data after the cluster analysis now contains 453 hCTs and 201 TFBS and the logistic regression analysis is applied. After the backward/forward selection procedures described in Section 2.5 are applied, the 16 main effects remaining in the final model are *ATF6, AhR, c-Ets-1(p54), c-Myc/Max, E2F, GATA_C, HSF_C, MEF2, Myogenin/NF-1, NFY, NFkcR_C, Spz1, STAT_C, STAT5_C, STAT5At, TATA*. The four interactions in the final model are [*c-ETs-1(p54)*GATA_C*], [*MEF2*STAT_C*], [*NFY*NFkcR_C*], [*NfkcR_C*STAT5At*], and [*STAT5At*TATA*]. There is also a significant three-way interaction: [*STAT5At*NFkcR_C*NFY*]. The final parameter estimates and standard errors are shown in the following SAS® output:

Parameter	Estimate	Standard Error
Intercept	-2.4154	0.2631
ATF6	3.1516	0.7471
AhR	-2.1887	0.9132
c-Ets-1 (p54)	1.6998	0.3863
c-Myc/Max	1.1200	0.3990
E2F	-1.3591	0.4489
GATA_C	-0.1321	0.5202
HSF_C	1.0521	0.5473
MEF2	-3.6921	1.4472
Myogenin/NF-1	2.0345	0.6718

[2] Collinear binding sites may be found at more than one location on a given gene.

Parameter	Estimate	Standard Error
NFY	-0.4989	0.6022
NFkcR_C	1.3382	0.4533
Spz1	-1.9450	0.6831
STAT_C	1.2222	0.4134
STAT5_C	-4.5466	2.1155
STAT5At	-5.5489	2.9048
TATA	0.9677	0.5738
c-ETs-1(p54)*GATA_C	-5.0556	1.5750
MEF2*STAT_C	3.7690	1.8587
NFY*NFkcR_C	-5.3897	2.0150
NFkcR_C*STAT5At	8.3482	2.8947
STAT5At*TATA	4.9338	2.2959
STAT5At*NFkcR_C*NFY	6.5811	2.7536

Logistic regression analysis can not only select the significant main effects (motifs) for the final model, but also detect the significant interactions among motifs. The final model indicates 5 pairs of positive interactions and a three- way interaction. *TATA* is the basic binding site for the RNA polymerase II, so its interaction with other motifs may indicate a general phenomena. *STAT5* and *NF-kappaB* have been reported to be cooperatively involved in the induction of *IL-6* in macrophage differentiation (Kawashima et al., 2001). Their significant interaction in our model implies cooperative roles in *Th* cell activation as well. The newly identified interaction between *MEF2* and *STAT_C*, and the three-way interaction need to be further validated experimentally to elucidate their potential roles in the *Th* cell activation process.

A goodness of fit test is conducted after the model of main effects is reduced from 201 predictors to 16:

$$(-2\log L_{(201)}) - (-2\log L_{(16)}) = 302.44 \sim \chi^2_{(185)}$$

This test yields a p-value>0.8. Since this *p*-value is not statistically significant the convergence criteria and fit is satisfied. The power also reduces for the same error rate when the number of variables is reduced, which is expected. Since the decrease is not dramatic (Figure 2), we can expect a relatively good model fit.

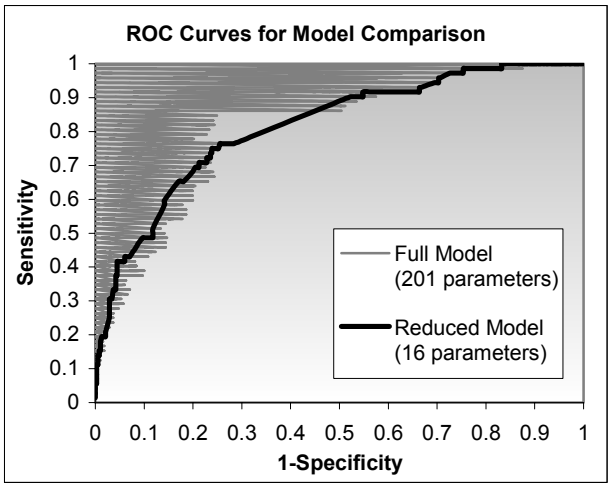

Figure 2: ROC curves to compare full and reduced models

In order to validate the results of the final model, a cutoff probability is estimated and used to calculate the final error rates. The final model chosen in the section above is used to predict the response (gene type). The predicted values are then used to create the following ROC curve (Figure 3).

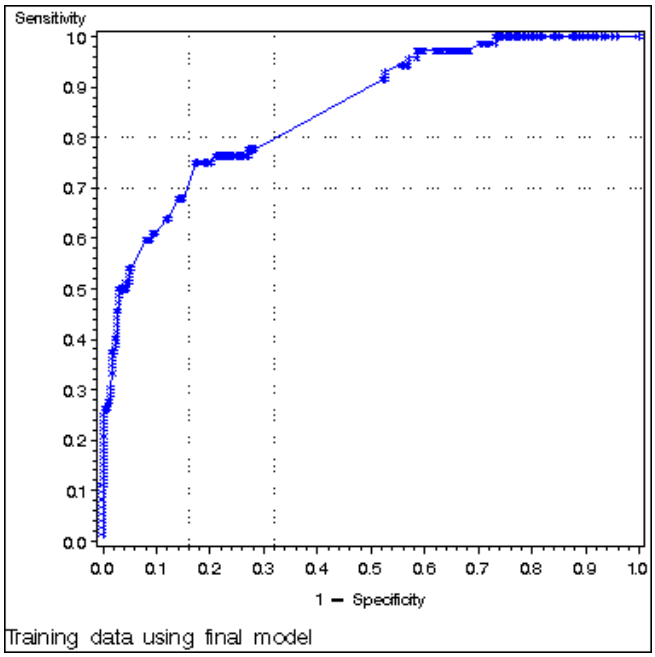

Figure 3: ROC curve for final training data model

The reference lines on the curve show a false positive rate between 0.16 and 0.32 yields sensitivity between 0.70 and 0.80. With that in mind, examining Figure 4, where the false positive rate and false negative rate are shown together, the error rates of 0.16 and 0.32 yield a probability level of 0.227. This value will be the cutoff value for the cross-validation analysis.

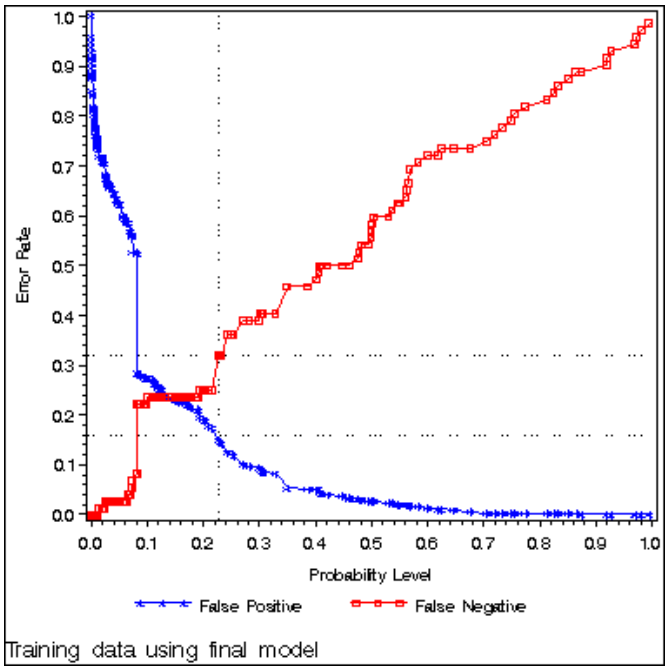

Figure 4: Error rates for final training data model

Once the predicted probabilities for the cross-validation are calculated in SAS® LOGISTIC, if the *P*(positive gene) exceeds 0.227, the response = 1 and likewise for response = 0 (see Table 3).

Predicted → Observed ↓	0	1	Total
0	309	72	381
1	30	42	72
Total	339	114	453

Table 3: Predicted vs. observed response for cross-validation

The table shows an 18.9% (72/381) Type I error (false positive) rate and 41.7% (30/72) Type II (false negative) error rate.

4. DISCUSSION

In order to implement efficient modeling methods for prediction and explanation of gene expression and TFBS data, given an established database of individual putative TFBS within genes, it is important to make use of a variety of statistical tools. The mixed model analysis, motif clustering, and logistic regression analyses described in this paper provide simple and alternative tools to accomplish this purpose. These methods allowed us to improve the quality of data for our analysis purposes, as well as provide insight into the data itself.

Logistic regression analysis of regulatory motifs has been applied previously for predicting tissue/signal specific transcription [4,20,21,22,23,24]. Most of these studies focused on binding sites known to be important to the transcription, while regression analysis enhanced the prediction by identifying the most significant sites. In our study, we have included all potential TFBS (total 227) in a human-mouse conserved segment for the initial dataset. In this dataset, redundancies exist for the same TFBS, caused by slightly different TRANSFAC matrix scanning. Transcription regulation in eukaryotes often involve the trans-regulatory action on a group of binding sites and there are only a few programs that have been developed to identify these cooperative binding sites, such as BioProspector [25] and CoBind [26]. Usually a collinearity analysis would discard redundant predictor variables to improve the model building procedures. Our cluster analysis effectively combines same or similar binding sites, so no information is discarded. Through cluster and logistic regression analyses, we have chosen TFBS, combinatorial pairs of TFBS, and the interactions that are considered significant for activated *T*-cell gene transcription.

Current experimental evidence suggests most *T*-cell activation-specific genes are regulated by the simultaneous binding of multiple transcription factors including members of *NF-kB/Rel/NF-AT*, *AP1*, *Ets1* and *CREB/ATF* families [27]. In the final model, the coefficients for *ATF6*, *c-Ets-1(p54)*, and *NFKcR_C* are all positive, which is also consistent with the existing knowledge of *T*-cell activation. As for *NF-AT* and *AP1*, the matrix used for scanning is very short and degenerated, consequently, they are not significantly represented in the regulated genes compared to controls in the data. Moreover, we may have identified some novel activated *T*-cell TFBS modules based on the strong significant effect of these sites in predicting the outcome. Nevertheless, all motifs in the final model are known to be immune related and reported to be critical for genes in the immune response. We believe our approach would be very useful for applications, where gene expression regulation is less understood and there is no prior information about significant binding sites.

5. ACKNOWLEDGMENTS

The authors would like to thank Li Liu, PhD, Arun Subramaniam, PhD, Asher Zilberstein, PhD, and Chang Hahn, PhD for their comments and contributions to this research. We would also like to thank Jay Saoud and Dennis Cosmatos, PhD for their support of this research.

6. REFERENCES

[1] Brazma, A., Vilo, J., Ukkonen, E., and Valtonen, K. Data mining for regulatory elements in yeast genome. In Proc. of 5th International Conference Intelligent Systems for Molecular Biology (ISMB 1997), AAAI Press, 65-74.

[2] Pilpel, Y., Sudarsanam, P. and Church, G. Identifying regulatory networks by combinatorial analysis of promotor elements. Nature Genetics (2001), 29, 153-159.

[3] Caselle, M., Di Cunto, F. and Provero, P. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. Bioinformatics (2002), 3(1), 7-19.

[4] Conlon, E., Liu, X., Lieb, J. and Liu, J. Integrating regulatory motif discovery and genome-wide expression analysis. PNAS (2003), 100(6), 3339-3344.

[5] Liu, X., Brutlag, D.L. and Liu, J.S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nature Biotechnology (2002), 20, 835-839.

[6] *Health on the Net Foundation* homepage: http://www.hon.ch.

[7] DeRisi, J.L., Iyer, V., and Brown, P. Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science (1997), 278(5338), 680-6.

[8] Quackenbush, J. Microarray data normalization and transformation. Nature Genetics (2002), 32 Suppl, 496-501. Review.

[9] Eickoff, B., Bernhard, K., Schick, M., Poustka, A. and van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. Nucleic Acids Res (1999), 27, e33.

[10] Richmond, C.S., Glasner, J., Mau, R., Jin, H. and Blattner, F. Genome-wide expression profiling in Escherichia Coli K-12. Nucleic Acid Research (1999), 27, 3821-3835.

[11] Beiβbarth, T., Fellenberg, K., Brors, B., Arribas-Prat, R., Boer, J.M., Hauser, N.C., Scheideler, M., Hoheisel, J.D., Schutz, G., Poustka, A. and Vingron, M. Processing and quality control of DNA array hybridization data. Bioinformatics (2000), 16, 1014-1022.

[12] Ge, N. Data pre-processing in microarray gene expression profiling. In Proceedings of ASA (2001). Section of Biopharmaceuticals.

[13] Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüß, M., Reuter, I., and Schacherer, F. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res (2000), 28, 316-319.

[14] Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E. and Kolchanov, N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucleic Acids Res (1995), 23, 4097-4103.

[15] Curran, M.D. Statistical modeling for genetics: Pharmacogenetics, molecular evolution and complex traits. DrPH thesis, Univeristy of North Carolina at Chapel Hill (2002).

[16] Katagiri M. and Kagawa N. The regulation of steroidogenesis by 17 alpha-hydroxylase/17,20-lyase (P450c17). Nippon Yakurigaku Zasshi (1998), 112(1), 43-50.

[17] Campbell, K.M. and Lumb, K.J. Structurally distinct mods of recognition of the KIX domain of CBP by Jun and CREB. Biochemistry (2002), 42(47), 13956-64.

[18] Swanson, H.I. and Yang, J.H. Specificity of DNA binding of the c-Myc/Max and ARNT/ARNT dimers at the CACGTG recognition site. Nucleic Acids Res (1999), 27(15), 3205-12.

[19] Chambers, A.E., Kotecha, S., Towers, N. and Mohun, T.J. Muscle-specific expression of SRF-related genes in the early embryo of Xenopus Laevis. EMBO J. (1992), 11(13), 4981-91.

[20] Wasserman, W. and Fickett, J. Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. (1998), 278, 167-181.

[21] Krivan W. and Wasserman, W. A predictive model for regulatory sequences directing liver-specific transcription. Genome Research (2001), 11, 1559-1566.

[22] Liu, R., McEachin, R.C., and States, D.J. Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. Genome Research (2003), 13(4), 654-61.

[23] Shelest, E., Kel A.E., Goessling E., and Wingender, E. Prediction of potential C/EBP/NF-kappaB composite elements using matrix-based search methods. In Silico Biol. (2003), 3(1-2), 71-9.

[24] Qiu, P., Qin, L., Sorrentino, R.P., Greene, J.R., Wang, L., and Partridge, N.C. Comparative promoter analysis and its application in analysis of PTH-regulated gene expression. J. Mol. Biol. (2003), 326(5), 1327-36.

[25] Liu, X., Brutlag, D.L. and Liu J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput (2001), 127-38.

[26] GuhaThakurta, D. and Stormo G.D. Identifying target sites for cooperatively binding factors. Bioinformatics (2001), 17(7), 608-21.

[27] Kuo, C.T. and Leiden, J.M. Transcriptional regulation of lymphocyte development and function. Annu. Rev. Immunol. (1999), 17, 149-87.

About the authors:

Marla D. Curran is a Senior Biostatistician at Aventis Pharmaceuticals. She currently conducts preclinical pharmaco-genomic research, as well as clinical research. She earned a DrPH from the Biostatistics Department at The University of North Carolina at Chapel Hill.

Hong Liu is a Principal Scientist at Aventis Pharmaceuticals. She currently conducts bioinformatics research in the Immunology area

Fan Long is a Scientist at Aventis Pharmaceuticals. She currently conducts bioinformatics research in the Immunology area

Nanxiang Ge is a Senior Principal Scientist at Aventis. He earned a PhD in Statistics from the University of Illinois at Urbana-Champaign and he is mainly interested in statistical genomics research.