Computational Methods for Preprocessing and Classifying Gene Expression Data- Survey

Ameer K. AL-Mashanji
University of Babylon
Hilla, Iraq
ameeruobabylon@gmail.com

Sura Z.AL-Rashi
University of Babylo
Hilla, Iraq
sura_os@itnet.uobabylon.edu.iq

***Abstract*— Microarray experiments generate data sets containing valuable information on the gene expression levels of millions of genes in the form of a set of biological samples. Inference of gene regulatory networks from microarray data has become an important research area in bioinformatics. Several computational methods have been proposed to infer important relationships between transcription factors with target genes from gene expression and transcription factor data sets. The inferences from these methods are consistent with the biological literature and can help researchers design reasonable and efficient drugs to resist different diseases. This paper is a survey of different methods used for preprocessing and inferring the gene regulatory networks from a data set of gene expression.**

Keywords*— Gene Expression, Transcription Factors (TFs), Gene Regulatory Networks, Data mining, Classification, Microarray Data Preprocessing,

I. INTRODUCTION

Bioinformatics is an interdisciplinary science that is interested with the collection, archiving, arranging and interpreting biological data. In another statement, it is considered as the implementation of computer techniques to the management of biological data. It can help to resolve more problems regarding biological field by using machine learning techniques and mathematical statistics [1][2]. The prediction processes are considered as significant processes in controlling many diseases, for example, cancer; precise prediction of the reasons of cancer will result in designing a strong and effective drug [3]. Various techniques have been proposed to emulate the behavior of gene regulatory networks such as Support Vector Machine (SVM), Bayesian Inference Model, Support Vector Regression (SVR) and Recurrent Neural Networks (RNNs). A first important factor in determining a good technique is the number of time series needed to infer the gene networks. Most of the algorithms require large data sets for inferring the regulatory network. The second factor is the noise in the gene expression data where some algorithms implement badly in the state of noisy data [4]. Microarray technique is typically a glass slide which considers a necessary tool that many biologists use to observe expression levels of genes in a specific organism. Therefore, Microarray provides researchers the chance to study the arranged behavior of genes and better conception the function of a gene in specific condition [5]. Microarray techniques are facing some challenges such as a high dimensional data problem which is considered as fundamental challenges in various datasets. It suffers from the redundant, irrelevant and noisy gene [6]. Gene selection process can be a resolution that may solve this significant problem, by reducing the number of genes [7]. Another challenge facing Microarray technique has been which experiments generated often contain multiple missing expression values [8]. Many algorithms which analyze data of gene expression that need a full data matrix as input, so that the missing values have to be estimated by using some proposed methods such as ROWaverage, KNNimpute, and LSimpute.

The rest of this paper is arranged as follows: Section II illustrates the concept of DNA microarray technology. Section III describes the preprocessing methods of microarray data. Section IV deals with microarray data classification techniques. Section V shows some conclusions.

II. DNA MICROARRAY TECHNOLOGY

Microarrays [9] are good techniques to observe the expression of many genes at the same time. The objective is to detect the set of genes that can pointedly appear particular disease cases or the genes that share common biological functions. Considerable information can be extracted from this set of genes by using data mining techniques. The final image is stored as a file for more analysis after applying microarray technique [10]. Many manufacturers of microarray slide supply their own software such as the puma package which is a tool to compute gene expression levels from raw image file data. This package is appropriate with Affy matrix Gene Chip data [11]. Microarray image file structure is shown in Fig.1.

978-1-7281-4425-2/19/$31.00 ©2019 IEEE

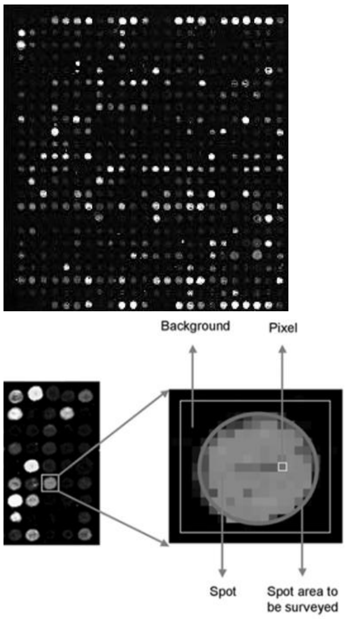

Fig .1 Microarray Image File Structure

The processed data from microarray image file can be represented in the shape of a matrix, often called a gene expression matrix that contains rows which express the genes and columns which express the special conditions [12]. Structure of the gene expression matrix is shown in Fig.2.

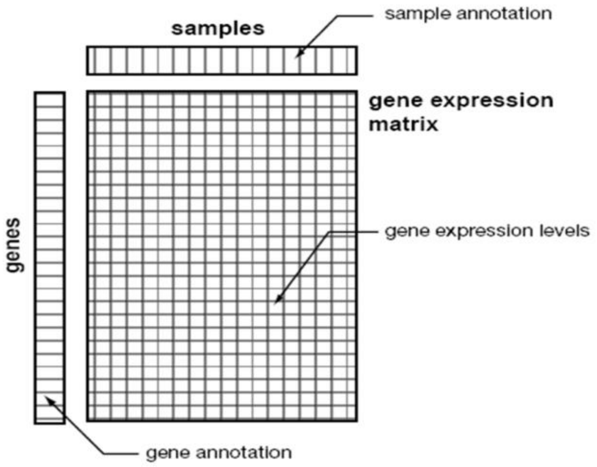

Fig .2 Gene expression matrix Structure

III. Microarray Data Preprocessing

Huge amounts of biological data are being generated via microarray technique. Data are usually characterized by an important proportion of missing values [13]. Some causes of missing values as below:

i. Corruption of image.

ii. Irregular spot.

iii. Dust and scratches in the image.

iv. Low intensity.

v. Insufficient resolution.

vi. Saturation.

vii. Spot variance.

In this survey, the most common methods for estimating missing values for gene expression data are described so that data will be ready to use for more analysis.

A. *Row Average*

Row Average is a traditional method that estimate the missing values in microarray data with the row average (gene vector). It does not take the correlation of data and feature that provided via the expression levels of other genes [14].

B. *KNNimpute algorithm*

The k nearest neighbor technique provides a way of estimating missing values for gene expression data sets. It is working to choose genes expression levels similar to the gene of interest to assign missing values [15]. The missing entry of data is estimated from neighboring genes and closeness between two genes is determined by using a proximity measures (e.g. Euclidean distance, Pearson correlation, correlation). The KNN technique is comparatively insensitive when the value of parameter K from 10-20 neighbors.

K-NN Imputation advantages: it doesn't need creating a predictive model for each feature with missing values in the data of gene expression. It is using the data correlation structure for estimating missing gene expression values. The disadvantages of this technique depend on the parameter k. It can be able to predict together, qualitative and quantitative features [13].

C. *LSimpute*

Simple linear regression is a method for the estimation of missing values in microarray data. It depends on the least-squares principle to estimate the missing values by using the correlations of genes and reducing the sum of squared errors [17]. The model of linear regression for y given x as

$$\hat{y} = \hat{\alpha} + \hat{\beta}\, x + e \tag{1}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\, \bar{x} \tag{2}$$

$$\hat{\beta} = \frac{Sxy}{Sxx} \tag{3}$$

$$Sxy = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \bar{x}\right)\left(y_j - \bar{y}\right) \tag{4}$$

$$Sxx = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2 \tag{5}$$

One of the other preprocessing tasks that is applied to gene expression data which are produced via microarray is data transformation. It is the implementation of mathematical function to each point in a data set so that the data appear to more closely with statistical inference [18]. It is especially valuable for many techniques for example, in the training phase will help to speed up the learning phase.

The most commonly procedures which are for transforming microarray DNA data, such as Log Transformation and Z-Score normalization, are reviewed.

D. Log Transformation

It is a procedure that takes the logarithm (base 2 or base 10) for values of the gene expression data. The logarithmic transformation of the most microarray gene expression data provides the best approximation of the normal distribution. Fig.3 shows an example of how a log transformation impact data distribution [19].

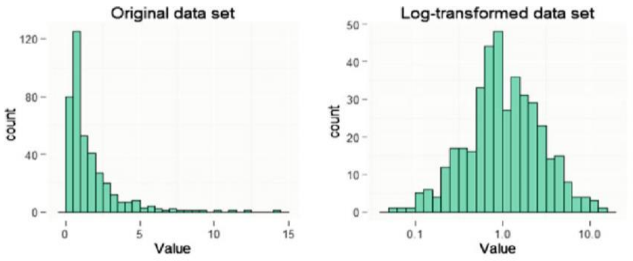

Fig. 3 Log Transformation measure

E. Z-Score normalization

Z-Score normalization is a common procedure for normalized the microarray gene expression data where the data are scaled so as the values fall within a specified domain [20]. For example, from −1.0 to 1.0, or from 0.0 to 1.0, as follows:

1)" transform the original expression data to the log scale given as:

$$\{ Ln(x1), Ln(x2),...,Ln(xs) \} \quad (6)$$

2)" The normalized of data is given as

$$\frac{\ln(x_i)-E}{V} \quad (7)$$

Where

$$E = \frac{1}{s}\sum_{j=1}^{s}\{\ln(x_i)\} \quad (8)$$

$$V = \sqrt{\frac{1}{S-1}\left\{\sum_{i=1}^{s}\ln(x_i)^2 - E^2\right\}} \quad (9)$$

Another important preprocessing task that is applied to gene expression data is gene selection. Thousands of genes are generated by using microarray technology. These genes may be irrelevant or redundant. With millions of genes, will appear the problem of high-dimensional samples which represents a challenge for learning most classification techniques. [21].

The methods of dimensionality reduction should eliminate genes that may be irrelevant and redundant. Selection of effective genes can help increasing the speed of subsequent tasks.

The most methods which use for genes selection from the DNA microarray data, such as Mutual information and T-Test will be reviewed

F. Mutual information

Several gene selection algorithms have used mutual information as an evaluation measure to eliminate the irrelevant and redundant gens. It is using the entropy and join-entropy of the random variables [22]. The mutual information of two random variables I and J can be defined according to the following equations.

$$MI\ (I, J) = H\ (I) + H\ (J) - H\ (I, J) \quad (10)$$

$$H(I) = -\sum_{i \in I} P(I)\ log\ p(j) \quad (11)$$

$$H(J) = -\sum_{j \in J} P(J)\ log\ p(J) \quad (12)$$

$$H(I,J) = -\sum_{i \in I}\sum_{j \in J} P(I,J)\ log(I,J) \quad (13)$$

Normalized mutual information ranges from 0 to 1, with the value 1 refers to that I and J are dependent this means high mutual while the value 0 refers to that I and J are independent this means low mutual [23].

G. T-Test

T-Test measure is one of the popular and effective ranking measures depending on the t-statistic between gene expressions and target class [24], is given as:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{S^2_1}{N_1} + \frac{S^2_1}{N_1}}} \quad (14)$$

Where

$$S^2 = \frac{\sum (x - \bar{x})^2}{N-1} \quad (15)$$

Where N_1 and N_2 refer to the number of samples in positive and negative classes, respectively, and $\overline{X_1}$ is the mean of the expressions in the positive class and $\overline{X_2}$ is the mean of the expressions in the negative class and S_1 is a standard deviation for positive class, S_2 standard deviation for negative class. A higher absolute T-Test indicates the higher importance of the gene for the prediction task [25]. Some preprocessing microarray data methods have been summarized in Table I.

TABLE I
SUMMARY OF THE SOME PREPROCESSING MICROARRAY DATA METHODS

S.No	Dataset used	Data Cleaning	Data Transformation	Genes Selection
[26]	Drosophila microarray time series.		Use Log2 Transformation	
[24]	C57BL/6J data set for mouse		Use Log2 Transformation	Applying T-Test
[25]	Breast Cancer data			Applying T-Test
[12]	GDS microarray data from the NCBI		Applying Z-Score	
[20]	GDS for Homo sapien from the NCBI		Applying Z-Score	

S.No	Dataset used	Data Cleaning	Data Transformatio n	Genes Selection
[27]	Time-course gene expression data	K nearest neighbor (KNN)	Use Log2 Transformation	Applying Mutual Information (MI)
[17]	NCI60 data set consists of 2069 genes and 64 simples	Row Average	Use Log2 Transformation	
	Lymphoma data set consists 2317 genes and 65 simples.	Simple Linear Regression (LSimpute)		
[22]	Breast cancer data set contains 7129 genes and 49 samples.		Use Log10 Transformation	Applying Mutual Information (MI)
	Colon cancer data set consist 2000 genes			

IV." Gene Expression Data Classification Techniques

There are many methods used for classifying microarray data. These methods use gene expression and TFs datasets to predict a set of transcription factors that cause such change gene expression and use them as a basis for classification of physiological or pathological processes. From the most used approaches are:

A. *Support Vector Machine (SVM)*

It is a supervised method to infer gene networks from the data of gene expression. It uses two of the data set as input. First, gene expression data for each gene in the organism second, list of known relationships between TFs and some genes. The primary goal is the suitable classification of unseen samples. SVM trains on known samples and then tests SVM on samples of unknown samples [28].

B. *Bayesian Inference*

Bayesian model is a method of statistical inference. It uses the principle of Bays theory to infer relationships between TFs protein and genes by using the datasets of gene expression and TFs when the gene regulatory network structure is known. The procedure is split into two stages. First; the training phase includes using training data to learn the model. The network structure is supposed to be known. Second, the prediction stage is applied to test data in order to identify the transcription factor for each gene [29].

C. *Support Vector Regression (SVR)*

SVR model uses the same basics as the SVM for classification. A result is a real number using in later analyzes. The data are divided into two data sets for training and testing. Where some of the genes are randomly selected to represent the training data and the rest represent the testing data. The model is trained through the use of training data and then applied to test data. The prediction precision is measured by using the Pearson correlation coefficient (PCC) between the expected expression values and expression values in the testing data [30].

D. *Recurrent Neural Networks (RNNs)*

Recurrent neural networks are the type of neural networks designed for capturing information from time-series data and can be thought of as several copies of the same network and with loops in them. A loop allows information to be passed from one step of the network to the next. The model of RNNs has a memory which takes information from previous steps. RNNs model emulates the topology of the gene regulatory network. The structure of RNNs is constructed to identify the regulatory interactions between genes and transcription factors from the microarray data [27].

TABLE II
SUMMARY OF THE SOME TECHNIQUES THAT USED TO INFER REGULATORY NETWORK FROM MICROARRAY DATA

S.No	Dataset used	Techniques and algorithms used	Result
[28]	Dataset expression profiles 500 genes	Support Vector Machine (SVM)	The results shown precision of prediction depends on the nature of the experimental condition and size of the network.
[29]	Artificial data set consists 649 genes and 19 transcription factors.	Bayesian Inference	The model is able to determine which transcription factors correctly regulate the target.
	Yeast cell cycle dataset consists of 6181 genes and 113 transcription factors.		
	Metabolic cycle dataset consists 3195 genes and 177 transcription factors .		
[30]	GSE8024 data set for mouse downloaded from the NCBI.	Support Vector Regression (SVR)	The model achieves a Pearson correlation coefficient of 0.77 between predicted and actual expression data
	RNA-seq data set for mouse cells that consist of 17560 genes.		

S.No	Dataset used	Techniques and algorithms used	Result
	The binding data TFs were downloaded from NCBI.		
[27]	Dataset of time course gene expression profiles .	Support Vector Machine (SVM)	The model achieves average test error 23.6% when used gene expression profile only. After combining the gene expression profile and binding sequence data as input, the test error is reduced to 15.8%.RNN is achieving 0.077 of The root mean square errors (RMSEs).
	Dataset of genome-wide location analysis.		
	Dataset of binding sequence from TRANSFAC database.	Recurrent Neural Network (RNN)	
	Gene ontology (GO) information dataset.		

V.ˮ CONCLUSION

This paper presents a survey of several methods that have been applied to microarray gene expression data. This study is showing that the KNNimpute method is more accurate and better than traditional row average method by taking a characteristic of the data correlation to estimate missing expression values. If we compare the estimation accuracy of the KNNimpute method with LSimpute method, it has been found the LSimpute method provides more accurate than KNNimpute. The comparison is shown in Table III.

TABLE III

COMPARISION BETWEEN KNNIMPUTE, LSIMPUTE, AND ROW AVERAGE METHODS

No	***Technique / Tool***	***Estimation Missing Values***	***S.No***
1	*LSimpute*	More accuracy	[17]
2	KNNimpute	Good accuracy	[27]
3	Row Average	Less accuracy	[14]

We recommend an LSimpute method for imputation of missing values. Also, we found Log Transformation procedure is mostly used for microarray data in order to transform data to facilitate the task on classification methods. It can reduce the effect of outliers and provide the best approximation for data to the normal distribution. Before and after using log transformation is shown in Fig. 4.

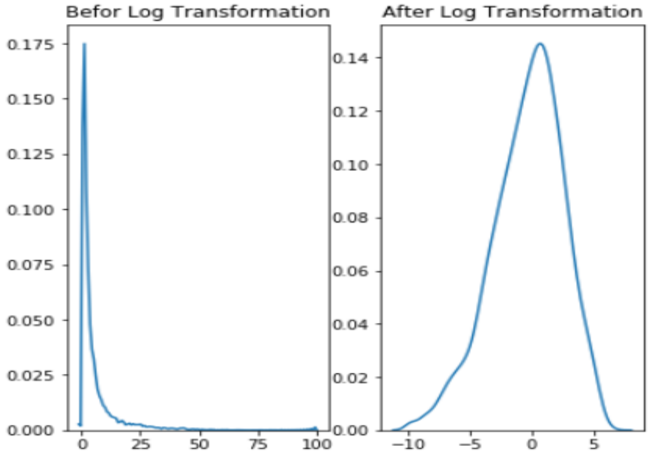

Fig. 4 The difference between before and after using Log transformation for a particular data set.

This survey demonstrates, most methods used to select genes related to the disease which are used the mutual information. Mutual information method is able to detect nonlinear relationships between gene expression data and insensitive to outlier data. So we recommend Mutual information to the gene selection process.

SVM in [28] demonstrates that the precision of an inference depends on the nature of experiential condition and size of the gene regulatory network. With the small size of the gene regulatory network, SVM exceeds unsupervised inference techniques. SVM is working better with the large size of the network.

Bayesian Inference technique in [29] illustrates that is able to infer regulatory relationships between several transcription factors and their target genes even from a short time series of gene expression with presence large amounts of genes.

SVR model achieves a Pearson correlation coefficient of 0.77 between predicted and actual expression data. The accuracy of prediction is getting better to 0.78 when used Two-Layers of SVR model, this demonstrates in [30].

When SVM used gene expression profile alone as input at [27], the average test error was 23.6%. After combining gene expression profile and binding sequence data as input, test error is reduced to 15.8%.

We conclude that SVM performs good performance with the small and large size of the gene regulatory network. Bayesian Inference model is good when gene expression data have a limited number of time series and a huge number of genes. Also, the prediction accuracy increases significantly when used several data sources as inputs, thus reducing the error factor.

REFERENCES

[1] K. Stankov, “Bioinformatic tools for cancer geneticists,” Arch. Oncol., vol. 13, no. 2, pp. 69–75, 2005.

[2] M. R. Barnes and I. C. Gray, Bioinformatics for geneticists. John Wiley & Sons, 2003.

[3] Y. Lu and J. Han, “Cancer classification using gene expression data,” Inf. Syst., vol. 28, no. 4, pp. 243–268, 2003.

[4] F. Rafii, M. A. Kbir, and B. D. R. Hassani, “Microarray Data Preprocessing To Improve Exploration on Biological Databases,”

in International Conference on Big Data, Cloud and Applications, Tetuan, Morocco, 2015, pp. 25–26.

[5] T. Schlitt and P. Kemmeren, "From microarray data to results: Workshop on Genomic Approaches to Microarray Data Analysis," EMBO Rep., vol. 5, no. 5, pp. 459–463, 2004.

[6] P. A. Mundra and J. C. Rajapakse, "Gene and sample selection using T-score with sample selection," J. Biomed. Inform., vol. 59, pp. 31–41, 2016.

[7] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data." 2013.

[8] A. W.-C. Liew, N.-F. Law, and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," Brief. Bioinform., vol. 12, no. 5, pp. 498–513, 2010.

[9] M. Dufva, "Introduction to microarray technology," in DNA Microarrays for Biomedical Research, Springer, 2009, pp. 1–22.

[10] M. M. Babu, "Introduction to microarray data analysis," Comput. genomics Theory Appl., vol. 17, no. 6, pp. 225–249, 2004.

[11] R. D. Pearson, X. Liu, G. Sanguinetti, M. Milo, N. D. Lawrence, and M. Rattray, "puma: a Bioconductor package for propagating uncertainty in microarray analysis," BMC Bioinformatics, vol. 10, no. 1, p. 211, 2009.

[12] Y. Li, W. Liu, Y. Jia, and H. Dong, "AWeighted Mutual Information Biclustering Algorithm for Gene Expression Data.," Comput. Sci. Inf. Syst., vol. 14, no. 3, 2017.

[13] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," Am. J. Obstet. Gynecol., vol. 195, no. 2, pp. 373–388, 2006.

[14] T. Aittokallio, "Dealing with missing values in large-scale studies: microarray data imputation and beyond," Brief. Bioinform., vol. 11, no. 2, pp. 253–264, 2009.

[15] H. De Silva and A. S. Perera, "Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data," in Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on, 2016, pp. 141–146.

[16] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.

[17] T. H. Bø, B. Dysvik, and I. Jonassen, "LSimpute: accurate estimation of missing values in microarray data with least squares methods," Nucleic Acids Res., vol. 32, no. 3, pp. e34–e34, 2004.

[18] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[19] J. Quackenbush, "Microarray data normalization and transformation," Nat. Genet., vol. 32, p. 496, 2002.

[20] M. Inoue and K. Horimoto, "Relationship between regulatory pattern of gene expression level and gene function," PLoS One, vol. 12, no. 5, p. e0177430, 2017.

[21] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in Advances in neural information processing systems, 2001, pp. 668–674.

[22] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," BMC Bioinformatics, vol. 6, no. 1, p. 76, 2005.

[23] S.-B. Guo, M. R. Lyu, and T.-M. Lok, "Gene selection based on mutual information for the classification of multi-class cancer," in International Conference on Intelligent Computing, 2006, pp. 454–463.

[24] K. Dimitrakopoulou et al., "Dynamic gene network reconstruction from gene expression data in mice after influenza A (H1N1) infection," J. Clin. Bioinforma., vol. 1, no. 1, p. 27, 2011.

[25] P. A. Mundra and J. C. Rajapakse, "Gene and sample selection using T-score with sample selection," J. Biomed. Inform., vol. 59, pp. 31–41, 2016.

[26] A. Honkela et al., "Model-based method for transcription factor target identification with limited data," Proc. Natl. Acad. Sci., 2010.

[27] Y. Zhang, J. Xuan, B. G. de los Reyes, R. Clarke, and H. W. Ressom, "Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data," BMC Bioinformatics, vol. 9, no. 1, p. 203, 2008.

[28] Z. Gillani, M. S. H. Akash, M. D. M. Rahaman, and M. Chen, "CompareSVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks," BMC Bioinformatics, vol. 15, no. 1, p. 395, 2014.

[29] M. K. Titsias, A. Honkela, N. D. Lawrence, and M. Rattray, "Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison," BMC Syst. Biol., vol. 6, no. 1, p. 53, 2012.

[30] C. Cheng and M. Gerstein, "Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells," Nucleic Acids Res., vol. 40, no. 2, pp. 553–568, 2011.