Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Correlated Discretized Expression Score: A Method for Identifying Gene Interaction Networks from Time Course Microarray Expression Data

Peter Larsen, Eyad Almasri, Guanrao Chen, Yang Dai*

Abstract— **One of the goals of genomic expression analysis is to construct gene interaction networks from microarray data. Time course microarray data is a common place to seek causal relationships between the expression of a regulator and its effect on the expression of its targets. By proposing gene expression patterns of regulator and target genes based on biological expectation of regulatory interactions, it is possible to propose a system to identify these patterns. This system is based on the Correlated Discretized Expression (CDE) score calculated from microarray time course data. The CDE-score is derived by discretizing microarray data to identify significant gene expression changes. The usefulness of this method is demonstrated using a set of hypothetical gene expression data and the analysis of S. cerevisiae cell cycle microarray data.**

I. INTRODUCTION

MICROARRAYS are routinely used to simultaneously assess the relative expression levels of many thousands of gene transcripts in biological samples. Increasingly sophisticated techniques have been developed to identify those transcripts among thousands whose expression have significantly changed in response to the selected experimental conditions. Yet it is through development of gene interaction networks, microarray data evolves from being descriptive to being a predictive tool. Time course microarray data is particularly appealing to identify such gene interaction networks as by following changes in gene expression over time allow the identification of causal relationships between a change in expression of a regulator and a subsequent change in expression of a target.

A number of methods for analysis of time course microarray data have been developed since Cho et al. classified 421 genes as periodic from visual inspection of microarray data [1]. Spellman, et al used a Fourier-like score computed for every gene [2]. Johansson et al. used partial least squares regression to fit data to sine curves [3]. Luan and Li followed a similar methodology, based on cubic splines rather than sine waves [4]. Zhao et al. proposed a statistical pulse model [5]. Lu et al. and Filkov et al. both proposed Bayesian modeling techniques [6, 7]. In this study, a specific biological expectation for gene interaction pairs is defined and a method for seeking these biologically relevant patterns is proposed.

A gene network, for the purpose of this study, is the sum of all gene interaction pairs in a system working together to drive some biological process. A gene interaction pair here is defined as a regulator gene that, through a change in expression, modulates the expression of a target gene. To identify such interactions, it is useful to consider time course microarray data to observe this causality in the expression data. To measure this form of gene interaction, we propose the Discretized Correlation Expression (CDE) score. Using expression data from a microarray experiment, CDE-scores can be calculated between every possible gene pair in an experimental set. A large, positive CDE-score suggests a positive regulatory interaction between the regulator and the target. A large, negative CDE-score suggests an inhibitory reaction. A CDE-score threshold is applied and every gene pair interaction with a CDE-score above the selected threshold is considered to be a part of the gene interaction network active in a system. Additional restrictions based on relevant biological information is used for further refine CDE-score derived gene networks.

The main components of the study are the following:

(1) Using hypothetical gene expression data, the CDE-score is calculated and demonstrated to be superior at identifying the proposed examples of regulator-target gene interaction pairs than the Pearson correlation coefficient method.

(2) The CDE-score method will be used with actual microarray data and compared to results from Pearson Correlation coefficient. A subset of a yeast cell cycle time course microarray experiment had been selected for this purpose.

(3) Gene interaction networks determined by CDE-score thresholds will be further refined with the addition of relevant biological information. For this purpose, information gathered from the published literature will be used to restrict potential gene interaction pairs to those that most resemble previously identified gene interaction pairs, Gene Ontology (GO) biological process

Manuscript received April 3, 2006. This work was supported in part by the National Science Foundation (EIA-022-0301) and Naval Research Laboratory (N00173-03-1-G016).

Peter Larsen is with the Core Genomics Laboratory at University of Illinois at Chicago, Chicago, University of Illinois at Chicago, 845 W Taylor St Chicago, IL 60607-7058, USA (e-mail: plarsen@uic.edu).

E. Almasri is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: ealmas1@uic.edu).

G. Cheng is with the Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607 USA. (e-mail: gchen4@uic.edu).

*Y. Dai is with the Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607 USA. (phone: 312-413-1487; fax: 312-413-2018; e-mail: yangdai@ uic.edu). Corresponding Author.

annotation, and identity of known transcription factors will be used.

II. CDE-score Calculation

A. Hypothetical Expression Data

To illustrate possible patterns of gene interactions, five hypothetical time course gene expression profiles are proposed (Fig. 1.A). The gene 'Regulator' shows a rise in expression, peaking at time 2. Hypothetical target genes are proposed. 'Target1' has an expression profile identical to 'Regulator', offset by one time unit. 'Target2' gene shows a much larger increase in expression, as might be expected of a gene's response to its transcription factor. 'Target3' shows a gene responding to an increase in 'Regulator' expression, though the response of 'Target3' persists after the increase in expression of 'Regulator' ends, as might be the case if the protein product of gene 'Regulator' remains active for some time after the gene's expression peaks. 'Target4' shows a gene whose expression is repressed by the 'Regulator' gene. The gene 'Unrelated' consists of randomly generated values from –2 to 2 and models a gene that may be significantly altered in expression, but without relationship to the expression of the 'Regulator' gene.

B. Pearson Correlation Coefficient

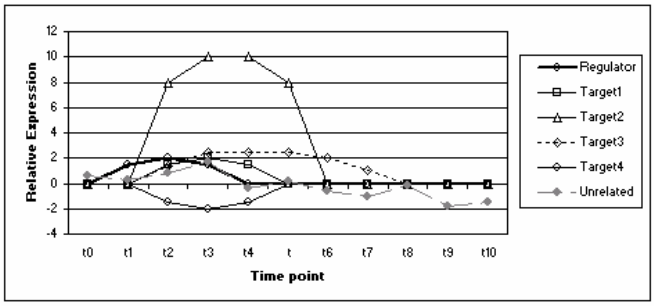
For the yeast dataset of 102 genes, a matrix CDE-scores was calculated for all 10302 possible gene pairs. To discretize the data, a gene was considered to be increased (1) in expression if the expression value was greater than one standard deviation above the gene's average signal across all time points, considered decreased (-1) if its expression value was one standard deviation below the gene's average signal across all time points, and considered unchanged (0) otherwise.

C. Correlation Discretized Expression Score

CDE-score, a novel method for scoring correlation of time course data, is proposed here to address the inability of correlation coefficients to accurately capture regulation interactions that follow biological expectations. The microarray data is first discretized into one of three values: 0 if there is no significant expression change, 1 if there is a significant increase in expression, and –1 if there is a significant decrease in expression. Given a time series of measurements of discretized data for a regulator and a target, the CDE-score of the possible relationship between regulator and target is:

$$CDE_{scoreR,T} = \sum_{t=1}^{t=t_{\max}-3} (R_t * T_{t+1}) + 0.5*(R_t * T_{t+2}) + 0.33(R_t * T_{t+3}) \quad (1)$$

where *Rt* is the discretized expression of the regulator at time *t, Tt* is the discretized expression of the target at time *t*, and *tmax* is the length of the time series. Coefficients in this equation were determined empirically. A large CDE-score suggests a positive regulation of Regulator to Target.



	t0	t1	t2	t3	t4	t	t6	t7	t8	t9	t10	Pearson	CDE-Score
Regulator	0	1.5	2	1.5	0	0	0	0	0	0	0		
Target1	0	0	1.5	2	1.5	0	0	0	0	0	0	**1.000**	**4.3**
Target2	0	0	8	10	10	8	0	0	0	0	0	**0.841**	**5.1**
Target3	0	0	1.4	2.5	2.5	2.5	2	1	0	0	0	**0.599**	**5.4**
Target4	0	0	-1.5	-2	-1.5	0	0	0	0	0	0	**-1.000**	**-4.3**
Unrelated	0.59	0.26	0.79	1.73	-0.4	0.21	-0.6	-1.1	-0.1	-1.8	-1.4	**0.669**	**1.5**

Fig. 1. (A) A series of hypothetical expression patterns of a 'Regulator' gene, four 'Targets' of the regulator, and a significantly differentially expressed 'Unrelated' gene not regulated by 'Regulator'. (B) Relative expression data of hypothetical genes for times 't0' to 't10'. Highlights indicate discretized values of expression data, where light gray with black text indicates a significant increase in expression discretized as 1 and dark gray with white text indicates a significant decrease in expression discretized as –1. 'Pearson' correlation coefficient calculated between 'Regulator' and all other gene expression data. Calculated 'CDE-score' between 'Regulator' and all other genes are calculated.

A negative CDE-score suggests a negative regulation or inhibition of Target by Regulator. By ranking CDE-scores between all possible gene-pair interactions, relative likelihood of a regulatory relationship between genes can be determined.

Returning to the hypothetical expression profiles, the expression values can be discretized and CDE-scores for 'Regulator' and target genes can be calculated (Fig. 1.B). By CDE scoring method, the randomly expressed gene has the lowest CDE-score, 'Target2' and 'Target3' have the highest CDE-scores, slightly ahead of the perfectly-matched, but perhaps biologically less interesting, expression profile of 'Target1'. 'Target4', the example of negative regulation, is also clearly identified by its negative score. For the hypothetical data, CDE-score performs better than Pearson correlation for detecting gene interactions.

III. Analysis of Yeast Cell Cycle Microarray Data

A. Dataset

The dataset used in this study is a 102-gene subset of the budding yeast *Saccharomyces cerevisiae* microarray experiments [2]. In this experiment, the gene expressions of alpha-factor cell cycle synchronized yeast cultures were collected over 18 time points taken at 7-minute intervals. The subset of 102 genes was selected to include 10 known transcription factors and their possible regulation targets.

In order to test the accuracy of the method proposed here, it is necessary to have some expectations as to the true gene interaction network. However, the 'true' gene network is not known and depends upon the chosen definition of gene interaction. For this study, the 'true' interactions were derived from examples in the published literature as identified by the bioinformatics tool, PathwayAssist 3.0 [9]. It should be noted that this is not a so-called 'golden standard' set for a true evaluation of the learning outcome. Nevertheless, this list of previously published gene interaction pairs can be reasonably considered to be 'true' gene interactions in this dataset. One hundred seventy one 'true' gene interaction pairs were found in the 102-gene subset by this method.

B. CDE-score Calculation for Microarray Data

Pearson correlation coefficient is a common way of measuring similarity between two gene expression profiles in a time series [8]. A Pearson correlation is close to one when there is good correlation between time series. Correlation is near negative one when there is a negative correlation, as might be the case for a regulator acting as an inhibitor of expression of a target. A correlation close to zero indicates that no correlation between expressions is observed. A Pearson correlation of gene expression values is calculated for 'Regulator' expression times 0 through 9, and all other expression profiles times 1 through 10 (Fig. 1.B). The matching expression profiles of 'Regulator' and 'Target1' have an excellent correlation coefficient, but this relationship of perfectly matched expression changes is not necessarily the most biologically interesting relationship between genes. Pearson's correlation performs only moderately well for the high-expressing 'Target2'. The random expression of 'Unrelated' has a higher correlation coefficient than the persistent expression of 'Target3', indicating a potential false positive and false negative. 'Target4', the example of negative regulation, is well identified by both Pearson similar to the example of positive regulation in 'Target1'.

C. Restriction of Possible Networks with Additional Biological Data

In the analysis of new microarray experimental data, the goal is rarely to solely generate networks of possible gene interactions, but to address some specific hypothesis in its biological context. Additional information can be applied to the analysis of these results, screening the list of expressed genes to eliminate genes not likely involved in the biological phenomenon studied in the experiment.

Using prior information in the PubMed database of scientific publications, information about previously observed gene interactions will be collected and used to generate a Likelihood of Interaction LOI-score [10]. If the new gene pair closely resembles gene interaction pairs frequently observed in the literature, it will be considered likely and have a high LOI-score. Gene pairs with a LOI-score greater than nine will be considered possible interaction pairs of the CDE-score determined gene interaction network.

Gene Ontology (GO) annotation, is a structured vocabulary for describing gene products [11] Of the 102 genes in the set, 78 have GO annotations of "cell cycle" according to the information in the Saccharomyces Genome Database SGD Gene Ontology Slim Mapper [12]. Potential regulators were limited to these 78 genes.

Of the 102 genes in the dataset, ten are known to be transcription factors [13]. The data was analyzed by CDE-score threshold restricting possible regulators to these ten transcription regulator genes.

This biological information can be combined to further restrict the set of allowed gene interaction pairs. LOI-score with transcription factor; LOI-score with "cell cycle" GO annotation; and LOI-score with transcription factors and "cell cycle' annotation" were used to restrict CDE-score determined networks.

D. F-score for Comparison of Networks and Selection of Thresholds

The metric selected for evaluating the proposed method is the F-score. The F-score combines the accuracy and the precision in the equation:

F-score = (2 * accuracy * precision) / (accuracy + precision) (2)

where precision is defined as the number of predicted previously published interactions over the total number of interactions predicted. Accuracy is defined as the number of predicted previously published interactions over the total number of published gene interactions.

CDE-score thresholds were selected by testing every threshold across the range of possible scores. The threshold that yielded network with the highest F-score was taken

IV. Results

Table I summarizes the F-scores of gene networks determined by CDE-score thresholds, and using additional restrictions from a gene pair LOI-score, cell cycle GO annotation, and known transcription factors. CDE-score alone produces the least accurate gene interaction networks with an F-score of 0.055. The addition of each form of relevant biological information improves the calculated networks markedly. The best gene interaction network uses information from GO annotation of cell cycle and known transcription factors to restrict potential gene interaction pairs from CDE-score thresholds, having an precision of 0.293, and accuracy of 0.380, and an over all F-score of 0.331.

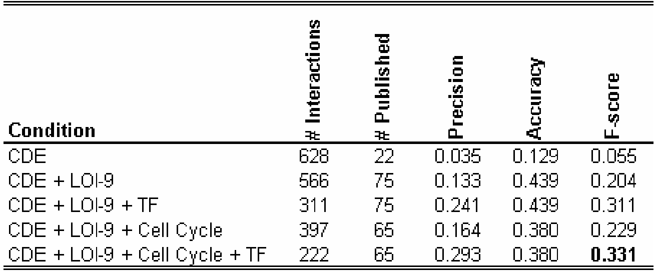
V. Discussion

A method for identifying gene interaction networks from time course microarray data has been shown. Proposed Correlated Discretized Expression (CDE) scoring metric was shown to be superior to Pearson correlation coefficient for identifying genes with correlated expression in the yeast cell cycle microarray dataset used for this study. Inclusion of additional, relevant biological data in the form of LOI-score, GO biological process annotation, or identity of known transcription factors increases the accuracy and precision of this method. The most favorable conditions of this method had a precision and accuracy of about 30%.

Though, very likely, many identified gene interactions of this method that are not represented in the literature are false positives, there are some reasons that can explain these results.

TABLE I

CDE-score Derived Gene Networks



Condition	# Interactions	# Published	Precision	Accuracy	F-score
CDE	628	22	0.035	0.129	0.055
CDE + LOI-9	566	75	0.133	0.439	0.204
CDE + LOI-9 + TF	311	75	0.241	0.439	0.311
CDE + LOI-9 + Cell Cycle	397	65	0.164	0.380	0.229
CDE + LOI-9 + Cell Cycle + TF	222	65	0.293	0.380	**0.331**

Gene networks with the best F-score by CDE-score threshold. 'LOI-9' indicates that network is restricted to include only gene interaction pairs with an LOI-score greater than 9. 'Cell Cycle' indicates that potential regulators of gene interaction pairs are restricted to genes with the Gene Ontology annotation of cell cycle. 'TF' indicates that potential regulators of gene interaction pairs are restricted to known transcription factors.

The dataset used was comprised of only 102 genes from the Saccharomyces cerevisiae genome of many thousands of genes. A chain of interactions, for example gene 'A' influences gene 'B', and gene 'B' influences gene 'C', might appear as gene 'A' directly influencing gene 'C' if gene 'B' is omitted from the dataset. This would not be a false positive, but accurate representation of a network that extends outside of the given set. An additional reason for identification of gene interaction pairs not predicted by previous experimental data is that the system may be detecting genuine, novel gene interactions observable in this dataset. Any potentially novel observations may be supported through additional analysis of the suggested gene interaction pairs and by seeking similar interactions in other cell cycle experiments or other model systems.

The method proposed here is computationally inexpensive and could easily be scaled up to accommodate much larger datasets, while more mathematically intensive procedures require large amounts of processing time. Alternatively, the method here might be used as an efficient pre-screening of data, limiting the dataset to a smaller set of high biological relevance before the data is given over to analysis by more computationally rigorous methodologies.

VI. Conclusion

The Saccharomyces cerevisiae cell cycle microarray experiments are appealing for development of analysis methods as this is a well-studied system with a large body of prior microarray data available for analysis. However, the ultimate goal of a method as is proposed here is to build a robust method capable of identifying novel gene interaction pairs from data of other, less well characterized systems of greater biological and scientific interest. Techniques here for identifying likely gene interaction pairs based on their molecular function can be extrapolated beyond yeast cell cycle. Certainly, the ability to select likely from unlikely gene interactions from a large search space of possibilities by their gene characteristics would be valuable in many systems. Methods for finding patterns in gene expression in time course data could be applied to other types of series, such as drug dosages or disease progression

Acknowledgment

Special thanks to Dr. Min Zou for providing list of genes for yeast 102-gene subset.

References

[1] Raymond J. Cho, Michael Campbell, Elizabeth A. Winzeler, Lars Steinmetz, Andrew Conway, Lisa Wodicka, Tyra G. Wolfsberg, Andrei E. Gabrielian, David Landsman, David J. Lockhart, and Ronald W. Davis. A Genomic-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell, 2:65-73, 1998.

[2] Paul T. Spellman, Gabin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell, 9:3273-3297, 1998.

[3] D. Johanasson, P. Lindegren, and A. Berglund. A Multivariate Approach Applied to Microarray Data for Identification of Genes With Cell Cycle-Coupled Transcription. Bioinformatics, 19:467-473, 2003.

[4] Y. Luan and H. Li. Model-Based Methods for Identifying Periodically Expressed Genes Based on Time Course Microarray Gene Expression Data. Bioinformatics, 20:332-339, 2004.

[5] L. P. Zhao, R. Prentice, and L. Breeden. Statistical Modeling of Large Microarray Data Sets to Identify Stimulus-Response Profiles. Proc. Natl Acad, Sci. USA, 98:5631-5636.

[6] X. Lu, W. Zhang, Z. S. Qin, K. E. Kwast, and J. S. Liu. Statistical Resynchronization and Bayesian Detection of Periodically Expressed Genes. Nucleic Acids Res, 32:447-455, 2004.

[7] Vladmir Filkov, Steven Skiena, and Jizu Zhi. Analysis Techniques for Microarray Time-Series Data. Journal of Computational Biology, 9:317-330, 2002.

[8] M. B. Eisen, P.T. Spellman, P.O. Brown, amd D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc. Natl. Acad. Sci., 85:14863-14868, 1998.

[9] Svetlana Novichkova, Sergei Egorov, and Nikolai Daraselia. MedScan, a Natural Language Processing for MEDLINE Abstracts. Bioinformatics, 19:1699-1706, 2003.

[10] Peter Larsen, Eyad Almasri, Guanrao Chen, and Yang Dai. Gene Ontology Improves Prediction of Genetic Interactions Based on Bayesian Network Approach. Unpublished.

[11] Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Genome Research, 11:1425-1433, 2001.

[12] http://db.yeastgenome.org/cgi-bin/GO/goTermMapper

[13] Lee, Tong Ihn, Nicola J. Rinaldi, Francois Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Tompson, Itamar Simon, Julia Zeitlinger, Erza G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Earnest Fraenkel, David K. Gifford, Richard A. Young. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science, 25:799-804, 2002.