2014 IEEE 27th International Symposium on Computer-Based Medical Systems

Discovering Differentially Expressed Genes in Yeast Stress Data

António Gonçalves*, Irene Ong†, Jeffrey A. Lewis‡ and Vítor Santos Costa*
*CRACS INESC-TEC and Department of Computer Science
Faculty of Sciences, Universidade do Porto, Portugal 4169-007
Email: {up201008720@alunos.dcc.fc.up.pt,vsc@dcc.fc.up.pt}
†Great Lakes Bioenergy Research Center
University of Wisconsin, Madison, WI 53706
Email: ong@cs.wisc.edu
‡Department of Biological Science
University of Arkansas, Fayetteville, AR 72701
Email: lewisja@uark.edu

***Abstract*—Transcriptional regulation plays an important role in every cellular decision. Gaining an understanding of the dynamics that govern how a cell will respond to diverse environmental cues is difficult using intuition alone. We try to discover how genes interact when submitted to stress by exploring techniques of gene expression data analysis. We use several types of data, including high-throughput data. These results will help us recreate plausible regulatory networks by using a probabilistic logical model. Hence, network hypotheses can be generated from existing gene expression data for use by experimental biologists.**

***Keywords*-Bioinformatics; Gene Regulation; Genomics; Data Analysis;**

I. Introduction

With the advent of high-throughput technologies and advanced measurement techniques, molecular biologists and biochemists are rapidly identifying components of gene networks and determining their biochemical activities. Understanding how these complex multicomponent networks govern how a cell will respond to diverse environmental cues is difficult using intuition alone. Towards this goal in this work: we want to validate prior work on genes involved in stress pathways [1], and we want to extend these results with information on non-coding genes.

In this work we used two types of data and applied two different methods to them: clustering based approaches/heatmaps and also genes differential expression (genes responding to external factors). The first method groups genes that have similar behavior/reaction and the second method allows us to see which genes respond to environmental changes. We are looking for pairs of coding/non-coding genes that are connected to stress functions/response. Discovering these pairs may lead to understanding how non-coding genes interact with the stress pathways.

II. Experimental Methodology

We used 2 types of data, microarray data [2] and RNA-seq data. Both datasets had gene expression data in which the genes were subjected to several types of stress. In the microarray data, the genes were exposed to *NaCl* (salt) stress, *EtOH* (ethanol) stress and *H2O2* (hydrogen peroxide) stress, as for the RNA-seq the genes were submitted to *heat* stress and *EtOH* stress.

First, we performed differential expression analysis to our data by using *edgeR* [3] with the RNA-seq data and *Limma* [4] with the microarray data.

Second, using our differential expression analysis we obtained 2708 genes with a p-value < 0.05 for the RNA-seq data, as for the microarray data we obtained 204 genes for the salt stress, 4697 genes for the ethanol stress and 4404 genes for the hydrogen peroxide stress, also with a p-value < 0.05.

Third, we intersected (Fig.2) the microarray differentially expressed genes in order to verify how many common genes we would find in all 3 stress situations.

Fourth, we generate all the possible pairs of coding and non-coding genes for the 2 data types, giving us a total of 1804660 pairs for the RNA-seq data and 2268 pair for the microarray data. As the number of pairs for the RNA-seq data is very high we applied filters (the results can be seen in Table I) to reduce the number of pairs. A threshold of a minimum correlation of 0.7 (absolute value) was defined. Both types of correlation are interesting, but if a gene has a very high negative correlation with another gene, we may assume that it may be repressing it or even overlapping it anti-sense.

By adding the non-coding information to the data obtained from [1] provided us with a better insight how genes are expressed differentially and will help us in our work by relating them to the coding genes and providing pairs of coding/non-coding genes. We are looking for non-coding genes that are highly negatively correlated with a coding gene, this will lead to information on non-coding genes that have a very high probability of being involved in stress pathways. In Table II we have the number of differentially expressed coding and non-coding genes for each of the

1063-7125/14 $31.00 © 2014 IEEE
DOI 10.1109/CBMS.2014.127

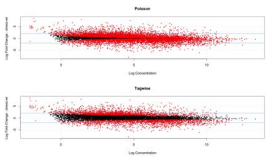

Figure 1. MA plot of the relationship between concentration and fold-change across the gene (RNA-seq data).

Figure 2. Number of genes common to all stress situations.

conditions yeast was submitted to, the middle column *ESR* is the number of genes that are known to react to external stress responses, these are only for the coding genes.

III. Discussion and Future Work

Using this type of analysis we were able to get interesting results regarding coding and non-coding genes that are involved in stress response. The low number of genes for the salt stress may be due to the fact that the experiments were done using a different methodology.

We can verify that we have a different number of pairs on both types of data, this is due to the initial intersection on the microarray data. Even so, we submitted our resulting pairs after the threshold filters to a verification in order to check if the coding gene of each pair is a known *ESR* (external stress response) gene and were able to obtain 236137 pairs for the RNA-seq data and 1197 pairs for the microarray data.

As a final step we checked the pairs for overlaps between the coding and non-coding gene and verified that in RNA-seq 184 pairs overlap and in microarray only 3 pairs overlap, this last value is certainly due to the intersection of the common genes, which led to a large cut in the possible pairs. From this resulting number of pairs we chose 22 RNA-seq pairs that seem good candidates, as they are highly negatively correlated (< -0.7) and overlap antisense, from microarray data we chose 2 pairs that seem good candidates, they are highly negatively correlated (< -0.7) in all 3 experiments and overlap antisense.

With the latest experiments, we found 29 genes (25 coding and 4 non coding) that are common to all conditions for both the RNA-seq data and the microarray data. Comparing this data with the processed data from [1] we found 4491 genes (2583 coding and 1908 non-coding) that are common in at least 2 of the conditions from all datasets, this validates previous results obtained by us.

Table I
Number of pairs after applying the correlation filter

RNA-seq		**Microarray**		
1 Experiment	*2 Experiments*	*1 Experiment*	*2 Experiments*	*3 experiments*
1562129	1055010	2267	2082	823

Table II
Number of differentially expressed genes

Condition	**Coding**	**ESR**	**non-Coding**
Salt	1130	141	957
DNA Damage	351	43	436
Alpha Factor	367	37	333
Sorbitol	188	25	208
Oxidative Stress	726	92	920
Heat Shock	442	52	363
Stationary Phase	1398	162	948
SC Media	389	40	452
SC Glycerol Media	143	20	264
High Calcium	279	45	390
Low Nitrogen	729	82	665
Low Phosphate	312	35	371
Calcofluor	117	11	213
Hydroxyurea	224	21	299
Grape Juice	1135	126	832
Benomyl	152	23	292
Congo Red	59	12	148

Acknowledgment

This work was funded by National Funds through the FCT - Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within project ADE: PTDC/EIA-EIA/121686/2010 (FCOMP-01-0124-FEDER-020575), project ABLe: PTDC/EEI-SII/2094/2012 (FCOMP-01-0124-FEDER-029010) and by the US 760 Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE BER 761 Office of Science DE-FC02-07ER64494).

References

[1] K. Waern and M. Snyder, "Extensive transcript diversity and novel upstream open reading frame regulation in yeast," *G3: Genes—Genomes—Genetics*, vol. 3, no. 2, pp. 343–352, 2013.

[2] J. A. Lewis and A. P. Gasch, "Natural variation in the yeast glucose-signaling network reveals a new role for the mig3p transcription factor," *G3: Genes—Genomes—Genetics*, vol. 2, no. 12, pp. 1607–1612, 2012.

[3] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[4] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Eds. New York: Springer, 2005, pp. 397–420.