BMC Systems Biology

Oral presentation

Effect of microarray data heterogeneity on regulatory gene module discovery

Alok Mishra* and Duncan Gillies

Address: Department of Computing, Imperial College, London, SW7 2RH, UK

Email: Alok Mishra* - alok.mishra@imperial.ac.uk

* Corresponding author

from BioSysBio 2007: Systems Biology, Bioinformatics and Synthetic Biology
Manchester, UK. 11–13 January 2007

Published: 8 May 2007

BMC Systems Biology 2007, **1**(Suppl 1):S2 doi:10.1186/1752-0509-1-S1-S2

This abstract is available from: http://www.biomedcentral.com/1752-0509/1?issue=S1

Introduction

An integrative genomics approach, in which data from different micro-array experiments are merged together to study regulatory networks [1], has been adopted in several recent research studies. However, we propose that blind use of this approach can be misleading. Our hypothesis is that as micro-array data from different experiments are merged, local patterns of activity, for example the cell cycle, can be masked by more global and dominant patterns such as stress reactions. We have carried out a systematic study in which data with increasing heterogeneity is clustered to determine groups of functionally related genes. These clusters are then tested for similarity to each other.

In order to validate our hypothesis, the primary requirement is to obtain the regulatory modules from various datasets and their mixtures and then measure their similarities to each other. A decreasing trend of similarity as we mix more and more heterogeneous data should confirm our hypothesis. A number of researchers have worked on the problem of finding regulatory networks, some of the most important ones being [2,3] where they have incorporated prior knowledge in the form of known transcription factors or DNA binding data to guide the clustering process. The results in these works have shown that the resulting clusters of regulated gene modules are biologically meaningful. We have used Module Networks algorithm [2] which is a well established approach and has had success in finding biologically relevant modules. For measuring the similarities among sets of regulated gene clusters resulting from this algorithm, we chose to use the *modified Rand Index* [4] which has been shown to be a very stable index of partition similarity.

Materials and methods

In order to validate our hypothesis we chose to work with two very diverse datasets from Stanford Microarray Database (SMD). One of them is when yeast is exposed to stress conditions while other is from cell-cycle related study. Expression of genes when stress conditions are created is much more drastic (both repressed and induced genes) when compared to cell-cycle experiments where optimal conditions are created for growth. We started with analysing data by individual researchers for experiments related to stress [5] in this paper referred as DS-STRESS1 (76 microarrays), [6] called DS-STRESS2 (49 microarrays) and [5] called DS-STRESS3 (41 microarrays). In the next stage we merged all the stress microarrays to create the data set we call DS-STRESS. To compare these clustering against an entirely different category, we took 93 microarray data sets for cell-cycle experiments [7] referred in this article as DS-CCYCLE. A further mixing of both stress and cell-cycle data was named DS-STRESS-CCYCLE. Finally, we extracted all available data (1082 microarrays) for yeast (not only stress/cell-cycle) named DS-ALL and compared the earlier results against it. In order to have statistical significance behind our results we also generated a random microarray dataset for all the genes by generating random numbers from a Gaussian

Table 1: Comparison of individual stress versus progressively mixed datasets

	DS-STRESS	DS-STRESS-CCYCLE	DS-ALL	DS-CCYCLE	DS-RANDOM
DS-STRESS1	0.1616	0.1368	0.1186	0.0354	0.0003
DS-STRESS2	0.0606	0.0555	0.0528	0.0176	0.0001
DS-STRESS3	0.1105	0.1109	0.0989	0.0309	0.0001

Table 2: Comparison of stress and cell-cycle (mixed) versus progressively mixed datasets

	DS-STRESS	DS-CCYCLE	DS-STRESS-CCYCLE	DS-ALL	DS-RANDOM
DS-CCYCLE	0.0418	0.2105	0.0783	0.0638	0.0007
DS-STRESS	0.2933	0.0418	0.2197	0.1784	0.0003

distribution with zero mean and unit standard deviation. This dataset was named DS-RANDOM.

For normalization, we use the assumption that the average log R/G ratio on the array should be zero. Further, we do filtering on the genes selected by choosing genes whose log(base2) of R/G ratio is greater than 2 times for at least one experiment. List of 145 transcription factors (TFs) as prior knowledge were taken from the Yeastract website http://yeastract.com/. We analysed all this data using the software package Genomica which has been provided by the authors of the Module Network.

Results

We compared each of the stress datasets against DS-STRESS-CCYCLE, DS-ALL, DS-CCYCLE which are increasingly distant from the stress datasets as described earlier. As reference, we also compared them against the two extremes of similarity – DS-STRESS which is a mixture of all the stress datasets and DS-RANDOM which is a random dataset. As seen from the results in table 1, different datasets show different similarity even to the DS-STRESS dataset. This suggests that DS-STRESS1 and DS-STRESS3 are more similar to each other than DS-STRESS2, the reason we think is that because they came from experiments related to common research. All the stress datasets' similarity to DS-CCYCLE is very low as we expected because of very different nature of expression in these diverse experiments. As expected, the similarity values for the random data-set are minuscule in all the cases.

The visible trend of similarity values gradually falling as we move from left to right indicates that similar data do keep the similarity among clusters higher while mixing with dissimilar data brings it down. We also did a *combined* data-set level comparison rather than individual data sets as done earlier. In this we compared the cell cycle and stress data-set with each other, DS-STRESS-CCYCLE, DS-ALL and DS-RANDOM. The results in table 2 generalise and substantiate our earlier observations as the same trends are even more robust here.

References

1. Tanay A, Steinfeld I, Kupiec M, Shamir R: **Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium.** *Mol Syst Biol* 2005, **1:**2005.0002-.
2. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nature Genetics* 2003, **34(2):**166-176.
3. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nature Biotechnology* 2003, **21(11):**1337-1342.
4. Hubert L, Arabie P: **Comparing Partitions.** *Journal of Classification* 1985.
5. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11(12):**4241-4257.
6. Saldanha AJ, Brauer MJ, Botstein D: **Nutritional Homeostasis in Batch and Steady-State Culture of Yeast.** *Mol Biol Cell* 2004, **15(9):**4089-4104.
7. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12):**3273-3297.