IDENTIFYING BIOLOGICAL PATHWAYS VIA PHASE DECOMPOSITION AND PROFILE EXTRACTION

Yi Zhang and Zhidong Deng*

Department of Computer Science, Tsinghua University
Beijing, 100084, China
**Email: michael@tsinghua.edu.cn*

Biological processes are always carried out through large numbers of genes (and their products) and these activities are often organized into different cellular pathways: sets of genes that cooperate to finish specific biological functions. Owing to the development of microarray technology and its ability to simultaneously measure the expression of thousands of genes, effective algorithms to reveal biologically significant pathways are possible. However, some open problems such as large amount of noise in microarrays and the fact that most biological processes are overlapping and active only on partial conditions pose great challenges to researchers. In this paper, we proposed a novel approach to identify overlapping pathways via extracting partial expression profiles from coherent cliques of clusters scattered on different conditions. We firstly decompose gene expression data into highly overlapping segments and partition genes into clusters in each segment; then we organize all the resulting clusters as a cluster graph and search coherent cliques of clusters; finally we extract expression profiles from coherent cliques and shape biological pathways as genes consistent with these profiles. We compare our algorithm with several recent models and the experimental results justify the superiorities of our approach: robustly identifying overlapping pathways in arbitrary set of conditions and consequently discovering more biologically significant pathways in terms of enrichment of gene functions.

1. INTRODUCTION

The rapid development of high-throughput techniques such as Oligonucleotide and cDNA microarrays [5] enable measuring the expression of thousands of genes simultaneously. This possibility offers an unprecedented opportunity to characterize the underlying mechanisms of a living cell. Activities of a living cell are so complex that different sets of genes participate in diverse biological processes to perform various cell functions. In this sense, identifying cellular pathways, sets of coherent genes that coordinate in biological processes to achieve specific functions, plays a considerable role in gaining an insight into the cell's activities.

Recently, researchers have made tremendous efforts to identify coherent gene groups [10]. Pioneering work includes agglomerative algorithm for hierarchical clustering [7], K-Means clustering of genes [17] and some graph-theoretical approaches for gene-based clustering [15]. Admittedly, applying traditional clustering algorithms on gene expression data can provide us with new perspectives on cellular processes. However, several problems in this process should be highlighted: (1) biological processes are active only on partial conditions. This fact renders clustering genes on entire conditions ineffective. (2) Extremely high noise exists in microarrays, which calls for robust models for pathway identification. (3) Partitioning genes into mutually exclusive clusters is unreasonable in that biological pathways are always overlapping.

Biclustering algorithms [14] are designed to capture biological processes active on part conditions. Different from traditional clustering methods, these models perform simultaneous clustering on both rows and columns and thus finally discover coherent submatrices where rows refer to genes and columns correspond to relevant conditions. One challenge to biclustering is that all the possible combinations of various genes and conditions are almost infinite.

Furthermore, overlapping property of cellular pathways is also mentioned by recent work. On one hand, some biclustering algorithms discover submatrices one after another and thus naturally yield non-exclusive biclusters. For instance, Cheng *et al* [6] mask the previous biclusters with random numbers and find other ones. Similarily, in [12] each bicluster merely deals with the "residual" expression of previous biclusters. On the other hand, algorithms aim to discover overlapping pathways simultaneously also existed. Battle et al [4] proposed a probabilistic model to discover overlapping biological processes concurrently.

Managing high noise in gene expression data is also indispensable for successfully determining

[*] Corresponding author.

coherent genes. In [2], the author engages robust similarity measure based on the rank of expression in each condition, rather than the accurate expression level, to model the similarity between expression profiles of genes. This sort of measures is robust in the sense that they mainly focus on the rough shape of expression profile and will not be affected by disturbances on accurate expression level. The consensus clustering algorithm in [9] combine different clustering results to form a clustering ensemble. The underlying idea of this ensemble learning approach is that integrating opinions of different “experts” can yield a robust estimation.

In fact, an algorithm address all of these open problems is highly desirable. In this paper, we propose a strategy that satisfies all these demands: robustly discovering overlapping pathways on partial conditions. Other than traditional approaches that directly seek grouping over genes, our algorithm identifies cellular pathways by robustly searching expression profiles over arbitrary set of conditions. The key ideas of our approach are: (1) decompose the entire conditions into highly overlapping segments and clustering genes over each segment; (2) manage all resultant clusters into a cluster graph and discover coherent cliques on cluster graph; (3) extract expression profiles over coherent cliques and shape overlapping pathways according to the these profiles. As a result, this algorithm is capable of robustly recognizing overlapping molecular pathways on partial conditions and thus furnishing biologically significant sets of genes in terms of enrichment of gene functions.

2. METHOD

Our pathway discovery algorithm consists of three steps: (1) decomposing conditions into overlapping segments and performing gene clustering on each segment; (2) construct a cluster graph from the resulting clusters over all segments and discovering coherent cliques on the graph; (3) extracting expression profile from each coherent clique and identifying biological pathway according to each profile. In the rest of this section, we examine these steps in section 2.1~2.3 and analyze the properties of this algorithm in section 2.4.

2.1. Phase Decomposition

In order to capture biological processes in partial conditions, we divide the entire conditions (i.e. columns) into highly overlapping segments. Each segment contains all the rows and a few columns in gene expression matrix. Then we discover co-expressed genes by gene-based clustering on each segment. Finally, large clusters are retained for later processing.

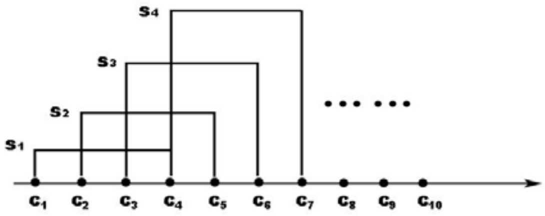

Fig. 1. Phase Decomposition

The first step is to decompose gene expression matrix into many segments. The goal of this decomposition is to ensure that biological process active on any partial conditions can be discovered by combining some segments. Note the term “segment” refers to a submatrix in gene expression matrix where all rows and a subset of columns are included. The decomposition strategy is shown in figure 1. Each segment covers fixed amount of conditions and advances a small step based on the previous segment. For instance, s_1 covers {c_1, c_2, c_3, c_4} and s_2 contains {c_2, c_3, c_4, c_5}. Two parameters are involved: (1) segment length L: the number of conditions covered by a segment. Note that using too large L loses the possibility to discover pathways active on short period; while engaging too small L makes clustering on each segment ineffective: co-expression on very short period always appears by chance. We set $L = 4$ in figure 1 only for illustration; such short segments will not be used in experiments. (2) step length $\triangle L$: in figure 1, we set $\triangle L = 1$, thus any biological process whose life-span is larger than L can be obtained from combining certain segments. For instance, period c_2 ~ c_6 can be captured by integrating two segments s_2 and s_3. One may choose larger step length in order to reduce the total number of segments. But fortunately, combinations of different segments can approximately represent any period larger than L, as long as segments are highly overlapping.

The second procedure is gene-based clustering on each segment so as to obtain the co-expression group. Here we use hierarchical clustering, with average link and Pearson correlation, to group genes on each segment. On each segment, cutting the hierarchical tree at specific threshold 1-c will produce many sets of co-expressed genes. Note c is a key parameter of our algorithm: two gene expression profiles are considered coherent when their Pearson correlation is larger than c, i.e. their distance is smaller than 1-c.

At last, clusters which contain less than 5 genes are discarded in that too small clusters are considered outliers or biologically insignificant groups.

2.2. Coherent Clique on Cluster Graph

After clustering on each segment as discussed in section 2.1, we obtain many co-expression gene clusters. Clusters on the same segment are mutually exclusive while clusters computed from different segments may be highly overlapping, especially when step length is small and thus adjacent segments may present similar structure on gene expression. In this section, we address the problem of how to utilize these clusters to discover biological processes that active on arbitrary period. For this purpose, we propose the concepts of cluster graph and coherent clique; then we focus on how to discover coherent cliques on cluster graph. Note that searching coherent cliques is to find possible biological processes.

Firstly, given two gene clusters *C and C'*, we define the *overlapping degree*, and use this definition to offer a distance measure between clusters. Note that $|C|$ is the amount of genes in cluster C.

- *Overlap(C, C')* = $\dfrac{|C \cap C'|}{|C \cup C'|}$
- *Distance(C, C')* = 1 – *Overlap(C, C')*

Secondly, after defining the distance between clusters, all clusters obtained from the procedures discussed in section 2.1 constitute a large *cluster graph*, which furnishes us with a global view of relationships among genes over different segments. *Cluster graph G(V, E)* is a complete graph where each node $v \in V$ refers to a cluster C and the weight of an edge $e = (v_1, v_2) \in E$ is the distance between two clusters corresponding to v_1 and v_2.

Thirdly, the concept of *β-coherent clique* is proposed as following: a *β-coherent clique Q(V', E')* in a cluster graph *G(V, E)* is a complete subgraph in *G*, satisfying that (1) any edge in *E'* has an weight less than β. (2) *V'* contains at least two nodes. Note that a β-coherent clique is biologically meaningful: (1) any two clusters in a coherent clique has a distance smaller than β, i.e. an overlapping degree larger than 1-β; (2) clusters in a β-coherent clique must come from different segments, since clusters from same segment are mutually exclusive; (3) The fact that several clusters from diverse segments share a large proportion of common co-expressioned genes indicates the existence of a biological process which is active on the period composed of these segments.

Finally, given a cluster graph, we want to discover β-coherent cliques on this graph. An effective algorithm to attain such goal is the hierarchical clustering with complete link [11]. Using this algorithm, we can get a hierarchical tree, and then cut the tree into many β-coherent cliques according to certain choice of β. The definition of complete link ensures that the resulting clusters on cluster graph are β-coherent cliques.

2.3. Profile and Pathway Extraction

In section 2.2, we partition the entire cluster graph into many β-coherent cliques. In this section, we discuss how to robustly extract expression profile of the biological process underlying each coherent clique. We also address how to discover cellular pathways, i.e. set of coordinated genes, from expression profiles.

To begin with, recall that a coherent clique is composed of a set of nodes in cluster graph and each node refers to a cluster from a segment. Since each cluster covers certain conditions: the conditions covered by the segment where this cluster is generated, thus we can define the *active period* of a coherent clique:

- The *active period P(Q)* of a coherent clique Q is all the conditions that covered by at least one cluster in Q.

See figure 1 for an illustration: Supposed that coherent clique Q is composed of three clusters, which are generated on segment s_1, s_2 and s_4, respectively. Then, the active period of Q is $\{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$.

Another important notion is the *core genes*:

- Gene g is the *core gene* of coherent clique Q if and only if g is the member of all clusters in Q.

Furthermore, the expression profile involving the underlying biological process of coherent cluster Q is:

- The expression profile of coherent cluster Q is defined on Q's active period and computed as the mean expression of all the core genes of Q.

Finally, we identify the cellular pathway corresponding to Q based on its expression profile:

- A gene g belongs to the cellular pathway of coherent clique Q if and only if the Pearson correlation between g's expression and Q's expression profile on Q's active period exceeds *c*, the coherence parameter mentioned in section 2.1. Note that *g*'s expression outside *Q*'s active period is not considered.

2.4. Further Analysis

In this section we mainly discuss three properties: (1) ability to discover pathways on partial conditions; (2) identifying overlapping pathways; (3) robustness.

Firstly, if the active period of a biological process *P* can be obtained by combining different segments produced in section 2.1, a corresponding coherent clique should be identified from cluster graph. This is based on two assumptions: (1) segment is defined in suitable granularity and highly overlapping; (2) genes cooperating in *P* should co-express in *P*'s active period.

Secondly, the overlapping property of pathways is ensured: (1) the active periods of diverse coherent cliques Q_1 *and* Q_2 differ, thus a gene *g* can be consistent with both Q_1's expression profile in Q_1's active period and Q_2's expression profile in Q_2's active period. (2) Even on same period, expression level of gene *g* can be consistent with different expression profiles.

Thirdly, the algorithm in this paper is robust, from four perspectives: (1) The definition of coherent clique is robust in that any two clusters in the clique have high overlapping degree and thus it is unlikely that an "outlier" cluster can be accommodated. (2) The computation of active period of each coherent clique is robust. See figure 1 for an illustration: supposed a biological process *P* is active on conditions c_1~c_7, then the corresponding coherent clique *Q* should have an active period as c_1~c_7. Ideally speaking *Q* ought to consist of four clusters which are located in segments s_1, s_2, s_3 and s_4, respectively, because *P* is active on these segments and genes participate in *P* should co-express on these segments. However, owing to high noise in microarrays, some clusters may be missed. As a result, *Q* may consist of only three clusters on s_1, s_2 and s_4, respectively. Even in this case, the active period of *Q* will be judged properly as c_1~c_7, based on s_1, s_2 and s_4. In short, segment overlapping ensures the robustness of active period estimation. (3) The choice of core genes in a coherent clique *Q* is robust in that each core gene must belong to all the clusters in *Q*. Since these clusters are located in different segments and obtained by clustering on each segment independently, it is unlikely that an outlier gene will belong to all these clusters. Admittedly, this selection strategy is so "cautious" that it may miss some core genes. However, only a subset of core genes is still sufficient to extract expression profile of the underlying biological process, because core genes are supposed to co-express well in their biological process. (4) At last, the quality of core gene selection and active period estimation ensures the quality of expression profile extraction and the resultant pathway.

3. EXPERIMENTAL RESULTS

In this section we present empirical results. Compared with several state-of-the-art models, our algorithm is more capable of identifying biologically significant pathways in terms of the enrichment of gene functions.

3.1. Dataset and Preprocessing

Two well-know datasets used in our experiments are yeast cell cycle data [16] and yeast stress data [8]. For preprocessing, we remove genes with more than 5% missing values and estimate missing values by KNNimpute [18]; then genes with small variance are removed. These steps result in 526 genes in cell cycle dataset and 659 genes in stress condition dataset.

3.2. Rival Methods

In this part, we introduce rival algorithms and their parameters: (1) **HClust** [7]: hierarchical clustering with average link and Pearson correlation. Finally 30 clusters are formed on both two datasets. (2) **Plaid** [12] is designed to discover biclusters one by one independently. The default parameters are used. We stop at 100 biclusters on both datasets. (3) **OP** [4] is a probabilistic model to search overlapping pathways simultaneously. The number of pathways is set as 30. (4) **PIPE**(Pathway Identification by Profile Extraction): our algorithm. The coherence threshold *c* mentioned in section 2.1 and 2.3 is 0.7; the parameter β used to define β-coherent clique is 0.7. On cell cycle dataset which contains 76 conditions, we set segment length *L* as 10 and step length $\triangle L = 2$; for yeast stress dataset contains 173 conditions, *L* is engaged as 20 and $\triangle L$ is set as 3.

3.3. Results on Cell Cycle

Running our algorithm on 526 genes in 76 conditions results in 162 coherent cliques, thus finally we obtain 162 cellular pathways. Note that these pathways are generated independently, thus the fact that 162

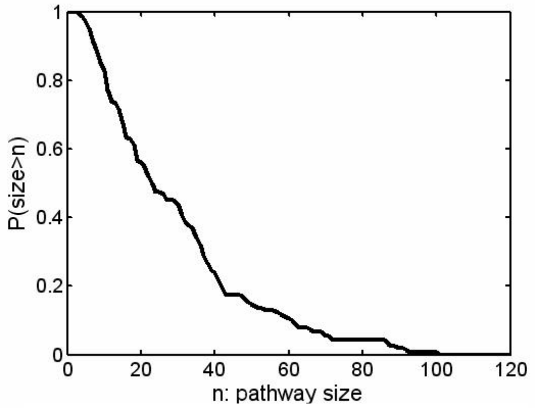

Fig. 2. Distribution of Pathway Size

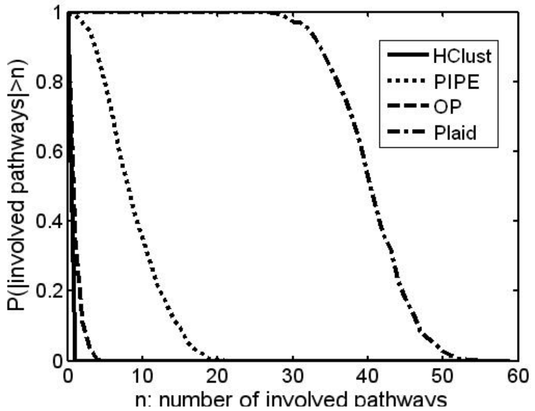

Fig. 3. Distribution of Gene Participation

pathways are produced from 526 genes does not indicate that the average size of pathways is quite small. In reality, the smallest pathway contains 4 genes and the largest one includes 101 genes. The distribution of pathway size is shown in figure 2. The X-axis is pathway size; and the Y-axis is the proportion between pathways larger than specific size and all the pathways. From figure 2 we can observe that more than 80% pathways contain more than 10 genes, while only about 20% pathways contains more than 40 genes. This result shows that the majority of pathways have moderate size.

More interestingly, we measure the overlapping among pathways. HClust will certainly generate 30 mutually exclusive pathways, and running OP model results in 30 slightly overlapping pathways; at last, plaid model find 100 biclusters one by one. Figure 3 shows the distribution of amount of pathways that each gene participates in. The X-axis is the number of pathways a gene joins; the Y-axis is the proportion between the genes which involve more than a specific amount of pathways and all the genes. Four algorithms present quite different properties in figure 3: (1) Since HClust merely produce mutually exclusive pathways, all genes take part in only one pathway. (2) For OP model, single gene takes part in at most 5 pathways, and only 19 out of 526 genes participate in more than three pathways. (3) In Plaid algorithm, pathways are excessively overlapping: almost all genes participate in more than 30 pathways. (4) For our PIPE method, the result in figure 3 shows a natural distribution that only a few genes throw themselves into more than 15 pathways.

According to many researches concentrating on scale-free topology of biological networks [3] and especially of genetic regulatory networks [13]: (1) there should be a few “hub” genes connected with many other genes and thus join a lot of biological processes; (2) most genes in network should not have large degrees and thus they participate in only a few biological processes. As shown in figure 3, only our algorithm generates results consistent with above conclusion. HClust and OP model can not produce “hub” gene, and Plaid produce too many “hubs”.

At last, to justify the biological significance of the pathways generated by these models, we test the enrichment of gene functions in GO categories [1]. For any pathway, the enrichment of a GO category is represented by p-value: the smaller the p-value, the better the enrichment. The p-value is computed based on Genomica [19]. For each GO category, we focus on the p-value of the pathway with best enrichment. For comparison, enrichment of all four algorithms are computed and listed in Table 1. Note that p-value larger than 0.001 is considered as a failure to find enrichment and is labeled as “---” in the Table. Among 117 GO categories listed in Table 1: (1) PIPE won 70 times, while OP, HClust and Plaid models won 25, 21 and 1 times, respectively. (2) PIPE failed 23 times, yet OP, HClust and Plaid failed 41, 58 and 90 times, respectively. Further examining the results listed in Table 1 will naturally yield to the conclusion that PIPE has identified more biologically significant pathways.

3.4. Results on Stress Condition

To further justify the superiority of PIPE, we engage yeast stress condition dataset [8] for another experiment. Running our algorithm on 659 genes and 173 conditions brings about 174 pathways. Pathway size distribution of PIPE is demonstrated in figure 4, where a few pathways contain more than 100 genes and the majority are in

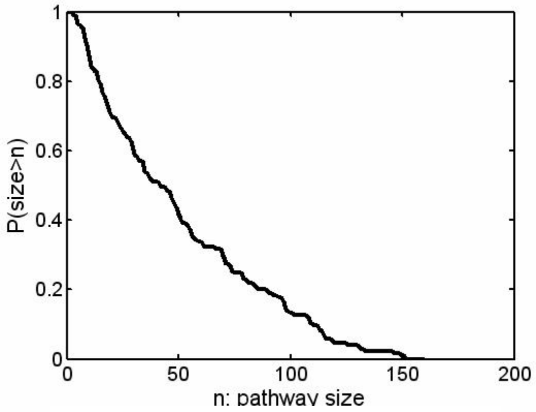

Fig. 4. Distribution of Pathway Size

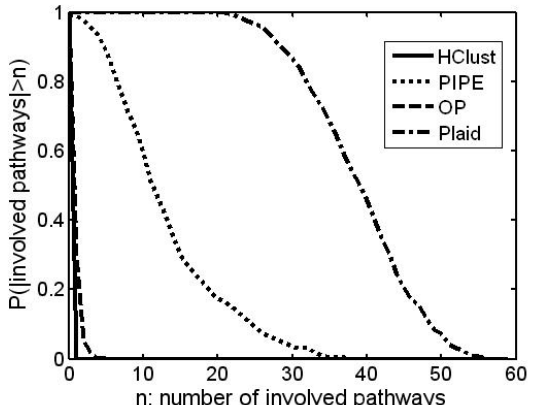

Fig. 5. Distribution of Gene Participation

moderate size. In addition, figure 5 shows similar results as observed in figure 3: HClust and OP can not discover any "hub" gene while Plaid considers most 659 genes as hubs that join a lot of pathways.

We also test the enrichment of GO categories and list the results in Table 2. From table 2 we can summarize that over 128 GO categories listed in Table2: (1) PIPE find best enrichment for 84 categories, while OP, HClust and Plaid find 18, 13 and 19 times, respectively. (2) PIPE failed to find enrichment for 11 GO terms, yet OP, HClust and Plaid failed 61, 81 and 68 times, respectively. In a word, PIPE has its own advantage to discover biologically meaningful pathways.

Another interesting fact is that Plaid algorithm performs much better on stress condition dataset than on cell cycle condition. One explanation for this result is the differences of regulation mechanism between endogenous phase (e.g. cell cycle and sporulation) and exogenous phase (e.g. stress condition, DNA damage and diauxic shift) [13]. In exogenous phase such as stress response, genes are often regulated by more transcriptional factor and participate in more processes than in endogenous phase such as cell cycle. Therefore, Plaid model, which tends to produce excessively overlapping pathways, results in more accurate results.

4. CONCLUSION

In this paper, we presented a new approach to discover cellular pathways. We firstly decompose gene expression matrix into highly overlapping segments and partition genes into clusters on each segment; then we organize all the resulting clusters into a cluster graph and identify coherent cliques; finally we extract expression profiles of coherent cliques and shape biological pathways from these profiles. We compare our algorithm with several recent models and the experimental results justify the superiorities of our approach: robustly identifying overlapping pathways on partial conditions and consequently discovering biologically significant pathways.

Acknowledgments

This work was supported in part by the National Science Foundation of China under Grant 60321002 and the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of MOE (TRAPOYT), China.

References

1 M. Ashburner, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 25: 25-29, 2000.

2 R. Balasubramaniyan, E. Hullermeier, N. Weskamp and J. Kamper. Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7): 1069-1077, 2005.

3 A.L. Barabasi and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5: 101, 2004.

4 A. Battle, E. Segal and D. Koller. Probabilistic discovery of overlapping cellular processes and their regulation. *Journal of Computational Biology*, 12(7): 909-927, 2005.

5 P.O. Brown and D. Bostein. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33-37, 1999.

6 Y. Cheng and G.M. Church. Biclustering of expression data. *ISMB*, 2000.

7 M.B. Eisen, P.T. Spellman, P.O. Brown and D. Bostein. Cluster analysis and display of genome-

wide expression patterns. *PNAS.* 95, 14863-14868, 1998.

8 A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11: 4241-4257, 2000.

9 T. Grotkjar, O. Winther, B. Regenberg, J. Nielsen and L.K. Hansen. Robust multi-scale clustering of large DNA microarray datasets with consensus algorithm. *Bioinformatics*, 22(1): 58-67, 2006.

10 D. Jiang, C. Tang and A. Zhang. Cluster analysis for gene expression data: a survey. *IEEE trans. on Knowledge and Data Engineering,* 16(11): 1379-1386, 2004.

11 B. King. Step-wise clustering procedures. J. Am. Stat. Assoc. 69, pages 86-101, 1967.

12 L. Lazzeroni and A. Owen. Plaid models for gene expression data. Technical report. Stanford Univ., 2000.

13 N.M. Luscombe, M.M. Babu, H. Yu, M. Snyder S.A. Teichmann and M. Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006): 308-312, 2004.

14 S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE trans. on Computational Biology and Bioinformatics,* 1(1): 24-45, 2004.

15 R. Sharan and R. Shamir. CLICK: a clustering algorithm with applications to gene expression analysis. *ISMB*, 2000.

16 P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.Q. Brown, D. Botstein and B. 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.

17 S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church. Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281-285, 1999.

18 O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17: 520-525, 2001.

19 http://genomica.weizmann.ac.il/index.html

Table 1. GO Categories Enrichment based on Cell Cycle Dataset

GO term	**HClust**	**PIPE**	**OP**	**Plaid**
35S Primary Transcript Processing	4.01E-6	**1.05E-7**	3.55E-5	---
amine biosynthesis	---	**1.91E-6**	7.02E-6	---
amine metabolism	---	2.48E-6	**2.45E-6**	---
amine transport	**7.56E-5**	---	---	---
aspartate family amino acid metabolism	---	---	**1.10E-7**	---
ATP dependent DNA helicase activity	---	**2.49E-6**	---	---
ATP dependent helicase activity	---	**2.14E-5**	---	---
beta-glucosidase activity	---	---	**5.77E-6**	---
bud neck	---	**5.29E-6**	---	---
carbohydrate metabolism	---	**1.25E-4**	---	---
carbohydrate transport	**1.70E-7**	3.20E-5	2.57E-6	6.88E-5
carrier activity	---	**2.58E-5**	2.84E-4	---
cation transport	---	2.57E-4	---	**2.60E-5**
cation transporter activity	---	**7.60E-5**	2.16E-4	1.98E-4
cell communication	---	**2.79E-5**	---	---
cell wall	7.94E-6	---	**2.89E-8**	3.06E-5
chromatin assembly or disassembly	4.66E-7	**8.15E-11**	3.49E-7	---
chromatin binding	3.95E-5	**4.33E-7**	3.95E-5	---
conjugation	---	**4.17E-5**	---	---
contractile ring	3.95E-5	**8.45E-7**	8.82E-5	---

cyclin-dependent protein kinase regular activity	---	**9.09E-5**	---	---
cytokinesis	2.12E-4	**4.59E-6**	---	---
cytokinesis, completion of separation	1.30E-5	1.97E-5	**1.11E-8**	---
cytoplasmic vesicle	---	---	**2.82E-5**	---
cytoskeleton organization and biogenesis	---	**1.05E-4**	---	---
cytosolic large ribosomal subnit (sensu Eukaryota)	2.86E-10	**1.18E-10**	6.26E-10	---
cytosolic small ribosomal subnit (sensu Eukaryota)	2.19E-5	**1.39E-5**	3.29E-5	---
development	---	**1.37E-4**	---	---
DNA binding	1.77E-4	**2.45E-8**	2.26E-6	2.87E-6
DNA helicase activity	---	**1.04E-6**	3.74E-5	---
DNA metabolism	**4.36E-9**	3.25E-8	2.82E-7	1.88E-5
DNA packaging	3.74E-5	**2.24E-8**	1.80E-4	---
DNA recombination	7.99E-6	**4.45E-10**	5.84E-5	---
DNA repair	1.62E-6	**8.37E-8**	2.56E-5	1.98E-4
DNA replication	---	**1.18E-4**	---	---
DNA replication initiation	---	**2.14E-5**	5.84E-5	---
DNA replication, synthesis, of RNA primer	---	**2.71E-4**	---	---
DNA strand elongation	**1.21E-6**	4.12E-6	1.20E-5	1.07E-4
DNA unwinding	2.12E-4	**2.49E-6**	2.12E-4	---
DNA-dependent ATPase activity	2.12E-4	**2.49E-6**	2.12E-4	---
DNA-directed DNA polymerase activity	6.00E-5	**9.86E-7**	2.53E-4	---
electron transporter activity	2.26E-4	---	**1.36E-4**	---
endoplasmic reticulum	---	**1.96E-4**	---	---
endosome	---	---	**2.82E-5**	---
energy pathways	**2.84E-6**	1.33E-4	3.05E-5	---
glucosidase activity	---	---	**7.63E-7**	---
glutamate metabolism	**4.13E-6**	9.92E-6	1.20E-4	---
glutamine family amino acid biosynthesis	**4.13E-6**	9.92E-6	1.20E-4	---
glutamine family amino acid metabolism	**3.52E-5**	8.28E-5	---	---
helicase activity	---	**5.51E-6**	---	---
hexose transport	**2.10E-8**	7.02E-6	3.26E-7	---
hydrolase activity	---	---	**1.16E-4**	---
ion transporter activity	---	1.10E-4	**3.16E-5**	---
iron ion transport	---	---	**2.82E-5**	---
iron ion transporter activity	---	---	**7.55E-7**	---
iron-siderochrome transport	---	---	**2.82E-5**	---
kinase inhibitor activity	---	---	**1.86E-4**	---
lagging strand elongation	---	**5.63E-6**	---	---
large ribosomal subunit	1.26E-9	**5.20E-10**	2.75E-9	2.42E-4
leading strand elongation	---	**2.71E-4**	---	---
main pathways of carbohydrate metabolism	**2.84E-6**	1.76E-4	4.63E-5	---
mannose transporter activity	**2.10E-8**	7.02E-6	3.26E-7	---
MCM complex	3.95E-5	**4.33E-7**	3.95E-5	---
metal ion transporter activity	---	---	**2.49E-5**	---
methionine metabolism	8.59E-5	---	**2.22E-8**	---
mismatch repair	8.29E-6	**1.42E-6**	4.71E-5	

mitochondrion	---	**1.31E-4**	---	2.76E-4
mitotic recombination	8.29E-6	**8.34E-9**	4.71E-5	---
nuclear organization and biogenesis	---	**1.48E-7**	---	---
nucleic acid binding	---	**1.57E-7**	4.52E-5	1.93E-6
nucleolus	**8.01E-23**	7.10E-20	2.42E-18	4.90E-9
nucleosome	5.46E-09	**1.35E-13**	3.90E-9	---
nucleotide-excision repair	---	**1.66E-5**	---	---
nucleotidyltransferase activity	---	**4.76E-5**	---	---
organic acid metabolism	---	9.04E-5	**1.80E-5**	---
polyamine transport	---	**9.98E-6**	---	---
pre-replicative complex	---	**1.40E-7**	3.13E-5	---
pre-replicative complex formation and maintenance	---	**1.40E-7**	3.13E-5	---
protein amino acid glycosylation	**1.42E-4**	---	---	---
protein binding	**1.54E-5**	4.69E-5	---	1.63E-4
protein biosynthesis	5.80E-13	**1.13E-13**	2.36E-12	8.46E-6
protein folding	**2.17E-12**	2.16E-10	1.07E-10	---
pyrophosphatase activity	---	**2.08E-4**	---	---
regulation of cell cycle	---	**3.51E-5**	---	---
regulation of cyclin dependent protein kinase activity	---	**1.18E-4**	---	---
regulation of enzyme activity	---	**1.18E-4**	---	---
regulation of metabolism	---	**3.46E-5**	---	---
regulation of transcription	---	**1.76E-5**	---	---
regulation of transcription from Pol 2 promoter	---	**3.06E-5**	---	---
replication fork	---	---	**4.71E-05**	---
response to DNA damage stimulus	**4.43E-9**	7.21E-9	2.41E-7	2.12E-5
response to stimulus	2.46E-5	**8.37E-6**	2.04E-4	---
response to stress	4.67E-5	**1.15E-5**	---	---
ribosomal large subunit biogenesis	9.54E-5	**6.84E-6**	---	---
ribosomal small subunit assemble and maintenance	---	**1.86E-4**	---	---
ribosomal subunit assembly	---	**9.98E-6**	---	---
RNA binding	**2.46E-4**	---	---	---
RNA metabolism	**2.46E-12**	2.61E-11	1.96E-9	1.07E-4
RNA processing	2.10E-12	**1.50E-12**	8.62E-10	1.46E-6
rRNA processing	**8.54E-14**	1.95E-13	3.73E-11	1.20E-5
siderochrome transport	1.65E-6	9.94E-5	**1.88E-8**	---
site of polarized growth	---	**2.36E-4**	---	---
small ribosomal sununit	2.19E-5	**1.39E-5**	3.29E-5	---
structural constituent of ribosome	1.12E-14	**2.15E-15**	4.63E-14	8.24E-7
sulfur amino acid biosynthesis	8.59E-5	---	**2.22E-8**	---
sulfur amino acid metabolism	7.89E-6	---	**1.70E-10**	---
sulfur metabolism	1.42E-4	---	**3.54E-9**	---
telomerase-independent telomere maintenance	6.00E-5	**5.18E-7**	2.53E-4	---
telomere maintenance	6.00E-5	**5.18E-7**	2.53E-4	---
transcription regulator activity	---	**2.78E-4**	---	---
transferase activity and phosphorus-contain group	---	**2.04E-4**	---	---
transcription metal ion transporter activity	---	---	**2.49E-5**	---

translational elongation	1.54E-6	**9.70E-7**	2.32E-6	---
tricarboxylic acid cycle	2.26E-4	---	**1.68E-4**	---
tricarboxylic acid cycle intermediate metabolism	**7.17E-7**	1.23E-4	4.17E-5	---
unfolded protein binding	**1.68E-9**	4.21E-7	4.14E-8	2.94E-5
vesicle-mediated transport	---	**1.11E-4**	---	---

Table 2. GO Categories Enrichment based on Stress Condition Dataset

GO term	**HClust**	**PIPE**	**OP**	**Plaid**
35S Primary Transcript Processing	3.11E-7	**4.27E-10**	6.35E-7	3.46E-9
acid phosphatase activity	---	1.18E-6	**4.21E-7**	---
aerobic respiration	---	**3.74E-6**	---	---
alcohol metabolism	7.70E-5	**1.48E-5**	7.70E-5	8.57E-5
amine biosynthesis	---	**8.56E-10**	2.54E-4	1.27E-9
amine metabolism	4.94E-4	1.73E-12	4.93E-4	**2.41E-15**
amine transport activity	**2.01E-4**	---	3.17E-4	---
aspartate family amino acid biosynthesis	---	**1.31E-5**	---	5.08E-4
aspartate family amino acid metabolism	9.58E-6	**3.02E-10**	9.58E-6	3.44E-5
ATP dependent DNA helicase activity	---	**2.13E-5**	---	1.64E-4
carbohydrate catabolism	---	---	**2.41E-4**	---
carbohydrate kinase activity	---	**3.03E-5**	---	---
carbohydrate metabolism	6.28E-10	5.67E-9	9.50E-7	**2.71E-10**
carbohydrate transport	---	**6.67E-7**	2.10E-4	---
cell communication	---	**2.06E-4**	---	---
cell wall	---	**1.72E-7**	1.77E-6	4.95E-4
cytosolic large ribosomal subnit (sensu Eukaryota)	1.65E-19	**1.50E-32**	1.35E-18	4.18E-20
cytosolic small ribosomal subnit (sensu Eukaryota)	2.02E-11	**2.36E-21**	6.56E-11	4.33E-12
disaccharide metabolism	---	**4.26E-5**	---	---
DNA binding	---	**1.55E-4**	---	---
DNA-directed RNA polymerase 1 complex	---	**2.92E-4**	---	---
endoplasmic reticulum	---	2.86E-5	**2.34E-5**	---
energy pathways	1.87E-12	1.64E-10	**5.30E-14**	9.49E-14
energy reserve metabolism	9.12E-5	**3.49E-6**	3.03E-5	3.55E-4
enzyme regulator activity	---	**1.74E-4**	---	---
galactose metabolism	**2.11E-8**	2.11E-7	**2.11E-8**	---
glucan metabolism	---	**5.67E-5**	6.53E-5	---
glucose metabolism	---	---	---	**8.08E-5**
glucosyltransferase activity	---	**3.42E-4**	---	---
glutamate metabolism	---	**9.34E-5**	---	---
glutamine family amino acid biosynthesis	---	**3.78E-7**	---	2.26E-4
glutamine family amino acid metabolism	---	**5.55E-6**	---	---
glutathione peroxidase activity	---	**1.38E-4**	---	---
glycogen metabolism	---	**5.67E-5**	6.53E-5	---
helicase activity	4.26E-4	**9.89E-7**	---	1.39E-4
heterocycle metabolism	4.72E-4	**1.43E-4**	---	---
hexose metabolism	**1.43E-5**	1.39E-4	**1.43E-5**	---

hexose transport	---	**9.66E-7**	6.53E-5	---
hexose transporter activity	---	**1.46E-4**	---	---
hydrolase activity acting on ester bonds	---	---	**2.73E-4**	---
ion transporter activity	---	**2.37E-4**	---	---
kinase activity	9.12E-5	**9.79E-6**	---	---
large ribosomal subunit	1.65E-19	**1.50E-32**	1.35E-18	4.18E-20
ligase activity, forming carbon-nitrogen bonds	---	---	---	**4.47E-4**
lipid metabolism	---	---	---	**4.19E-4**
main pathways of carbohydrate metabolism	3.59E-4	2.51E-6	**5.29E-8**	7.70E-6
mannose transporter activity	---	**3.64E-4**	---	---
methionine metabolism	1.18E-6	**4.47E-7**	1.18E-6	---
methyltransferase activity	---	**1.24E-5**	---	8.48E-5
mitochondrion	---	4.83E-5	4.40E-4	**2.51E-7**
Noc complex	---	**2.25E-4**	---	---
non-membrane-bound organelle	---	**1.73E-4**	---	---
nucleic acid binding	**9.82E-12**	6.29E-8	6.44E-11	2.17E-11
nucleolus	6.66E-21	**7.55E-26**	6.73E-20	1.51E-21
nucleotide biosynthesis	**1.74E-5**	**1.74E-5**	2.76E-4	---
nucleotide metabolism	**1.14E-4**	**1.14E-4**	---	---
organic acid metabolism	---	1.72E-10	2.73E-4	**1.33E-13**
oxidoreductase activity	4.65E-4	1.17E-5	2.73E-4	**1.85E-6**
oridoreductase activity on CH-OH group of donors	---	**6.02E-5**	---	3.33E-4
pentose metabolism	---	1.93E-4	**6.26E-5**	---
peroxidase activity	---	**1.16E-9**	1.34E-7	7.03E-5
peroxisomal matrix	---	---	3.76E-4	---
phosphoric ester hydrolase activity	---	3.18E-4	**1.16E-4**	---
polysaccharide metabolism	---	**1.28E-4**	2.10E-4	---
processing of 20S pre-rRNA	1.79E-9	7.70E-11	4.69E-9	**4.39E-11**
protein binding	---	1.74E-7	**3.44E-11**	2.52E-4
protein biosynthesis	2.82E-29	**3.41E-60**	1.64E-27	1.82E-27
protein folding	2.36E-7	1.95E-15	6.67E-17	2.61E-8
purine nucleotide metabolism	**1.74E-5**	**1.74E-5**	2.76E-4	---
pyrophosphatase activity	---	3.55E-4	---	**3.51E-5**
regulation of biosynthesis	---	**2.85E-6**	---	---
regulation of catabolism	---	**1.38E-4**	---	---
regulation of cell redox homeostasis	---	**1.34E-7**	---	---
regulation of cellular process	---	**2.61E-7**	---	---
regulation of translation	---	**6.97E-8**	---	---
regulation of translational fidelity	---	**1.33E-6**	---	---
respiratory chain complex 3	---	**4.72E-5**	1.18E-4	---
response to biotic stimulus	---	8.64E-5	---	**5.28E-5**
response to osmotic stress	---	---	**1.74E-5**	---
response to stimulus	---	1.29E-5	4.22E-8	**6.96E-11**
response to stress	---	2.62E-6	8.32E-9	**6.39E-12**
ribonucleoprotein complex	**3.74E-9**	8.38E-8	1.14E-8	2.44E-8
ribosomal large sununit assembly and maintenance	1.34E-6	1.24E-5	2.54E-6	**8.76E-7**

ribosomal large subunit biogenesis	---	**4.22E-6**	---	4.92E-4
ribosome	---	**6.72E-5**	---	---
ribosome biogenesis	4.97E-5	1.04E-4	**3.65E-5**	7.95E-5
RNA binding	**3.75E-14**	5.21E-11	2.09E-13	2.54E-13
RNA helicase activity	4.26E-4	**9.89E-7**	---	1.38E-4
RNA ligase activity	4.97E-5	**8.23E-10**	7.95E-5	2.22E-7
RNA metabolism	**5.78E-30**	2.02E-20	5.09E-27	1.60E-28
RNA methyltransferase activity	---	**1.19E-4**	---	---
RNA modification	**7.88E-9**	3.03E-8	1.92E-8	1.22E-7
RNA polymerase complex	---	**2.92E-4**	---	---
RNA processing	2.59E-22	3.68E-22	2.92E-21	**5.29E-23**
RNA-dependent ATPase activity	---	**2.13E-5**	---	1.64E-4
rRNA binding	---	**2.89E-5**	---	1.64E-4
rRNA modification	---	**3.41E-6**	---	---
rRNA processing	2.59E-22	3.68E-22	2.92E-21	**5.29E-23**
S-adenosylmethionine-dependent methyltransferease	---	---	---	**4.46E-4**
signal transduction	---	**9.34E-5**	---	---
small nuclear ribonucleoprotein complex	---	**1.73E-4**	---	---
small nucleolar ribonucleoprotein complex	1.50E-7	1.08E-7	3.17E-7	**1.16E-9**
small ribosomal sununit	2.01E-11	**2.32E-21**	6.56E-11	4.33E-12
snoRNA binding	3.11E-7	2.10E-8	6.53E-7	**7.56E-11**
SRP-dependent protein-membrane target	---	---	**5.48E-5**	---
structural constituent of ribosome	2.29E-29	**2.55E-61**	7.00E-28	7.76E-31
succinate dehydrogenase (ubiquinone) activity	---	**7.03E-6**	---	---
sulfur amino acid metabolism	**3.48E-6**	**3.48E-6**	**3.48E-6**	1.45E-4
sulfur metabolism	**7.67E-6**	1.45E-5	**7.67E-6**	---
thioredoxin peroxidase activity	---	**6.72E-6**	---	---
transcription from Pol 1 promoter	---	**2.92E-4**	---	---
transcription regulator activity	---	**1.55E-4**	---	---
transcription, DNA-dependent	---	**2.92E-4**	---	---
transferase activity, transferase acyl groups	---	**3.23E-5**	---	---
transferase activity and phosphorus-contain group	---	**3.60E-4**	---	---
transcription factor activity, nucleic acid binding	---	**5.01E-6**	---	---
transcription initiation factor activity	---	**1.44E-4**	---	---
transcription regulator activity	---	**5.01E-6**	---	---
translational elongation	---	**8.46E-7**	---	---
trehalose metabolism	---	**4.26E-5**	---	---
tricarboxylic acid cycle	---	**1.26E-8**	2.77E-5	---
tricarboxylic acid cycle intermediate metabolism	---	---	**7.61E-5**	3.55E-4
tRNA aminoacylation	2.09E-4	**2.81E-8**	3.12E-4	2.06E-6
tRNA metabolism	2.76E-6	**3.80E-8**	5.08E-6	1.75E-5
tRNA methyltransferase activity	---	**6.85E-5**	---	---
tRNA modification	2.76E-6	**3.80E-8**	5.08E-6	1.79E-5
UDP-glycosyltransferase activity	---	**3.42E-4**	---	---
unfolded protein binding	3.66E-4	1.87E-7	**1.05E-11**	1.14E-4