On the Society of Genome: Social Affiliation Network Analysis of Microarray Data

Jung Hun Ohn[1], Jihoon Kim[1], and Ju Han Kim[1,2,⋆]

[1] Seoul National University Biomedical Informatics (SNUBI)
[2] Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea
jhohn2@snu.ac.kr, hoonie.kim@gmail.com, juhan@snu.ac.kr

Abstract. To investigate the structure of the genomic interaction network built from yeast gene-expression compendium dataset of hundreds of systematic perturbations, social affiliation network analysis methodologies were applied through quantifying various density, closeness and centrality measures and exploring core-periphery structures. Genes affected by a larger number of perturbations were found to be involved in responses to various environmental challenges. Deletion of essential genes was suggested to cause larger number of genes to be significantly up or down regulated. We explored the network structure made up of several sub-networks using core-periphery models to find ancient pathways. Glycolysis and TCA cycle have relatively core positions in the energy-related processes of yeast.

1 Introduction

The current way of describing cellular processes are based on mechanical concepts and each cellular process is regarded as a conveyer belt on which many workers, i.e. proteins, work to give products for the survival of a large factory or a cell. Biology books are full of many such schematic figures, which is, of course, useful for illustrating life phenomena. However, this may mislead. Each gene product or protein has no concept of such processes as DNA replication, apoptosis or signal transduction. They are just interacting with each other without the intention of replicating DNA or transducing signals. These purposeless interactions form the basis of life and may in fact be a better description of life. Complex information exchanges between cellular components keep life go on.

How can we describe this aspect of life? Let us pick the wisdom of social analogy. We endow each gene with its functions from the point of cellular processes like DNA replication and cell cycle control, just as we have our own social roles defined with respect to the social groups like families and jobs. We are in contact with people who share with us the same group memberships, which is the basis of our personal contact and information exchange. One interacts with its group members directly or indirectly and the members are quite important in understanding him: we can know a man by the company he keeps!

⋆ Corresponding author.

L. Wang et al. (Eds.): FSKD 2006, LNAI 4223, pp. 1062–1071, 2006.

Describing the properties of individuals through its social relationship with others has been the subject of study for social network analysts. [1] They try to find social 'stars' in different aspects and to describe the network structure through various centrality measures and navigate its unique structures by graph theoretic approaches. In its graph representation, each node represents an individual and each edge social interaction between two individuals. The presence or absence of interaction between N individuals can be expressed as an N-by-N binary matrix, i.e. 1 for the presence and 0 for the absence of interaction. This matrix is called one-mode matrix. On the other hand, two-mode network represents the affiliation of a set of actors with a set of social occasions. Many social network relations consist of the linkages among actors through their joint participation in social activities or membership in collectivities (i.e. events). Such networks of actors tied to each other through their participation in events and events linked through multiple memberships of actors, are referred to as affiliation networks. [1][2]

Affiliation network is represented as a matrix with binary relationship between actors and events. If an actor is affiliated with an event, the binary relation is given by 1 and otherwise 0 (see methods). Figure 1 shows an example of such affiliation matrices with 18 actors and 12 events.

		1E	2E	3E	4E	5E	6E	7E	8E	9E	10E	11E	12E
1	A1	1	1	1	0	1	1	0	1	1	0	1	0
2	A2	0	0	1	0	1	0	0	1	0	0	0	1
3	A3	0	1	1	1	1	1	0	1	0	1	0	0	1
4	A4	0	0	0	0	0	1	0	1	1	1	0	0	0
5	A5	0	0	0	0	1	1	1	0	0	0	0	0	1
6	A6	0	1	0	0	1	1	1	0	1	0	0	0	0
7	A7	0	0	0	0	1	1	1	0	0	0	0	0	1
8	A8	0	0	1	1	1	1	1	0	1	0	1	0	0
9	A9	0	0	1	0	0	1	1	0	0	0	0	0	0
10	A10	0	0	1	0	0	1	1	0	0	1	1	1	1
11	A11	0	0	1	1	1	0	0	0	0	0	0	0	0
12	A12	0	0	1	0	0	0	1	0	0	0	0	0	0
13	A13	0	1	0	0	0	0	0	0	1	1	1	1	1
14	A14	1	0	1	0	0	0	0	0	0	0	1	1	1
15	A15	0	0	1	1	0	1	1	1	0	0	0	0	0
16	A16	0	1	1	0	1	1	1	0	0	0	1	0
17	A17	0	0	0	1	1	0	0	0	1	0	0	1
18	A18	1	0	1	1	0	0	0	0	0	0	0	0

Fig. 1. An example of affiliation matrix with 18 actors and 12 events

In the present study, we binarized a yeast microarray dataset to build an affiliation matrix and tried to describe and analyze social behavior of yeast genes. With the classical notation of social network analysis, a gene corresponds to an actor and a group of genes to an event.

Rosetta yeast compendium dataset [3] is hitherto the most systematic approach to profile yeast genes. Gene expression levels were measured in 300 different conditions to investigate the impact of uncharacterized perturbations on the cell like deletion mutations and drug treatments. In the original article authors newly annotated eight deleted genes by hierarchical clustering analysis and confirmed it experimentally.

Cohen *et al.* referred to the 'molecular phenotype' of a gene as the constellation of changes in gene expression profile after deletion of the gene.[4] The molecular phenotype is a group of genes that are significantly up or down regulated by a gene deletion or chemical treatment. Rosetta compendium dataset has gene

expression profiles in 300 different gene deletion mutations and drug treatments. A drug treatment works like a gene deletion as it usually blocks the action of several gene products it binds to. Each gene deletion or chemical treatment condition assigns more than 6,000 genes into two groups, molecular phenotype and non-molecular phenotype. This is why the Rosetta compendium dataset of yeast genes is well suited for our purpose. Genes or actors that are differentially transcribed following deletion of one common gene or common chemical treatment belongs to one group and the group is analogous to the events in social network analysis terminology. Genes affected by common perturbations can be assumed to communicate with each other directly or indirectly. This assumption well justifies our approach to analyze social behavior of genes from the perspective of social affiliation network.

This structural uniqueness of the Rosetta dataset led Rung *et al.* to construct, what they called, disruption networks and they analyzed yeast genome graph theoretically and showed that disruption network is scale-free.[5] The social network analysis framework gives additional insights into gene-to-gene communications.

2 Data and Methods

2.1 Data Preprocessing and Determination of Molecular Phenotypes

Rosetta Compendium dataset was downloaded from ExpressDB. [6] It is a compendium of expression profiles corresponding to 300 diverse mutations and chemical treatments (276 deletion mutants, 11 tetracycline regulatable essential genes, 13 chemical treatments) in S. cerevisiae. Excluding genes that have more than 20 missing values left 6,152 genes for analysis. A data matrix containing log expression ratio in each condition was used for analysis. The matrix was normalized with respect to conditions such that mean and standard deviation of each column log ratio value was set to 0 and 1, respectively.

Generally whether a gene is differentially expressed in a condition is determined in a biological sense by its fold ratio. Statistical significance has also been used as a means of selecting differentially expressed gene in a large dataset. [7] We pooled the log ratio values to get a cutoff for binarization process. We obtained 5% quantile (Q0.05) and 95% quantile (Q0.95) (i.e. -1.24 and 1.33, respectively) for the above normalized log ratio values and used them for the cutoff value determining significant log ratio.

2.2 Binarization

Let Eij be the normalized (with respect to condition) log expression ratio of gene i in gene disruption or chemical treatment condition j above. New data matrix A with Aij as its element is given by: A = <Aij>,

$$A_{ij} = \begin{cases} 1 \text{ (if } Q0.05 < Eij < Q0.95) \\ 0 \text{ (if } Q0.05 > Eij \text{ or } Q0.95 < Eij) \end{cases}$$

Gene or actor i is affiliated with the molecular phenotype of gene mutation or drug treatment condition j if A$ij = 1$ and is not affiliated if A$ij = 0$. A is the **affiliation matrix** shown in figure 1.

2.3 Analysis of Affiliation Network

Bipartite matrix. Affiliation matrix A is tranformed into a bipartite *square* matrix B given by, (given N actors, M events and O representing zero matrix)

$$\mathbf{B} = \begin{pmatrix} O(NXN) & A(NXM) \\ A(MXN) & O(MXM) \end{pmatrix}$$

Bipartite graph. A graph is bipartite if the vertices are partitioned in two mutually exclusive sets such that there are no ties within either set and every edge in the graph is an unordered pair of nodes in which one node is in one vertex set and the other in the other vertex set. Bipartie graph is very useful in representing two-mode network.

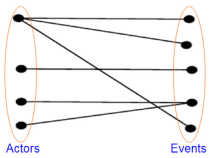

Fig. 2. Bipartite graph representation of an affiliation matrix. Left vertices are actors and right ones events.

Geodesic distance. A shortest path between two nodes is referred to as a geodesic. A geodesic distance matrix G =<Gij> represents geodesic distances between all pairs of nodes in the bipartite graph.

Rates of participation. Rate of participation of actor i is given by

$$\sum_j A_{ij} \quad .$$

which implies how many events an actor participates in. The more sociable an actor is, the more events will he or she participate in.

Size of events. Size of event j is given by

$$\sum_i A_{ij} \quad .$$

which implies how many actors participate in the event j.

Node centrality measures and group centralization measures. For detailed description of the concept of centrality, refer to [1][2][8] and [9]. The origin of this idea in social network analysis can be found in the concept of the 'star'- the person who is the most 'popular' in his or her group or who stands at the center of attention. Group centralization index measures the extent to which the graph is a *star graph - there is one central node with the remaining nodes considerably less central.* Centrality measures were calculated using the UCINET 6.0 software.[10]

Node Degree Centrality. This is the simplest definition of node centrality. The central node must be the one who have the most ties to other nodes in the network. In the two-mode data, actor degree centrality is the number of events an actor attended and event degree centrality is the number of actors participating in the event. Degree centrality of an actor i is given by

$$\sum_j B_{ij} \quad .$$

Node Closeness Centrality. This measures how close a node is to all the other nodes. In two-mode network represented by a bipartite graph, all paths consist of an alternating series of nodes and edges of the form u-v-u'-v' and so on where u and u' are from one vertex set and v and v' from the other. The closeness centrality of a node was defined by Freeman and is inversely proportional to the total geodesic distance from the node to all other nodes in the network.[11] Closeness centrality of an actor i is given by

$$\Big[\sum_j G_{ij}\Big]^{-1} \quad .$$

Node Betweenness Centrality. This measures the probability that a communication or simply a path from node j to node k takes a particular route through a node i. All lines are assumed to have equal weights. Let g_{jk} be the number of geodesics linking the two nodes j and k. Let g_{jk}(i) be the number of geodesics linking the two nodes that contain node i. In two-mode network, betweenness centrality is a function of paths from actors to actors, events to events, actors to events and vice versa. Betweenness centrality of an actor i is given by

$$\sum_{j<k} g_{jk}(i)/g_{jk} \quad .$$

Group Centralization Measures (Degree, Closeness or Betweenness). Group centralization measure is a group level measure of centrality. Let C(i) be a node centrality index (degree, closeness or betweenness) and C(i)* be the largest value of the indices across all nodes. The general form of group centralization index is given by:

$$C = \frac{\sum_i [C(i)^* - C(i)]}{max \sum_i [C(i)^* - C(i)]} \quad .$$

2.4 Core/Periphery Structures

A common notion in social network analysis is the concept of a core/periphery structure and a dense, cohesive core and a sparse, unconnected periphery are sought. Borgatti *et al.* formalized the notion of core/periphery structure and suggested both discrete and continuous models in detecting core/periphery structure in network data and the computer package UCINET 6 incorporates the model.[12] We adopted the continuous model, which assumes the network has one core and assigns each node a measure of 'coreness'. In UCINET 6, the value of coreness of node i, ci, is obtained so as to maximize the matrix correlation between the data matrix (in affiliation network, the bipartite matrix) and the pattern matrix, P, the element of which is $p_{ij} = c_i * c_j$.

3 Results

3.1 Whole Genome View

Rate of participation. Rate of participation of an actor counts the number of events an actor participates in. Actors that participate in a large number of events are regarded as sociable actors. Genes that are differentially expressed in more than 150 out of 300 perturbing conditions are as follows.

YBR072W, YBR145W, YBR296C, YCL018W, YER069W, YFL014W, YFR030W, YFR053C, YGL255W, YHR018C, YHR137W, YHR215W, YIR034C, YJL088W, YJL153C, YJR025C, YML123C, YMR062C, YMR094W, YMR095C, YMR096W, YMR105C, YNL036W, YNL160W, YOL058W, YPL019C, YPR160W, YPR167C, YJR109C, YKL001C, YKL096W, YLR303W

These genes are "social stars" in yeast genome in that they are parts of a large number of molecular phenotypes and in biological sense, are very sensitive to external perturbations. The MIPS functional classifications[13] of these 'star' genes are 1) Stress response, 2) Amino acid biosynthesis, 3) C-compound and carbohydrate biosynthesis, 4) Small molecule transport, 5) Osmoregulation. The functions are important for the survival of the yeast against various environmental challenges. It is natural to suppose that genes that are involved in the processes related to *interaction with cellular environments* will be frequently up or down regulated by external perturbations.

Size of events. Size of an event implies how many actors participate in the event. Examples of gene deletions or drug treatments with large sizes are:

yor078w, erp4, ymr141c, kar2, yef3, cdc42, rpl12a, cla4-haploid, ymr014w, arg5,6, gyp1, dfr1, rps24a, hes1-haploid, idi1, ymr030w, kre1, bub3, yhr011w, ste20, erg11, 2-deoxy-D-glucose, TUNICAMYCIN, she4, yor006c, pac2, mak10, cue1, cat8, hat2, sir1, ymr285c, ade16, phd1-haploid, bub1-haploid, erg4-haploid, yer041w, prb1, aqy2, yml003w, rml2, hir2, msu1, yml011c, top1-haploid, pma1, rnr1-haploid, yor072w, yel033w, sap30

The functional categories are 1) ribosome biogenesis, 2) lipid, fatty-acid and isoprenoid biogenesis, 3) transport, 4) transcriptional control, 5) cell cycle 6) DNA synthesis and replication, 7) budding and pheromone response. The specific kinds of genes giving rise to a large size of perturbation are somewhat different from those found by Featherstone *et al.* and Rung *et al.* because of different normalization process. But the above functional categories lead to the similar conclusion; genes whose deletion strongly 'wiggles' the whole cellular transcriptional system are 'essential' cellular processes that are always switched on irrespective of environmental stimuli.[5][14] The perturbation may be the direct result of the deletion itself or the indirect one of the triggered mechanisms in compensation for the gene disruption to keep one yeast from being lethal. [14][15]

3.2 Analysis of Genes Participating in Energy Related Processes

We explore the structure of a specific network made up of several sub-networks. The MIPS database provides a catalogue of functional categories which groups together genes with similar functions and we explored the network of genes known to participate in 'Energy' related processes.[13] The energy related gene network is composed of 10 subgroups of genes assigned to the following functional categories. A total of 208 genes were included. The number in the parenthesis is the number of genes participating in the process. These genes have no missing values in Rosetta compendium dataset and errors from missing data were excluded.

1. Oxidation of fatty acid (6)
2. Fermentation (28)
3. Glycolysis and gluconeogenesis (28)
4. Glyoxylate cycle (5)
5. Pentose-phosphate pathway (9)
6. Metabolism of energy reserves (glycogen, trehalose) (33)
7. Respiration (70)
8. TCA cycle (20)
9. Other energy generation activities (13)
10. Electron transport (2)

Core/Periphery structure of Energy related genes. In the Energy related affiliation network, the core/periphery structure is investigated. (See methods for details) Figure 3 shows the distribution of coreness scores of genes in each functional categories (Genes with higher coreness scores form the core). Genes participating in fatty acid oxidation and energy transport are mostly placed in the periphery, whereas, glucose metabolism related process (categories 3 and 6) contain core genes in energy process and ATP generating processes (categories 2 and 8) occupy intermediate position. ATP consuming process (Respiration) related genes have relatively peripheral placement. Ancient pathways like Glycolysis and TCA cycle have relatively core positions in the network.[16]

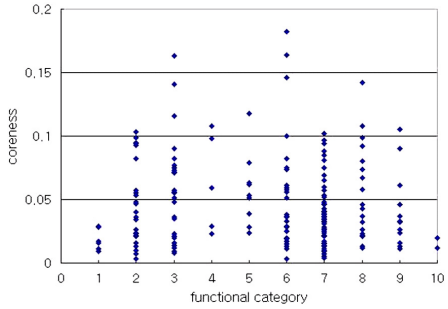

Fig. 3. Core/Periphery structure of energy-related pathways in yeast

Graph centralization index. A graph with higher centralization index is more like a 'star' graph. Table 1 shows the glyoxylate cycle gene group has the highest degree and closeness centralization indices and forms the most 'star' like graph. In contrast, respiration process has the smallest degree, closeness and betweenness centralization indices and has the least star-like structure.

Table 1. Graph centralization index

Process	Degree	Closeness	Betweenness
fatty acid oxidation	26.22	26.99	**66.13**
fermentation	28.32	18.99	16.29
glycolysis	45.81	24.84	24.74
glyoxylate cycle	**52.39**	33.23	49.35
pentose phosphate	45.56	30.84	46.38
energy reserves	49.94	31.23	24.50
respiration	22.78	21.16	8.57
TCA cycle	43.79	26.99	33.68
All actors	26.49	21.59	4.11

Fatty acid oxidation has relatively small degree and closeness centralization indices but it has unusually high betweenness centralization index. YLR284C (ECI1) has the largest betweenness centrality score of all the actors which means other actors depend on this gene to communicate with each other and this gene product might have some control over the interactions.

3.3 TCA Cycle

Now let us focus on one of the sub-networks of energy related processes, or TCA cycle. Borgatti *et al.* pointed out geodesic distance matrix (see methods) as an input for multidimensional scaling gives good visualization results and makes it easy to draw rough conclusions at a glance.[9] Figure 4. shows multidimesional scaling representation of TCA cycle related genes and conditions after geodesic distance matrix is formed from the affiliation matrix. Genes are coded with its enzyme names and 52 conditions (labeled with numbers) were those that contain more than 5 participating genes out of 20 TCA cycle related genes.

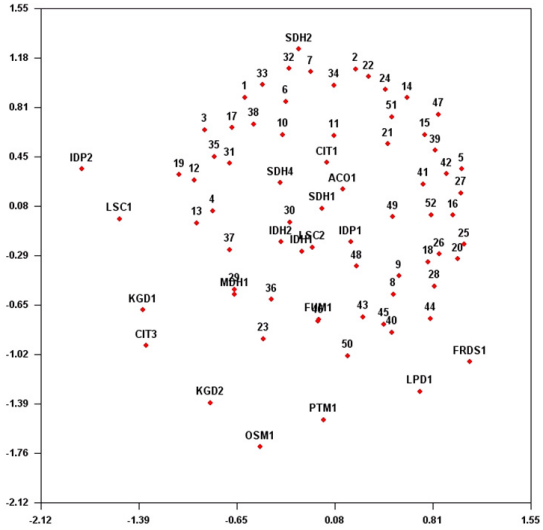

Fig. 4. Multidimesional-scaling representation of TCA cycle-related genes and conditions

We can find a core/periphery structure especially among genes or actors. The core group contains succinate dehydrogenase complex (SDH1, SDH2 and SDH4) and isocitrate dehydrogense complex (IDH1 and IDH2), fumarase, aconitase and citrate synthase. The core group genes are well known TCA cycle related genes and the peripheral genes have hitherto unspecified role in TCA cycle.[17] This might mean the core genes are more exclusively dedicated to a specific process than the peripheral genes.

4 Discussion

The function of a protein is a contextual attribute of strict and quantifiable patterns of interactions between the myriad of cellular constituents.[18] Large-scale gene expression profile was investigated in the context of the social network analysis where genes are regarded as actors, conditions as events, and the network topology as variety of centrality and relatedness indices.

The analysis demonstrated some important features such as core-peripheral players and significant intermediary actors that may be critical for the control of the system and useful for the development of valuable therapeutic substances.

Acknowledgement

This study was supported by a grant from Korea Health 21 R and D Project, Ministry of Health and Welfare, Republic of Korea (A060711).

References

1. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
2. Scott J: Social Network Analysis. SAGE publications (2000)
3. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H.: Functional discovery via a compendium of expression profiles. Cell. 102(2000) 109–126
4. Cohen, B.A., Pilpel, Y., Mitra, R.D., Church, G.M.: Discrimination between Paralogs using Microarray Analysis: Application to the Yap1p and Yap2p Transcriptional Networks. Mol Biol Cell. 13(5) (2002) 1608–1614
5. Rung, J., Schlitt, T., Brazma, A., Freivalds, K., Vilo, J.: Building and analysing genome-wide gene disruption networks. Bioinformatics. suppl. 2 (2002) 202–210
6. Asch, J., Rindone, W., Church, G.M.: Systematic management and analysis of yeast gene expression data. Genome Research 10 (2000) 431–445
7. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N.: Revealing modular organization in the yeast transcriptional network. Nature Genetics. 31 (2002) 370–377
8. Faust, K.: Centrality in affiliation networks. Social Networks. 19 (1997) 157–191
9. Borgatti, S.P., Everett, M.G.: Network analysis of 2-mode data. Social Networks. 19 (1997) 243–269
10. Borgatti, S.P., Everett, M.G., Freeman, L.C.: Ucinet for Windows: Software for Social Network Analysis. Harvard Analytic Technologies USA (2002)
11. Freeman, L.C.: Centrality in social networks. Social Networks. 1 (1979) 215–239
12. Borgatti, S.P., Everett, M.G.: Models of Core/Periphery Structures. Social Networks. 21 (1999) 375–395
13. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., Weil, B.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30 (2002) 31–34
14. Featherstone, D.E., Broadie, K.: Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. BioEssays. 24 (2002) 267–274
15. Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature. 411 (2001) 41–42
16. Wagner, A., Fell, D.A.: The small world inside large metabolic networks. Proc. R. Soc. Lond. 268 (2001) 1803–1810
17. Przybyla-Zawislak B, Gadde DM, Ducharme K, McCammon MT: Genetic and biochemical interactions involving tricarboxylic acid cycle (TCA) function using a collection of mutants defective in all TCA cycle genes. Genetics 1999, 152(1): 153–166
18. Barabasi, A.L., Oltvai, Z.N.: Network Biology: Understanding the cell's functional organization. Nature genetics. 5 (2004) 101–113