이학석사학위논문

Exploration of Affiliation Networks for Biologically Relevant Structures

소속 연결망 탐색을 통한 생물학적으로 의미 있는 구조의 발견

2005년 2월

서울대학교 대학원

협동과정 생물정보학 전공

온정헌

Exploration of Affiliation Networks for Biologically Relevant Structures

by

Jung Hun Ohn

A thesis submitted in fulfillment of
the requirement for the degree of
Master of Science
in Bioinformatics
Seoul National University, Seoul, Korea

December, 2004

Doctoral committee:

Professor ____________________ Chairman

Professor ____________________ Vice chairman

Professor ____________________

Exploration of Affiliation Networks for Biologically Relevant Structures

지도교수 김 각 균

이 논문을 이학석사 학위논문으로 제출함
2004 년 10 월

서울대학교 대학원
협동과정 생물정보학 전공

온 정 헌

온 정 헌 의 이학석사 학위논문을 인준함
2004년 12월

위 원 장	김 주 한	(인)
부위원장	김 각 균	(인)
위 원	조 광 현	(인)

학위논문 원문제공 서비스에 대한 동의서

논문제목 : 소속 연결망 탐색을 통한 생물학적으로 의미 있는 구조의 발견
학과(부) : 협동과정 생물정보학
학　번 : 2002-20630
연 락 처 :
제 출 일 : 2004 년 12 월 일
저 작 자 : 온 정 헌 (인)

ABSTRACT

In social network analysis terminology, affiliation networks are networks with two distinct groups of nodes and can be found in many biological networks. We explored three different kinds of affiliation networks and extracted biologically relevant structures.

One affiliation network is from a systematic yeast gene perturbation microarray experiment and we applied social network analysis methodologies, quantifying various density, coreness and centrality measures. Genes participating in larger number of processes were found to have functions important for the survival of the yeast against 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 cycles have relatively core positions in the energy-related processes of yeast.

Another affiliation network is formed from a systematic protein complex profiling experiment and their combinatorial property was investigated. Cell cycle, signaling and cell structure associated protein complexes share little proteins within each functional group while maintaining diverse component overlaps with complexes of other functional groups. On the other hand, RNA metabolism, protein synthesis and transcription associated complexes utilize many proteins in common both within each functional group and beyond each functional group. And the former functional groups are placed in the periphery while the latter functional groups form a core in a higher order organization map of the yeast proteome based on the combinatorial nature of protein complexes.

The other affiliation network is from chemotherapeutic susceptibility profile data of cancer cell lines to different anti-cancer drugs. In the multidimensional scaling visualization of cell lines and drugs using a geodesic distance as a distance measure, drugs with the same known mechanism of action are clustered together.

Key words: social affiliation network analysis, centrality, core-periphery model, protein complex, geodesic distance
Student Number: 2002-20630

CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 products or proteins have 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 and indirectly and the members are quite important in understanding him: we can know a man by the company he keeps!

Describing the properties of individuals through its social relationship with others has been the subject of study for social network analysts (Wasserman and Faust, 1994). 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

(Wasserman and Faust, 1994; Scott, 2000).

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.

In the present study, we explored three different kinds of affiliation networks to discover biologically relevant structures.

One affiliation network is derived from binarization of Rosetta yeast compendium dataset. Rosetta yeast compendium dataset (Hughes *et al.*, 2000) is hitherto the most systematic approach to profile transcriptional behaviour of 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. Drug treatment works like gene deletion as it usually blocks cellular processes just as a gene deletion blocks cellular processes it is involved in.

Cohen *et al.* introduces the concept of 'molecular phenotype' of a gene as the constellation of changes in gene profile after deletion of the gene (Cohen *et al.*, 2002). In Rosetta compendium dataset, each perturbation assigns more than 6,000 genes into two groups, molecular phenotype of the disrupted gene or non-molecular phenotype. This is why the Rosetta compendium dataset of yeast genes is well suited for derivation of affiliation network. Genes or *actors* will participate in one or more of the molecular phenotypes, or *events*! Genes belonging to one molecular phenotype are assumed to communicate with each other directly or indirectly because they are transcriptionally related.

This structural uniqueness of the Rosetta dataset led Rung *et al.* to construct, what they called, disruption networks and they analysed yeast genome graph theoretically and showed that disruption network is scale-free (Rung *et al.*, 2002). Our social affiliation network analysis well includes the results and gives additional insights into gene-to-gene communications.

Another affiliation network is formed from the protein complex profiling data by Gavin *et al.* (Gavin *et al.*, 2002). In contrast to the first one, this network reflects direct or indirect physical interactions among yeast genes. Systematic analyses of protein complexes have been tried, the most extensive of which is the profiling by Gavin AC *et al.* and Ho *et al* (Gavin *et al.*, 2002; Ho *et al.*, 2002). Gavin *et al.* used tandem-affinity purification (TAP) and mass spectrometry on a large-scale to identify and characterize protein complexes in yeast. Here, protein complexes correspond to *the events* of social affiliation network in which genes or actors physically interact. Most cellular processes are carried out by protein complexes and its identification and characterization gives insights into how the proteins are organized into functional units (Dezso *et al.*, 2003). Until recently, protein complexes like spliceosome, cyclosome, proteasome and nuclear pore complexes are among the well-known (Rout *et al.*, 2000; Zachariae *et al.*, 1996; Neubauer *et al.*, 1997; Verma *et al.*, 2000).

Meanwhile, a particular complex is not necessarily composed of invariable protein members nor is any constituting molecule involved uniquely in that specific complex. Gavin *et al.* illustrated this aspect by linking complexes that share components and derived a higher order network of multi-protein complexes. This combinatorial aspect of utilizing molecular components can be seen in the design of a protein molecule itself composed of conserved domains or motifs that also comprise other kinds of protein molecules giving rise to different functionality. In a likely manner, different molecular machines often use the same protein to exert different functions (Gavin *et al.*, 2003).

We investigated this combinatorial nature at the level of protein complexes by quantifying the degree to which a given complex is made up of exclusively participating protein subunits both all through the functional groups and within each functional group and present a higher order organization map of nine functional groups based on protein component sharing.

Lastly, another advantage of affiliation network analysis is the two modes can be concurrently considered to give insights into their relationship. There have been attempts to classify or cluster tissues or genes based on transcription profiles from microarray data. But they were within the group of tissues or genes themselves and the trials to pool the two groups together and then cluster or classify them have been relatively rare because of the difficulty in measurement of distances between the two heterogeneous groups (Butte *et al.*, 2000). Using geodesic distance as distance measure, we made an attempt to investigate the association among entities of the two distinct groups and draw biologically and clinically relevant structures using the chemotherapeutic susceptibility data of NCI 60 cancer cell lines to 118 anticancer drugs (Scherf *et al.*, 2000).

2. DATA AND METHODS

2.1. Data pre-processing and formation of affiliation matrix

2.1.1. Rosetta compendium dataset

Rosetta Compendium dataset was downloaded from ExpressDB (Asch *et al.*, 2000). 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 (Ihmels *et al.*, 2002). We pooled the log ratio values to get a cutoff for binarization process. Arbitrarily 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.

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

$$A_{ij} = \begin{cases} 0 & (\text{if} \quad Q_{0.05} \le E_{ij} \le Q_{0.95}) \\ \\ 1 & (\text{if } E_{ij} < Q_{0.05} \text{ or } E_{ij} > Q_{0.95}) \end{cases}$$

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.

2.1.2. Multi-protein complex dataset

Gavin *et al.* used tandem-affinity purification (TAP) and mass spectrometry on a large scale and identified 232 distinct protein complexes in yeast (Gavin *et al.,* 2002). A total of 1353 genes constitute the complexes. The 232 protein complexes are roughly assigned into nine functional groups according to YPD and by literature mining in the original article. The numbers in the parentheses are the number of protein complexes within each functional group.

(1) Cell cycle (13)
(2) Polarity and structure (8)
(3) Intermediate and energy metabolism (43)
(4) Membrane biogenesis and turnover (20)
(5) Protein synthesis and turnover (33)
(6) Protein/RNA transport (12)
(7) RNA metabolism (28)
(8) Signaling (20)
(9) Transcription/ DNA maintenance/chromatin structure (55)

A protein complex is defined to be 'isolated' if its composing subunit proteins participate 'exclusively' in the specific complex. A protein complex is less isolated if its building blocks are parts of other complexes as well.

To quantify the degree to which a protein complex is isolated, the 'Isolation Index' is assigned for each complex *t*.

For a complex *t* there exists a set of genes, C_t = {gene | gene is a sub-component of protein complex t} (t=1,2, ··· , 232) and the 232 protein complexes are partitioned into nine subsets of different functional groups, S_f, f=1,2,···,9 as depicted above.

First of all, from the original protein complex dataset we generate a binary 1353-by-232 matrix **A =** $<a_{ij}>$,

$$a_{ij} = \begin{cases} 1 \text{ , if the gene i participates in the complex j.} \\ 0 \text{ , if the gene i does not participate in the complex j.} \end{cases}$$

2.1.2.1 Whole category Isolation Index ($I_{w,t}$)

Whole category Isolation Index is defined for each complex. For a protein complex *t*, $\mathbf{D_t}$ is the sub-matrix of **A**, where $\mathbf{D_t} = <a_{ij}>$**,** $i \in C_t$ and j = 1, 2,···, 232. We calculated the distance of the data matrix $\mathbf{D_t}$ from the ideal pattern matrix $\mathbf{P_t} = <p_{ij}>$,

$$p_{ij} = \begin{cases} 1 \text{ , } i \in C_t \text{ and } j = t. \\ 0 \text{ , } i \in C_t \text{ and } j \neq t. \end{cases}$$

, where $\mathbf{P_t}$ represents a data matrix of an ideal protein complex whose components are parts of only the protein complex out of 232 complexes.

Pearson correlation coefficient is used as the distance measure and the whole category Isolation Index ($I_{w,t}$) is defined as the Pearson correlation coefficient between $\mathbf{D_t}$ and $\mathbf{P_t}$. It ranges from −1 to 1 and the value closer to 1 signifies that the specific complex is more isolated among the 232 protein complexes.

2.1.2.2 Intra category Isolation Index ($I_{i,t}$)

In contrast to the whole category Isolation Index, intra category Isolation Index is defined considering group membership of each complex. Assuming a protein complex *t* belongs to the functional group *k*, the data matrix $\mathbf{D_t}'$ is the sub-matrix of **A**, where $\mathbf{D_t}'$ = $<a_{ij}>$, $i \in C_t$ and $j \in S_k$. The ideal pattern matrix is $\mathbf{P_t}'$ = $<p_{ij}'>$,

$$p_{ij'} = \begin{cases} 1 \text{ , } i \in C_t \text{ and } j = t. \\ 0 \text{ , } i \in C_t \text{ and } j \neq t \text{ and } j \in S_k. \end{cases}$$

The Pearson correlation coefficient between the matrices $\mathbf{D_t}'$ and $\mathbf{P_t}'$ is the intra category Isolation Index ($I_{i,t}$) and the index closer to 1 means that the protein complex is isolated and has less subunit proteins in common with other protein complexes within the same functional category.

2.1.3. Chemotherapeutic susceptibility data of NCI 60 cell lines

Chemotherapeutic susceptibility of NCI 60 cancer cell lines to 118 anticancer drugs were measured to give 60 by 118 matrix. The matrix elements are $-\log_{10}(GI_{50})$ where GI_{50} is the 50% growth inhibitory activities of the 118 drugs in each cell line and the 118 drugs are chosen because their mechanisms of action are putatively understood and used for cluster analysis (Scherf *et al.*, 2000).

A higher $-\log_{10}(GI_{50})$ value means the cell line is more susceptible to the drug and the $-\log_{10}(GI_{50})$ data matrix is binarized into affiliation matrix by giving 1 if a chemotherapeutic susceptibility of a cell line is more than the (0.8*s.d.+ mean) value for each drug and 0 otherwise.(Staunton *et al.,* 2001) Geodesic distance matrix is formed from the affiliation matrix. Based on the geodesic distance matrix, metric multidimensional scaling into 2 dimensions was performed for visualization.

2.2. Analysis of affiliation network

Affiliation matrix. Generally, an affiliation network consists of two key elements: a set of actors and a collection of subsets of actors (called *events*). It is represented as affiliation matrix A with elements,

$$A_{ij} = \begin{cases} 1, \text{ if actor i participates in event j} \\ \\ 0, \text{ if actor i does not participate in event j} \end{cases}$$

(see Fig 1. for example)

Bipartite matrix. In analysis of affiliation network, this matrix A is tranformed into a bipartite square matrix B given by, (given N actors, M events and O representing zero matrix)

$$\begin{pmatrix} O(N \times N) & A(N \times M) \\ \\ A'(M \times N) & O(M \times M) \end{pmatrix}$$

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. This representation preserves the whole informations in two-mode network. Figure 2 shows an example of the bipartite graph representation. This representation preserves the whole informations in two-mode network.

Geodesic. 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 graph. Geodesic distance matrix can also be drawn from a bipartite graph. (See figure 3. for geodesic distance matrix of data in figure 1.)

2.2.1. Rates of participation

Rate of participation of actor *i* is given by $\sum_j A_{ij}$ which implies how many events an actor participates in. The more sociable an actor is, the more events will he or she participate in. Likewise, the more interactions a gene has, the more cellular processes will it participate in and might have a longer evolutionary history.

2.2.2. Size of events

Size of event *j* is given by $\sum_i A_{ij}$ which implies how many actors participate in the event *j*. The larger events or cellular processes may be those that facilitate interactions among actors or genes.

2.2.3. Node centrality measures and group centralization measures

For detailed description of the concept of centrality, refer to (Wasserman and Faust, 1994;Scott, 2000;Faust, 1997;Borgatti and Everett, 1997). 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 (Borgatti *et al.*, 2002).

2.2.3.1. Node Degree centrality

This is the simplest definition of node centrality. The cental 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 $\sum_{j} B_{ij}$.

2.2.3.2. 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.(Freeman, 1979)

Closeness centrality of an actor *i* is given by $\left[\sum_{j} G_{ij}\right]^{-1}$.

2.2.3.3 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 ***gjk*** be the number of geodesics linking the two nodes *j* and k. Let ***gjk***(*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 $\sum_{j<k} g_{jk}(i) / g_{jk}$.

2.2.4. 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:

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

The maximum is taken over all possible graphs and the measure is, of course, always between 0 and 1. The value 1 is attained when the graph is of the form in figure 4(a) and 0 is assigned for a graph with the form in figure 4(b).

2.2.5. 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 (Borgatti and Everett, 1999). 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, c_i, 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. On the society of yeast genome as revealed by perturbations

3.1.1. Whole genome view

3.1.1.1 Rate of participation

Table 1 shows the MIPS functional classifications of genes differentially expressed in more than 150 out of 300 conditions. These genes are "social stars" in yeast genome in that they participate in a large number of events and are more likely to interact with larger number of other genes. The functional categories of these 'star' genes are like the following.

1) Stress response
2) Amino acid biosysthesis
3) C-compound and carbohydrate biosynthesis
4) Small molecule transport
5) Osmoregulation

These functions are important for the survival of the yeast against various *environmental challenges* and these genes may have longer evolutionary history leading to large number of interaction with different kinds of genes (Park and Bolser, 2001). As Rung *et al.* showed in their disruption network approach, this rate of participation follows a power law (Rung *et al.*, 2002). The scale-free model predicts that the nodes that appeared early in the history of the network are the most connected ones (Barabasi and Albert, 1999).

3.1.1.2. Size of events

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 etc.*

Functional categories of deleted genes are mostly related to: 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. Although the specific kinds of genes whose deletion strongly 'wiggles' the whole cellular transcriptional system is somewhat different from those found by Featherstone et al. and Rung et al.

because we used different normalization process, it is not surprising that these are essential cellular processes that are always switched on irrespective of environmental stimuli (Featherstone and Broadie, 2002; Rung et al., 2002). A larger number of genes may well be up or down regulated by the knock-out of essential genes and 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 (Jeong et al., 2001). But this direct and indirect impact is difficult to differentiate (Featherstone and Broadie, 2002).

3.1.2. Analysis of genes participating in Energy related processes

We wanted to 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. 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) Energy transport (2)

3.1.2.1. Core/Periphery structure of Energy related genes

Figure 5 shows the distribution of coreness scores of genes in each functional categories. The core/periphery model looks for strongly connected component in a network and gives higher coreness scores for the more connected actors and events. 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 (Wagner and Fell, 2001).

3.1.2.2. Graph centralization index

A graph with higher centralization index is more like a 'star' graph (see figure 4a). According to table 2, the glyoxylate cycle gene group has the highest degree and closeness centrality indices and has the most 'star' like graph. In contrast, respiration process has the smallest degree, closeness and betweenness centrality indices and has the least star-like structure.

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.

3.1.3. TCA cycle

Now let us focus on one of the sub-networks of energy related processes, or TCA cycle. Figure 6. shows multidimesional scaling representation of TCA cycle related genes and conditions. Genes are given in its enzyme names and 52 conditions (labeled with numbers for simplicity) were those that contain more than 5 participating genes out of 20 genes.

Borgatti *et al.* pointed out geodesic distance matrix as an input for MDS gives good visualization results and makes it easy to draw rough conclusions at a glance (Borgatti and Everett, 1997). You'll find a core/periphery structure especially among genes or actors. Table 3 lists the coreness scores of the 20 actors. The core group contains succinate dehydrogenase complex (SDH1, SDH2 and SDH4; SDH3 had missing value in expression data and excluded in the analysis) and isocitrate dehydrogense complex (IDH1 and IDH2), fumarase, aconitase and citrate synthase. Genes assigned to the core group are well known TCA cycle related genes and those in the periphery group have hitherto unspecified role in TCA cycle (see figure 7) and refer to (Przybyla-Zawislak *et al.,* 1999). This might mean the core genes are more exclusively dedicated to a specific process than the peripheral genes.

3.2. Exploring the combinatorics of protein complexes

Figure 8(a) shows the scatter plots of $I_{w,t}$ and $I_{i,t}$ for each protein complex according to its functional group. Interestingly, we can observe two different kinds of complexes. Some protein complexes have low whole category Isolation Index but high intra category Isolation Index and are scattered in left upper panel in each plot. Other complexes have little difference in the two indices and are spread around the Y = X axis.

The former patterns of complexes are mostly found in processes like cell cycle, polarity and structure and signaling, while the latter patterns of complexes are the norms in the processes of transcription/DNA maintenance/chromatin structure and RNA metabolism.

Cell cycle and signaling associated proteins generally form small complexes of regulatory units. Their ultimate goal is to preserve, transfer or amplify signals. The enrichment of these processes in complexes that have significantly higher intra category Isolation Indices than the whole category Isolation Indices gives the lesson; though the same protein component seems to be used in different complexes with other functional annotations, the sharing of components with complexes within the same functional category is supposedly deterred to maximize and ensure the specificity of signaling.

On the other hand, the processes like transcription, mRNA processing and translation are carried out by large molecular machines of various transcriptional and translational complexes. These molecular machines carry out large numbers of functions in concert both in time and space. For example, in RNA polymerase II complex (reflected in a complex of 9th functional group whose whole category and intra category Isolation Indices are 0.746267 and 0.805473 respectively), different combinations of five protein components build the structures called jaw, clamp, cleft, shelf and funnel that works for its attachment to DNA, the maintenance of the RNA-DNA duplex, the access of template strand to the active site and its translocation along the strand (Cramer *et al.*, 2001; Gavin *et al.,* 2003).

From the point of parsimony or considering the diverse number of functions to be carried out with the limited number of protein components, it is of course waste of resources to invent different exclusively composed machines to perform so many different functions. Instead, the cell seems to have evolved to utilize different combinations of protein components to perform various tasks to keep its life go on.

Figure 8(b) gives a higher order map of nine functional groups; the closer the groups are, the more protein components are held in common. Transcription, RNA metabolism and protein synthesis and turnover are strongly interconnected from the point of shared components and occupy core positions. But the processes with the significantly different whole category and intra category Isolation Indices like polarity and structure, cell cycle and signaling are placed in the periphery.

This map of functional groups is in contrast with its counterpart map of functional groups based on direct binary protein interactions suggested by Schwikowski *et al.* (Schwikowski *et al.*, 2000). They found cell cycle control process shows the most interactions with other classes and is placed in core positions in the network of different functional groups. The different aspects reflected in each map account for the discordance.

Quantifying the combinatorial property of gene groups or protein groups such as clusters from microarray experiments or pathways with the above-introduced method will help in entangling the basically combinatorial phenomenon, the ‘pleiotropy’.

3.3. The chemotherapeutic susceptibility of cancer cell lines

Figure 9(a) shows the multidimensional scaling map of the geodesic distance among cell lines and chemotherapeutic drugs based on GI50 data and figure 9(b) gives the legends for each cell line and drug of figure 9(a). The strategy of multidimensional scaling of geodesic distance matrix has often been used to explore gene expression data.(Nilsson *et al.*, 2004)

In figure 9(a), plus signs are cell lines while the rectangular drugs are colored in different ways after their known mechanisms of action. According to the hierarchical clustering result by Staunton *et al.,* drugs with similar known mechanism of action assemble together and our multidimensional map expectedly also provides visual confirmation of this point in two dimensions (Staunton *et al.,* 2001). Like the result by Staunton *et al.,* tubulin inhibitors (in red) form a tight cluster. Geldanamycin (No. 115) and Bisantrene (No. 49) are among the tubulin inhibitors and this is proposed to be from the involvement of Geldanamycin in G1 cell cycle arrest like taxane, the popular tubulin inhibitor. And 5-FU (No. 86), an anti-metabolite, is known to inhibit the RNA synthesis not just the DNA synthesis and is placed close to the Rs (RNS synthesis inhibitor, brown color). Topoisomerase 1 inhibitors (T1) are scattered near DNA synthesis inhibitors (Ds) which reflects the fact Camptothecin's cytotoxicity is based on DNA replication process or 'replication fork encounter lesions.'

We can further investigate the relationship between the two different groups of cell lines and drugs. For example, 5-FU is adjacent to the colon cancer cell lines as if reflecting its common use in patients with colon cancer. The usually prescribed primary regimen for ovarian cancer is the combination of paclitaxel, cisplatin, carboplatin and ovarian cancer cell lines are mixed with taxol associated tubulin inhibitor (TU) group and cisplatin related alkylating agent (A7) group. But in determining the chemotherapeutic regimen for treating cancer patients, various factors other than cancer types are taken into account and these multidimensional scaling results have limitations in that it is solely based on the *in vitro* chemotherapeutic susceptibility of cell lines to drugs.

4. DISCUSSION

We have investigated various affiliation networks from systematic profiling experiments like the perturbation based microarray experiment, multi-protein complex purifications and chemotherapeutic susceptibility profiling of cancer cell lines. And we extracted strongly connected group in disruption networks by applying core/periphery model of social network analysis methodology, quantified the isolatedness of a molecular complex and visualized the relationship between cancer cell lines and drugs. These approaches have revealed biologically relevant structures in various respects.

Of course, in addition to the hitherto investigated networks, there are numerous other two-moded data in biological literatures; transcriptional factors *vs.* their downstream regulated genes, structural motifs *vs.* protein molecules and many others.

Network analyses are usually conducted within each one mode and the structures between two modes are ignored after all. But thus-far introduced two modes investigation methodologies will add up insights from different viewpoints and be an indispensable step in deciphering the system of life.

5. REFERENCES

Asch,J., Rindone,W., Church,G.M. (2000) Systematic management and analysis of yeast gene expression data, *Genome Research,* **10**, 431-445.

Barabasi,A.L., Albert,R. (1999) Emergence of scaling in random networks, *Science,* **286**, 509-512, 1999

Barabasi,A.L., Oltvai,Z.N. (2004) Network Biology: Understanding the cell's functional organization, *Nature genetics,* **5**, 101-113.

Borgatti,S.P., Everett,M.G. (1997) Network analysis of 2-mode data, *Social Networks,* **19**, 243-269.

Borgatti,S.P., Everett,M.G. (1999) Models of Core/Periphery Structures, *Social Networks,* **21**, 375-395.

Borgatti,S.P., Everett,M.G., Freeman,L.C. (2002) Ucinet for Windows: Software for Social Network Analysis, Harvard Analytic Technologies, USA.

Butte,A.J., Tamayo,P., Slonim,D., Golub,T.R., Kohane,I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proc Natl Acad Sci U S A.* **97**(22), 12182-12186.

Cohen,B.A., Pilpel,Y., Mitra,R.D., Church,G.M. (2002) Discrimination between Paralogs using Microarray Analysis: Application to the Yap1p and Yap2p Transcriptional Networks, *Mol Biol Cell,***13**(5), 1608-1614.

Cramer,P., Bushnell,D.A., Kornberg,R.D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution, *Science* **292**(5523), 1863-1876.

Dezso,Z., Oltvai,Z.N., Barabasi,A.L. (2003) Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Res.,* **13**(11), 2450-2454.

Faust,K. (1997) Centrality in affiliation networks, *Social Networks,* **19**, 157-191.

Featherstone,D.E., Broadie,K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network, *BioEssays* **24**, 267-274.

Freeman,L.C. (1979) Centrality in social networks, *Social Networks,* **1**, 215-239.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M., Remor,M., Hofert,C., Schelder,M., Brajenovic,M., Ruffner,H., Merino,A., Klein,K., Hudak,M., Dickson,D., Rudi,T., Gnau,V., Bauch,A., Bastuck,S., Huhse,B., Leutwein,C., Heurtier,M.A., Copley,R.R., Edelmann,A., Querfurth,E., Rybin,V., Drewes,G., Raida,M., Bouwmeester,T., Bork,P., Seraphin,B., Kuster,B., Neubauer,G., Superti-Furga,G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,***415**(6868), 141-147.

Gavin,A.C., Superti-Furga,G. (2003) Protein complexes and proteome organization from yeast to man. *Curr Opin Chem Biol.,***7**(1), 21-27.

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. (2000) Functional discovery via a compendium of expression profiles, *Cell,***102**, 109-126.

Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K., Yang,L., Wolting,C., Donaldson,I., Schandorff,S., Shewnarane,J., Vo,M., Taggart,J., Goudreault,M., Muskat,B., Alfarano,C., Dewar,D., Lin,Z., Michalickova,K., Willems,A.R., Sassi,H., Nielsen,P.A., Rasmussen,K.J., Andersen,J.R., Johansen,L.E., Hansen,L.H., Jespersen,H., Podtelejnikov,A., Nielsen,E., Crawford,J., Poulsen,V., Sorensen,B.D., Matthiesen,J., Hendrickson,R.C., Gleeson,F., Pawson,T., Moran,M.F., Durocher,D., Mann,M., Hogue,C.W., Figeys,D., Tyers,M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature,***415**(6868), 180-183.

Ihmels,J., Friedlander,G., Bergmann,S., Sarig,O., Ziv,Y., Barkai,N. (2002) Revealing modular organization in the yeast transcriptional network, *Nature Genetics,* **31**, 370-377.

Jeong,H., Mason,S.P., Barabasi,A.L., Oltvai,Z.N. (2001) Lethality and centrality in protein networks, *Nature,* **411**, 41-42.

Maslov,S.A., Sneppen,K. (2002) Specificity and stability in topology of protein networks, *Science,* **296**, 910-913.

Neubauer,G., Gottschalk,A., Fabrizio,P., Seraphin,B., Luhrmann,R., Mann,M. (1997) Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc Natl Acad Sci U S A* **94**(2), 385-390.

Nilsson,J., Fioretos,T., Hoglund,M., Fontes,M. (2004) Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics* **20**(6), 874-880.

Park,J., Bolser,D. (2001) Conservation of Protein Interaction Network in Evolution, *Genome Informatics,* **12**, 135-140.

Przybyla-Zawislak,B., Gadde,D.M., Ducharme,K., McCammon,M.T. (1999) Genetic and biochemical interactions involving tricarboxylic acid cycle (TCA) function using a collection of mutants defective in all TCA cycle genes, *Genetics,***152**(1), 153-166.

Rout,M.P., Aitchison,J.D., Suprapto,A., Hjertaas,K., Zhao,Y, Chait,B.T. (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**(4), 635-651.

Rung,J., Schlitt,T., Brazma,A., Freivalds,K., Vilo,J. (2002) Building and analysing genome-wide gene disruption networks, *Bioinformatics,* **18** suppl. 2, 202-210.

Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T., Scudiero,D.A., Eisen,M.B., Sausville,E.A., Pommier,Y., Botstein,D., Brown,P.O., Weinstein,J.N. (2000) A gene expression database for the molecular pharmacology of cancer by John Weinstein, *Nat Genet.* **24**, 236-244.

Schwikowski,B., Uetz,P., Fields,S. (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol.*, **18**(12), 1257-1261.

Scott,J. (2000) Social Network Analysis, SAGE publications, USA.

Shmulevich,I., Zhang,W. (2002) Binary analysis and optimization-based normalization of gene expression data, *Bioinformatics,* **18**(4), 555-565.

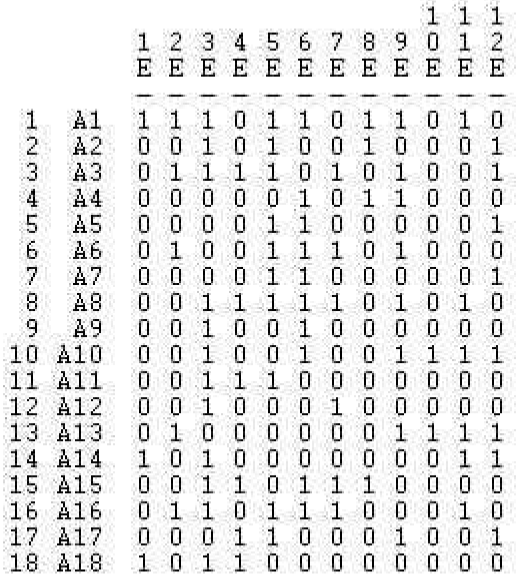
Staunton,J.E., Slonim,D.K., Coller,H.A., Tamayo,P., Angelo,M.J., Park,J., Scherf,U., Lee,J.K., Reinhold,W.O., Weinstein,J.N., Mesirov,J.P., Lander,E.S., Golub,T.R. (2001) Chemosensitivity prediction by transcriptional profiling, *Proc Natl Acad Sci U S A.* **98**(19), 10787-10792.

Verma,R., Chen,S., Feldman,R., Schieltz,D., Yates,J., Dohmen,J., Deshaies,R.J. (2000) Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol Biol Cell* **11**(10), 3425-3439.

Wagner,A., Fell,D.A. (2001) The small world inside large metabolic networks, *Proc. R. Soc. Lond.,* B **268**, 1803-1810.

Wasserman,S. and Faust,K. (1994) Social Network Analysis, Cambridge University Press, U.K.

Zachariae,W., Shin,T.H., Galova,M., Obermaier,B., Nasmyth,K. (1996) Identification of subunits of the anaphase-promoting complex of Saccharomyces cerevisiae. *Science* **274**(5290), 1201-1204.



```
                                    1 1 1
         1 2 3 4 5 6 7 8 9 0 1 2
         E E E E E E E E E E E E
         - - - - - - - - - - - -
 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 0 1 0 1 0 0 1
 4   A4  0 0 0 0 0 1 0 1 1 0 0 0
 5   A5  0 0 0 0 1 1 0 0 0 0 0 1
 6   A6  0 1 0 0 1 1 1 0 1 0 0 0
 7   A7  0 0 0 0 1 1 0 0 0 0 0 1
 8   A8  0 0 1 1 1 1 1 0 1 0 1 0
 9   A9  0 0 1 0 0 1 0 0 0 0 0 0
10  A10  0 0 1 0 0 1 0 0 1 1 1 1
11  A11  0 0 1 1 1 0 0 0 0 0 0 0
12  A12  0 0 1 0 0 0 1 0 0 0 0 0
13  A13  0 1 0 0 0 0 0 0 1 1 1 1
14  A14  1 0 1 0 0 0 0 0 0 0 1 1
15  A15  0 0 1 1 0 1 1 1 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. One example of an affiliation matrix with 18 actors and 12 events.

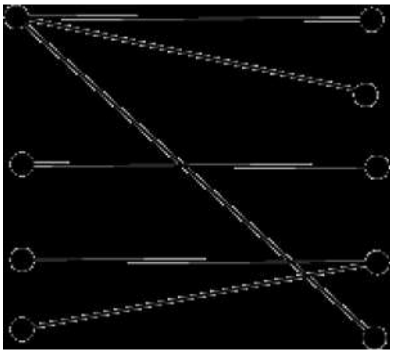
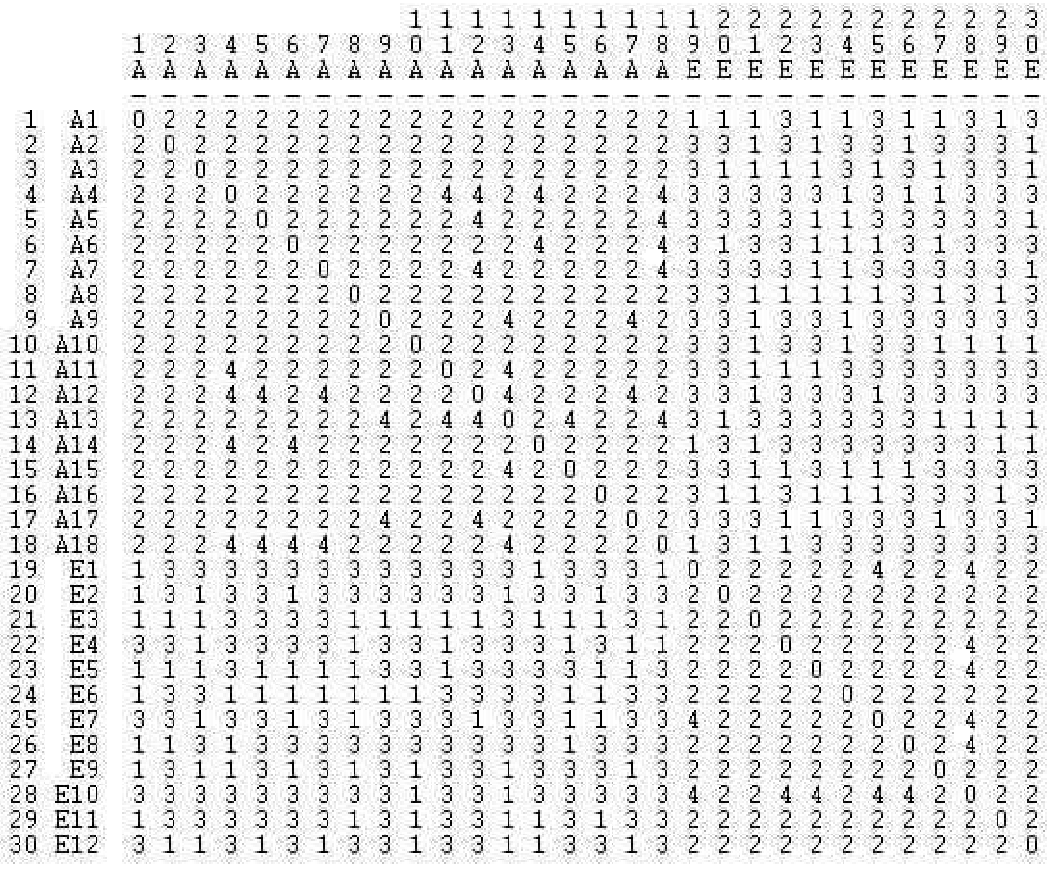
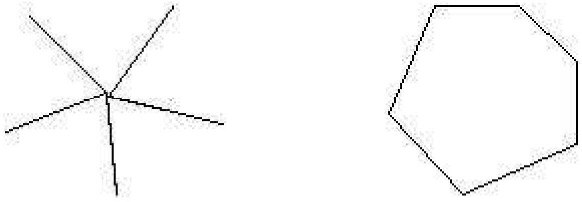


Fig. 2. Bipartite graph representation of an affiliation matrix. Left vertices are actors and right ones events.



```
                           1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3
         1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0
         A A A A A A A A A A A A A A A A A A E E E E E E E E E E E E
         - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 1   A1  0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 3 1 1 3 1 1 3 1 3
 2   A2  2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 1 3 3 1 3 3 3 1
 3   A3  2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 1 1 3 1 3 1 3 3 1
 4   A4  2 2 2 0 2 2 2 2 2 2 4 4 2 4 2 2 2 4 3 3 3 3 3 1 3 1 1 3 3 3
 5   A5  2 2 2 2 0 2 2 2 2 2 2 4 2 2 2 2 2 4 3 3 3 3 1 1 3 3 3 3 3 1
 6   A6  2 2 2 2 2 0 2 2 2 2 2 2 2 4 2 2 2 4 3 1 3 3 1 1 1 3 1 3 3 3
 7   A7  2 2 2 2 2 2 0 2 2 2 2 4 2 2 2 2 2 4 3 3 3 3 1 1 3 3 3 3 3 1
 8   A8  2 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 2 3 3 1 1 1 1 1 3 1 3 1 3
 9   A9  2 2 2 2 2 2 2 2 0 2 2 2 4 2 2 2 4 2 3 3 1 3 3 1 3 3 3 3 3 3
10  A10  2 2 2 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 3 3 1 3 3 1 3 3 1 1 1 1
11  A11  2 2 2 4 2 2 2 2 2 2 0 2 4 2 2 2 2 2 3 3 1 1 1 3 3 3 3 3 3 3
12  A12  2 2 2 4 4 2 4 2 2 2 2 0 4 2 2 2 4 2 3 3 1 3 3 3 1 3 3 3 3 3
13  A13  2 2 2 2 2 2 2 2 4 2 4 4 0 2 4 2 2 4 3 1 3 3 3 3 3 3 1 1 1 1
14  A14  2 2 2 4 2 4 2 2 2 2 2 2 2 0 2 2 2 2 1 3 1 3 3 3 3 3 3 3 1 1
15  A15  2 2 2 2 2 2 2 2 2 2 2 2 4 2 0 2 2 2 3 3 1 1 3 1 1 1 3 3 3 3
16  A16  2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 3 1 1 3 1 1 1 3 3 3 1 3
17  A17  2 2 2 2 2 2 2 2 4 2 2 4 2 2 2 2 0 2 3 3 3 1 1 3 3 3 1 3 3 1
18  A18  2 2 2 4 4 4 4 2 2 2 2 2 4 2 2 2 2 0 1 3 1 1 3 3 3 3 3 3 3 3
19   E1  1 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 1 0 2 2 2 2 2 4 2 2 4 2 2
20   E2  1 3 1 3 3 1 3 3 3 3 3 3 1 3 3 1 3 3 2 0 2 2 2 2 2 2 2 2 2 2
21   E3  1 1 1 3 3 3 3 1 1 1 1 1 3 1 1 1 3 1 2 2 0 2 2 2 2 2 2 2 2 2
22   E4  3 3 1 3 3 3 3 1 3 3 1 3 3 3 1 3 1 1 2 2 2 0 2 2 2 2 2 4 2 2
23   E5  1 1 1 3 1 1 1 1 3 3 1 3 3 3 3 1 1 3 2 2 2 2 0 2 2 2 2 4 2 2
24   E6  1 3 3 1 1 1 1 1 1 1 3 3 3 3 1 1 3 3 2 2 2 2 2 0 2 2 2 2 2 2
25   E7  3 3 1 3 3 1 3 1 3 3 3 1 3 3 1 1 3 3 4 2 2 2 2 2 0 2 2 4 2 2
26   E8  1 1 3 1 3 3 3 3 3 3 3 3 3 3 1 3 3 3 2 2 2 2 2 2 2 0 2 4 2 2
27   E9  1 3 1 1 3 1 3 1 3 1 3 3 1 3 3 3 1 3 2 2 2 2 2 2 2 2 0 2 2 2
28  E10  3 3 3 3 3 3 3 3 3 1 3 3 1 3 3 3 3 3 4 2 2 4 4 2 4 4 2 0 2 2
29  E11  1 3 3 3 3 3 3 1 3 1 3 3 1 1 3 1 3 3 2 2 2 2 2 2 2 2 2 2 0 2
30  E12  3 1 1 3 1 3 1 3 3 1 3 3 1 1 3 3 1 3 2 2 2 2 2 2 2 2 2 2 2 0
```

Fig. 3. Geodesic distance matrix.



(a)star graph(5 nodes)   (b)circular graph(5 nodes)

Fig. 4. Graph structure and centralization index.

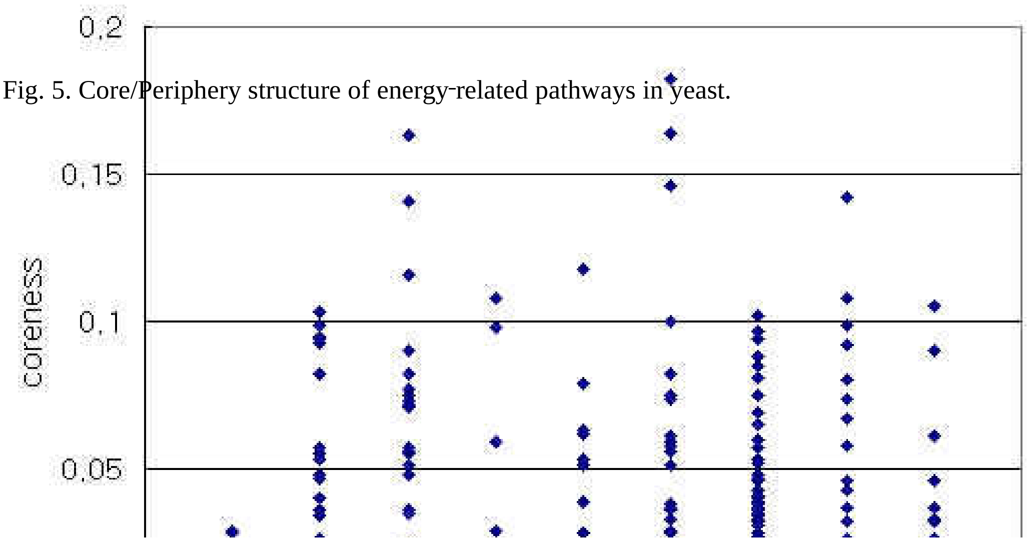


Fig. 5. Core/Periphery structure of energy-related pathways in yeast.

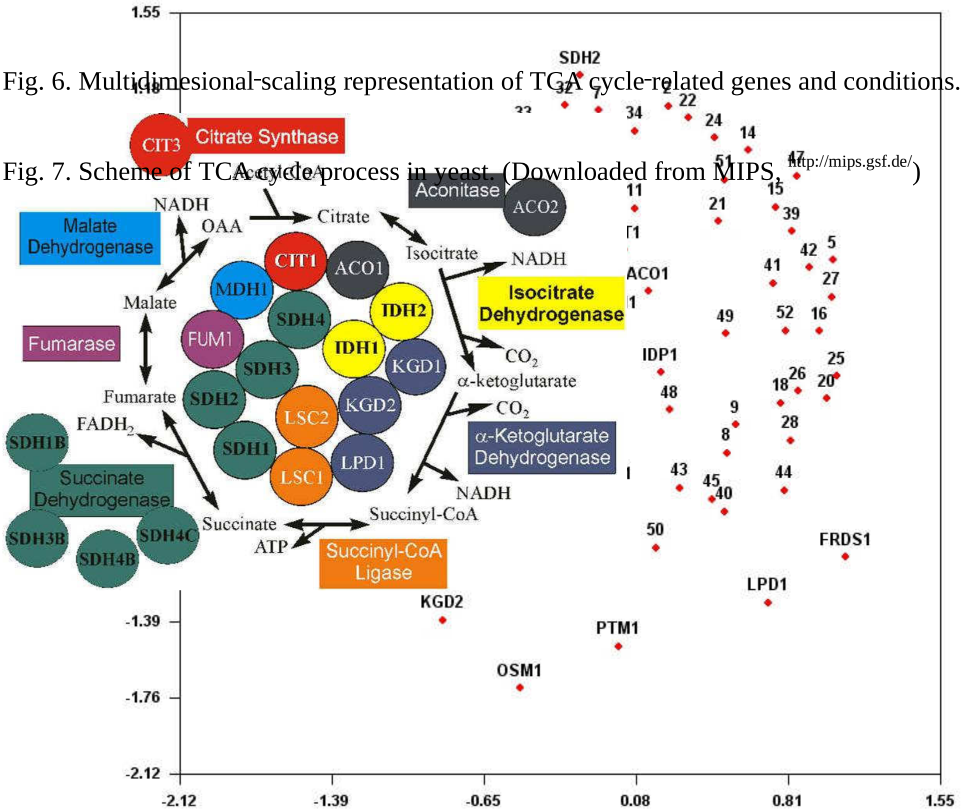
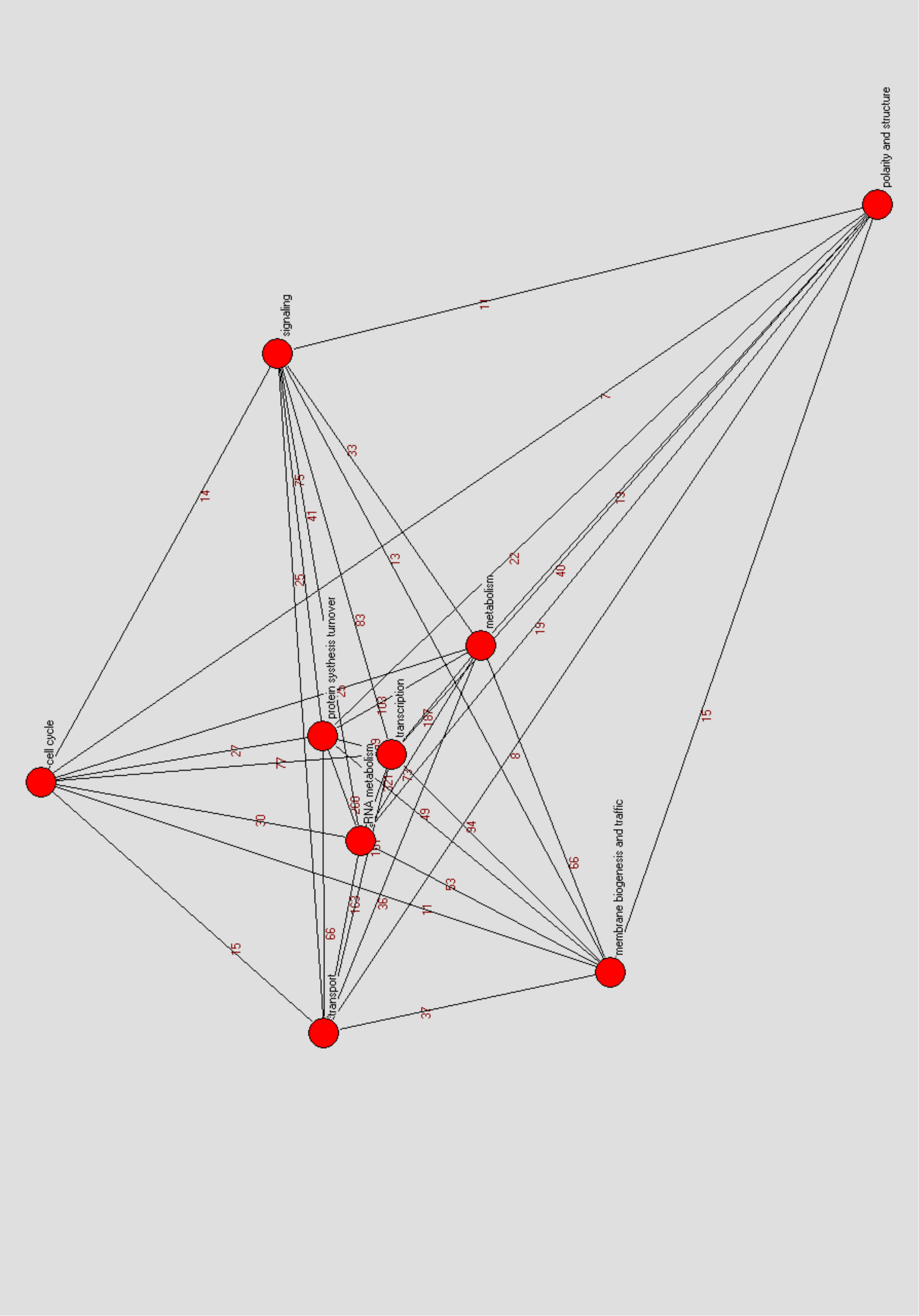


Fig. 6. Multidimesional-scaling representation of TCA cycle-related genes and conditions.

Fig. 7. Scheme of TCA cycle process in yeast. (Downloaded from MIPS, http://mips.gsf.de/)

Fig. ８（ａ）. Scatter plots of whole category ($I_{w,t}$) and intra category ($I_{i,t}$) isolation indices for each complex according to nine functional groups.


polarity and structure
signaling
metabolism
protein systhesis turnover
transcription
RNA metabolism
cell cycle
membrane biogenesis and trafic
transport
11
7
33
14
75
41
13
13
22
40
19
25
83
29
103
187
27
77
73
121
49
8
34
66
260
101
30
53
11
163
36
66
15
15
37

of

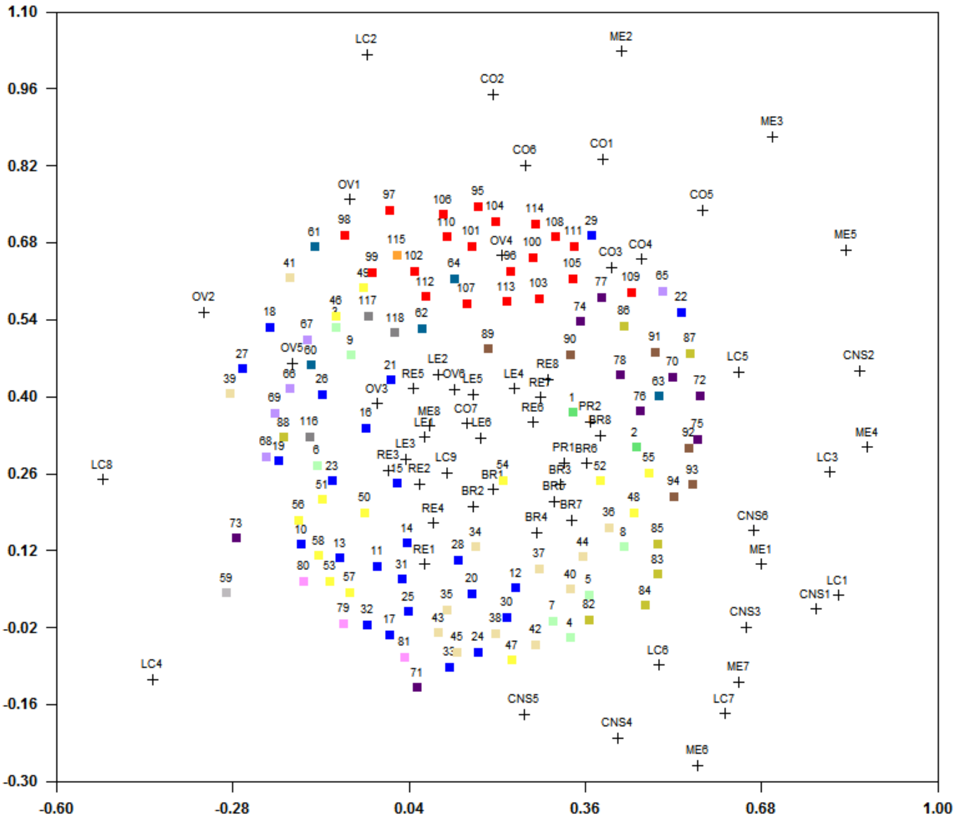


Fig. 9(a). Metric 2 dimensional MDS of geodesic distance matrix

1:A2, Mitomycin
2:A2, Porfiromycin
3:A6, Carmustine (BCNU)
4:A6, Chlorozotocin
5:A6, Clomesone
6:A6, Lomustine (CCNU)
7:A6, Mitozolamide
8:A6, PCNU
9:A6, Semustine (MeCCNU)
10:A7 ,Asaley
11:A7, Busulfan
12:A7, Carboplatin
13:A7, Chlorambucil
14:A7, Cisplatin
15:A7, Cyclodisone
16:A7, Diaminocyclohexyl-Pt-II
17:A7, Dianhydrogalactitol
18:A7, Diaziridinylbenzoquinone
19:A7, Fluorodopan
20:A7, Hepsulfam
21:A7, Iproplatin
22:A7, Mechlorethamine
23:A7, Melphalan
24:A7, Piperazine mustard
25:A7, Piperazinedione
26:A7, Pipobroman
27:A7, Spiromustine
28:A7, Teroxirone
29:A7, Tetraplatin
30:A7, Thiotepa
31:A7, Triethylenemelamine
32:A7, Uracil mustard
33:A7, Yoshi-864

60:Db, Cyanomorpholinodoxorubicin
61:Db, Hycanthone
62:Db, Morpholino-adriamycin
63:Db, N-N-Dibenzyl-daunomycin
64:Db, Pyrazoloacridine
65:Di, 5-6-Dihydro-5-azacytidine
66:Di, alpha-2'-Deoxythioguanosine
67:Di, Azacytidine
68:Di, beta-2'-Deoxythioguanosine
69:Di, Thioguanine
70:Df, Aminopterin
71:Df, Aminopterin-derivative
72:Df, Aminopterin-derivative
73:Df, an-antifol
74:Df, an-antifol
75:Df, Baker's-soluble-antifolate
76:Df, Methotrexate
77:Df, Methotrexate-derivative
78:Df, Trimetrexate
79:Dr, Guanazole
80:Dr, Hydroxyurea
81:Dr, Pyrazoloimidazole
82:Ds, Aphidicolin-glycinate
83:Ds, Cyclocytidine
84:Ds, Cytarabine (araC)
85:Ds, Floxuridine (FUdR)
86:Ds, Fluorouracil (5FU)
87:Ds, Ftorafur
88:Ds, Thiopurine (6MP)
89:Rs, Aciicin
90:Rs, Dichloroallyl-lawsone
91:Rs, DUP785 (brequinar)
92:Rs, L-Alanosine

BR1-BR:MCF7
BR2-BR:MCF7/ADF-RES
BR3-BR:MDA-MB-231/ATCC
BR4-BR:HS578T
BR5-BR:MDA-MB-435
BR6-BR:MDA-N
BR7-BR:BT-549
BR8-BR:T-47D
CNS1-CNS:SNB-19
CNS2-CNS:SNB-75
CNS3-CNS:U251
CNS4-CNS:SF-268
CNS5-CNS:SF-295
CNS6-CNS:SF-539
CO1-CO:HT29
CO2-CO:HCC-2998
CO3-CO:HCT-116
CO4-CO:SW-620
CO5-CO:HCT-15
CO6-CO:KM12
CO7-CO:COLO205
LC1-LC:NCI-H23
LC2-LC:NCI-H522
LC3-LC:A549/ATCC
LC4-LC:EKVX
LC5-LC:NCI-H322M
LC6-LC:NCI-H460
LC7-LC:HOP-62
LC8-LC:HOP-92
LC9-LC:NCI-H226
LE1-LE:CCRF-CEM
LE2-LE:K-562
LE3-LE:MOLT-4

34:T1, Camptothecin
35:T1, Camptothecin,7-Cl
36:T1, Camptothecin,9-MeO
37:T1, Camptothecin,9-NH2 (RS)
38:T1, Camptothecin,9-NH2 (S)
39:T1, Camptothecin,10-OH
40:T1, Camptothecin,11-formyl (RS)
41:T1, Camptothecin,11-HOMe (RS)
42:T1, Camptothecin,20-ester (S)
43:T1, Camptothecin,20-ester (S)
44:T1, Camptothecin,20-ester (S)
45:T1, Camptothecin,20-ester (S)
46:T2, Amonafide
47:T2, Amscrine
48:T2, Anthrapyrazole-derivative
49:T2, Bisantrene
50:T2, Daunorubicin
51:T2, Deoxydoxorubicin
52:T2, Doxorubicin
53:T2, Etoposide
54:T2, Menogaril
55:T2, Mitoxantrone
56:T2, Oxanthrazole (piroxantrone)
57:T2, Teniposide
58:T2, Zorubicin (Rubidazone)
59:Pi, L-Asparaginase

93:Rs, N-phosphonoacetyl-L-aspartic-acid
94:Rs, Pyrazofurin
95:TU, Colchicine
96:TU, Colchicine-derivative
97:TU, Dolastatin-10
98:TU, Halichondrin B
99:TU, Maytansine
100:TU, Trityl-cysteine
101:TU, Vinblastine-sulfate
102:TU, Vincristine-sulfate
103:TU, Taxol (Paclitaxel)
104:TU, Taxol analog
105:TU, Taxol analog
106:TU, Taxol analog
107:TU, Taxol analog
108:TU, Taxol analog
109:TU, Taxol analog
110:TU, Taxol analog
111:TU, Taxol analog
112:TU, Taxol analog
113:TU, Taxol analog
114:TU, Taxol analog
115:P90, Geldanamycin
116:Uk, 3-Hydropicolinaldehyde-thiosemicarbazone
117:Uk, 5-Hydroxypicolinaldehyde-thiosemicarbazone
118:Uk, Inosine-glycodialdehyde

LE4-LE:SR
LE5-LE:HL-60
LE6-LE:RPMI-8226
ME1-ME:LOXIMVI
ME2-ME:MALME-3M
ME3-ME:SK-MEL-2
ME4-ME:SK-MEL-5
ME5-ME:SK-MEL-28
ME6-ME:M14
ME7-ME:UACC-62
ME8-ME:UACC-257
OV1-OV:OVCAR-3
OV2-OV:OVCAR-4
OV3-OV:OVCAR-8
OV4-OV:IGROV1
OV5-OV:SK-OV-3
OV6-OV:OVCAR-5
PR1-PR:PC-3
PR2-PR:DU-145
RE1-RE:UO-31
RE2-RE:SN12C
RE3-RE:A498
RE4-RE:CAKI-1
RE5-RE:RXF-393
RE6-RE:786-0
RE7-RE:ACHN

RE8-RE:TK-10

Alkylating agents: A2, A7 = alkylating at N-2, N-7 position of guanine, respectively; A6 = alkylating at O-6 position of guanine; T1 = topoisomerase I inhibitor; T2 = topoisomerase II inhibitor; Db = DNA binder; Di = DNA incorporation; Df: antifols; Dr = ribonucleotide reductase inhibitor; Ds = DNA synthesis inhibitor; Rs = RNA synthesis inhibitor; Tu = tubulin-active antimitotic agents; Pi = protein synthesis inhibitor; P90 = hsp90 binder; Uk = unknown

BR: breast cancer CNS: CNS tumor CO: colon cancer LC: lung cancer LE: leukemia ME: melanoma OV: ovarian cancer PR: prostate cancer

RE: renal cell cancinoma   Cell lines are represented with plus signs

Fig. 9(b). Legend for figure 9(a)

Table1. MIPS functional category of genes differentially expressed in more than 150 conditions

YBR072W	
11.01	STRESS RESPONSE
YBR145W	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
2.16	FERMENTATION
11.07	DETOXIFICATION
YBR296C	
01.04.07	PHOSPHATE TRANSPORT
13.01.01.99	HOMEOSTASIS OF OTHER CATIONS
13.01.03.03	HOMEOSTASIS OF PHOSPHATE
67.04.07	ANION TRANSPORTERS (CL, SO4, PO4, ETC.)
YCL018W	
01.01.01	AMINO ACID BIOSYNTHESIS
YER069W	
01.01.01	AMINO ACID BIOSYNTHESIS
YFL014W	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
01.06.07	LIPID, FATTY-ACID AND ISOPRENOID UTILIZATION
03.01.05.01	DNA REPAIR
11.01	STRESS RESPONSE
13.11.03.13	OSMOSENSING
YFR030W	
01.01.01	AMINO ACID BIOSYNTHESIS
01.02.01	NITROGEN AND SULFUR UTILIZATION
YFR053C	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
2.01	GLYCOLYSIS AND GLUCONEOGENESIS
YGL255W	
13.01.01.01	HOMEOSTASIS OF METAL IONS (NA, K, CA ETC.)
67.04.01.01	HEAVY METAL ION TRANSPORTERS (CU, FE, ETC.)
YHR018C	
01.01.01	AMINO ACID BIOSYNTHESIS
YHR137W	
01.01.04	REGULATION OF AMINO ACID METABOLISM
YHR215W	
01.04.01	PHOSPHATE UTILIZATION
40.27	EXTRACELLULAR / SECRETION PROTEINS
YIR034C	
01.01.01	AMINO ACID BIOSYNTHESIS
YJL088W	
01.01.01	AMINO ACID BIOSYNTHESIS
YJL153C	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
YJR025C	
01.01.10	AMINO ACID DEGRADATION (CATABOLISM)
01.07.01	BIOSYNTHESIS OF VITAMINS, COFACTORS, AND PROSTHETIC GROUPS
YML123C	
01.04.07	PHOSPHATE TRANSPORT
8.19	CELLULAR IMPORT
13.01.01.03	HOMEOSTASIS OF PROTONS
13.01.03.03	HOMEOSTASIS OF PHOSPHATE
67.04.07	ANION TRANSPORTERS (CL, SO4, PO4, ETC.)
67.07	C-COMPOUND AND CARBOHYDRATE TRANSPORTERS
YMR062C	
01.01.01	AMINO ACID BIOSYNTHESIS
YMR094W	
03.03.01	MITOTIC CELL CYCLE AND CELL CYCLE CONTROL
YMR095C	
11.01	STRESS RESPONSE
YMR096W	
3.99	OTHER CELL DIVISION AND DNA SYNTHESIS ACTIVITIES
11.01	STRESS RESPONSE
YMR105C	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
2.19	METABOLISM OF ENERGY RESERVES (GLYCOGEN, TREHALOSE)
YNL036W	
8.16	EXTRACELLULAR TRANSPORT, EXOCYTOSIS AND SECRETION
YNL160W	
11.01	STRESS RESPONSE
40.27	EXTRACELLULAR / SECRETION PROTEINS
YOL058W	
01.01.01	AMINO ACID BIOSYNTHESIS
01.02.01	NITROGEN AND SULFUR UTILIZATION
YPL019C	
8.13	VACUOLAR TRANSPORT
YPR160W	
01.05.01	C-COMPOUND AND CARBOHYDRATE UTILIZATION
2.19	METABOLISM OF ENERGY RESERVES (GLYCOGEN, TREHALOSE)
YPR167C	
01.01.01	AMINO ACID BIOSYNTHESIS
01.02.01	NITROGEN AND SULFUR UTILIZATION
YJR109C	
01.01.01	AMINO ACID BIOSYNTHESIS
YKL001C	
01.01.01	AMINO ACID BIOSYNTHESIS
01.02.01	NITROGEN AND SULFUR UTILIZATION
YKL096W	
40.01	CELL WALL
YLR303W	
01.01.01	AMINO ACID BIOSYNTHESIS

Table 2. Graph centralization index

Process	Degree centr.(%)(homogeniety)	Closeness centr.(%)	Betweenness centr.(%)
fatty acid oxidation	26.22 (5.22%)	26.99	**66.13**
fermentation	28.32 (1.46%)	18.99	16.29
glycolysis	45.81 (1.50%)	24.84	24.74
glyoxylate cycle	**52.39** (6.66%)	33.23	49.35
pentose phosphate	45.56 (3.54%)	30.84	46.38
energy reserves	49.94 (1.41%)	31.23	24.50
respiration	22.78 (0.64%)	21.16	8.57
TCA cycle	43.79 (1.96%)	26.99	33.68
All actors	26.49 (0.29%)	21.59	4.11

Table 3. Coreness scores of 20 actors of TCA cycle.

Core		Periphery	
SDH1	0.308	FRDS1	0.144
ACO1	0.308	LPD1	0.144
CIT1	0.298	MDH1	0.135
IDP1	0.288	LSC1	0.125
LSC2	0.269	PTM1	0.106
IDH2	0.259	KGD1	0.086
IDH1	0.211	KGD2	0.067
SDH4	0.202	CIT3	0.067
FUM1	0.192	OSM1	0.048
SDH2	0.163	IDP2	0.048

초록

사회 연결망 분석 용어에서, 소속 연결망이란 두 가지 상이한 노드 집단의 연결망을 일컬으며 이는 다양한 생물학 연결망에서 발견된다. 본 연구에서는 세가지 종류의 소속 연결망을 탐색하여 생물학적으로 의미있는 구조를 추출해 내었다.

첫 번째 소속 연결망은 총체적으로 효모 유전자에 동요를 일으켜 전, 후의 유전자의 발현량을 측정한 마이크로어레이 실험으로부터 구성되며 다양한 사회 연결망 분석의 도구들, 예를 들어 밀도, 핵심성 및 중심성 측도들을 적용하였다. 많은 동요에 영향 받는 유전자들은 여러 가지 환경적 자극으로부터 효모의 생존을 보장하기 위한 기능들을 가지고 있다. 반면에 효모의 일반적인 생명 유지를 위한 프로세스에 참여하는 유전자들을 돌연변이 시켰을 때 동요의 파급 범위가 컸다. 여러 개의 서브 연결망으로 이루어진 연결망을 핵심-주변부 모델을 적용하여 진화적 기원이 오랜 프로세스들이 핵심에 존재함을 발견했다. 해당작용 및 크렙스 회로는 에너지 연관된 프로세스에서 상대적으로 핵심적 위치를 점하고 있다.

두 번째 소속 연결망은 총체적인 단백질 복합체 분리 실험에서 구성했으며 이를 통해 세포 구성의 조합적 특성을 탐구하였다. 세포 주기, 신호 전달 및 세포 구조 관련 단백질 복합체들은 각 기능 군 내에서 상호 단백질 공유가 적은 반면에 단백질 합성, 전사 및 RNA 대사 관련 단백질 복합체들은 기능 군 내, 외에서 공히 단백질 공유가 많았다. 이러한 조합적 특성에 기반하여 고차원적인 각 기능군의 지도를 그렸을 때 전자의 복합체들은 주변부에 존재한 반면, 후자의 복합체들은 핵심부에 존재하였다.

마지막 소속 연결망은 세포 주의 화학요법에 대한 순응도 데이터로부터 구성했으며, 소속 연결망의 최단거리 행렬을 다차원척도 기법으로 시각화했을 때, 비슷한 작용 메커니즘을 갖는 약물들이 군집을 이루었다.

주요어: 사회 소속 연결망 분석, 중심성, 핵심 주변부 모델, 단백질 복합체, 최단 거리
학번: 2002-20630

<감사의 글>

의과대학 졸업반 시절이었습니다. 재학 기간 중 방학이면 연구실에 나가 다양한 연구에 참여하면서 기존의 분자생물학의 환원론적 접근방식만으로는 적절한 치료법을 개발하는 데 한계가 있으며, 생명현상을 시스템 수준에서 이해하는 접근이 필수적이라는 것을 느꼈습니다. 하지만 6년간 환자 치료하는 법만을 배운 저로서는 이러한 접근법에 필수적인 수리, 전산, 통계분야의 기초가 없었고 따라서 관련 저널을 이해하는 데조차 큰 어려움이 있었습니다.

뜻이 있는 곳에 길이 있다고, 마침 미국에서 귀국하신 김주한 교수님께 연이 닿아 이듬해에는 통계학과, 컴퓨터 공학과의 핵심 교과목들을 수강하며 생물정보학을 공부할 기회를 얻게 되었습니다. 일년간 다양한 분야의 다양한 사람들을 접하면서 앞으로 시스템 수준에서 생명현상을 이해할 수 있는 기초를 쌓아나갈 수 있었습니다.

의사로서의 수련을 잠시 중단하고 공부에 매진했지만 진료능력이 부족하다는 점은 저를 초조하게 만들었고 이듬해 인턴 수련을 받게 되었습니다. 의사로서의 능력을 키울 수 있었던 2003년 한 해는 대학원 시절만큼이나 보람찬 시기였습니다.

인턴과정을 수료하고 지금은 공중보건의사로 근무하고 있습니다. 낮에 진료하고 저녁에는 연구와 논문작업으로 바쁜 한 해였지만 부족하나마 그 결실을 보게 되었습니다. 생물정보학 논문을 이해조차 못하던 3년 전 저의 모습이 생각납니다. 아직 많이 부족하지만 오늘의 제가 있기까지 도움을 주셨던 수많은 분들에게 이 자리를 빌어 깊은 감사의 뜻을 전하고 싶습니다.

우선, 생물정보학이란 학문과 인연을 맺게 해주시고 대학원 기간 내내 아낌없는 조언으로 이끌어주시고 지켜봐 주신 김주한 교수님께 감사를 드립니다. 또한 지도교수로서 학문 내, 외로 도움을 많이 주신 김각균 교수님 그리고 심사위원으로서 발전적인 평가와 조언으로 자리를 빛내주신 조광현 교수님께 감사를 드립니다.

또한 진정한 동료로서 학문을 향한 열정을 함께한 스누비 연구실 사람들 모두에게 특별한 감사를 드립니다. 지금은 위스콘신에서 학문의 열정을 불태우고 있을 후니 김지훈 형, 같은 의사로서 연구에 매진하고 있으며 항상 든든한 친구인 김민구, 너털웃음이 여유로워 좋은 정희준, 공간적 거리로 토론이 너무 아쉬웠던 정태수 박사님, 가건물의 연구실로 갈 때면 항상 밤 늦게 같이 자리를 지켜주던 한미령 씨, 정말 착한 김미현 씨, 생물정보학에 대한 열정이 인상 깊은 나영지 씨, 가끔은 엉뚱하시지만 정말 따뜻한 서화정 누나, 스누비의 큰 형님 김지훈 형, 스누비의 잘생긴 기둥 박찬희 형, 만년 소녀 박유랑 씨, 재치 만점에 항상 웃는 모습인 이혜원 씨, 나만큼 방황하던 천재 김기원, 얘기는 별로 나눠보지 못했지만 날카로운 질문으로 세미나 분위기를 돋우는 김옥구 씨, 말이 필요 없는 우리 누님 윤혜성 누나, 목소리 만큼이나 마음도 따뜻하신 조성범 선생님, 항상 이십 대이실 것 같은 박지연 누나, 아기 얘기만 나오면 눈이 빛나시는 이정애 누나, 지금은 텍사스에서 역시 학문의 열정을 불태우고 있을 이석호 박사님, 부드러운 부산 사투리가 인상 깊은 김세영 씨, 한 번 뵈었지만 열정을 느낄 수 있었던 홍승권 선생님, 아직 한번도 이야기를 못 해본 이영주 씨... 이분들과 더불어 행복했던 연구실 생활이 기억납니다.

같이 생물정보학이란 미지의 세계로 도전장을 던진 협동 과정 1기 친구들, 사려 깊고 순수한 수학자 조병훈 씨, 항상 올 에이로 성적표를 도배하던 김수영, 고등학교 동창인 걸

뒤늦게 알고 서로 놀란 윤규만, 재치 만점에 고전 미인 윤단규 씨, 항상 웃으시며 따뜻하신 정제균 형, 털털한 가운데 샤프한 허인애 씨, 소의 난자 채취로 항상 고생하시던 권용삼 형, 항상 웃는 얼굴의 남동현 형, 지금은 사업에 몸담고 있으며 앞으로 우리의 밥값을 책임질 미래의 갑부 김호규, 그리고 마음이 정말 바다같이 넓은 이수연... 어떻게 공부할 지 몰라 같이 헤매던 이들 덕에 의미 있는 대학원 시절을 보낼 수 있었습니다.

또한 올 한해 평택구치소에서 공중보건의사로 일하면서 여러모로 배려해 주신 최동한 과장님, 항상 쾌활한 이명혜 간호사님 그리고 묵묵히 의무과의 업무를 훌륭히 처리하시는 권윤식 주임님께도 감사의 뜻을 전합니다.

그리고 우리 사랑하는 가족들을 빼놓을 수 없습니다. 남들과 조금 다른 길을 선택할 때 아들을 믿어주고 뒷바라지 해주신 부모님, 멀리 미국의 아이오와에서 학문의 길을 가고 있는 누나 그리고 항상 든든한 동생에게 깊은 감사를 표합니다.

마지막으로, 제 삶의 단계단계마다 보이지 않는 손길로 이끌어주시는 하느님께 감사 드립니다.