# Network Descriptors to Capture the Dynamic Changes of Networks in Temporal Microarray Data Sets

Chen-hsiung Chan[1], Kuan-Yeu Pan[2], and Cheng-Yan Kao[1, 3]
*[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, [2]Bioengineering Research Center, National Tsing Hua University, Hsinchu, Taiwan, and [3]Institute for Information Industry, Taiwan*
*cykao@csie.ntu.edu.tw*

## ABSTRACT

*The analysis of temporal microarray datasets usually focused on the expression patterns of genes, and genes with similar expression patterns are clustered. However, the relationships among the clustered genes are not clear except for the plausible co-regulated connections. Also, there are no established methods to quantify the changes among the data sets in different stages. We have combined protein-protein interaction data with micorarray expression profiles to reveal the possible interactions among genes with/without similar expression patterns. Based on the features of networks, we propose a quantitative measure to evaluate the changes between two microarray datasets in different conditions or time points. The quantified scores could in turn be used as a criterion to identify significant sub-networks, which may be used to establish the functional connections to the specified conditions or factors.*

## 1: INTRODUCTIONS

The advances in microarray techniques have enabled high-throughput monitoring of state changes on various types of cells. Often, cells under different conditions (e.g. cancer *v.s.* normal or treated *v.s.* untreated) are monitored and compared for differences in mRNA expression levels. Most microarray studies focus on the expression patterns of genes, i.e. whether a set of genes show similar expression levels under similar conditions, and whether these genes are regulated by the same transcription factors. Only a few studies integrate microarray data and network information (regulatory network, metabolic pathways, or protein-protein interaction networks) and examine the relationships among a set of genes in interest [1].

As more and more biological data accumulated, network biology is becoming critical in the new systems biology era. Network biology deals with biological networks, whether they are regulatory networks, signal transduction networks, metabolic pathways, or physical interaction networks. Understanding the connections among network nodes (genes, proteins, or metabolites) will reveal insights to the underlying biological mechanisms of life.

In this work, we have employed human protein-protein interaction networks in the analysis of microarray data. The expression levels of mRNA (and thus protein) are obtained from temporal microarray experiments, and are used as modifiers of the protein interaction network. Considering these changes in expression levels, the protein-protein interaction networks change as well, resulting in dynamic networks corresponding to changes in biological conditions.

However, dynamic networks are not trivial to analyze. For example, how to quantify changes or differences between two networks under different conditions? Without a quantitative measure, it is not possible to evaluate such changes; and only qualitative or descriptive terms can be used. Most studies on network features and topologies focus on the ‘global’ characteristics of networks [2, 3]. However, not all biological processes have global effects. It is necessary to have a measure on both global and sub-network dynamics.

In this work, we proposed novel network descriptors capable of quantitative evaluation of dynamic biological network changes. We have incorporated concepts from information theory [4]. The nodes (proteins) and edges (interactions) are assigned predefined states. Based on the distribution of these states, two novel scores, node and edge entropies, can be calculated. These two scores reflect the quantitative changes of proteins and their interactions in different conditions. With these scores, it is now possible to evaluate whether a perturbation (treat with drugs or knock out genes) have global or local influences, and to what degree. Also, these scores can be used to identify sub-networks with maximum changes. That is, such sub-networks might be close related to the specified conditional changes or perturbations.

The rest of this paper is organized as follows: Section 2 outlines materials and methods used to derive our scores. Section 3 illustrates the application of our scores in a real biological scenario, the pitfalls and advantages of the new scores are also discussed. Section 4 concludes the novelty and potential impacts of this work.

## 2: MATERIALS AND METHODS

### 2.1: TEMPORAL MICROARRAY DATA SETS

Temporal microarray data are obtained from James *et al.* [5]. Chondrocytes from mouse were incubated and monitored with microarray. The differentiation and maturation of chondrocytes are critical to skeletal development. These chondrocytes were monitored every 3 days for 15 days, resulting in five microarray images. The microarray data were deposited into the GEO database [6]. GEO (Gene Expression Omnibus) provides a platform to share microarray data. The five sets of microarrays collected in day 3, 6, 9, 12, and 15 are designated as stage 0, 1, 2, 3, 4 in this work, respectively.

The microarray used in [5] are manufactured by Affymetrix. There are 22,690 probes representing nearly 14,000 mouse genes. A ruby (http://www.ruby-lang.org/) script was written to parse the raw data. Mouse genes are mapped to homologous human genes since only a few mouse protein-protein interaction data are available. The conversion is performed with conversion table provided by NCBI. Near 11,700 genes are converted into their human homologs.

In addition to the mammalian time course experiments, we also applied our method to a yeast time course dataset. Data from the yeast cell cycle experiment conducted by Spellman *et al.* [7] has been analyzed with our method. In these experiments, four datasets are available based on the synchronization scheme: α Factor, CDC15, CDC28, and elutriation. For brevity, we only show the results from the α Factor dataset. There are 18 time points in the α Factor dataset. These time points are spaced in 7 minutes. 800 cell cycle regulated genes were identified by Spellman *et al.* The sub-network formed by these cell cycle genes are constructed and compared to the entire yeast interactome.

### 2.2: PROTEIN INTERACTION DATA

Human protein-protein interaction data are obtained from the POINT database [8]. POINT contains more than 150,000 protein-protein interactions from various model organisms. There are more than 37,000 human protein-protein interaction data in POINT, participated by nearly human 10,000 proteins. POINT also contains more than 25,000 predicted protein interactions, but these predicted data are not used in this work.

The global network, referred as 'Human Proteome', is constructed with all available human protein-protein interactions.

Five sets of proteins are identified through literatures and by expert. These proteins are served as 'seeds' or 'queries' for a sub-network connected to these proteins. These proteins are closely related to chondrocyte differentiation. Therefore the sub-network formed by these proteins should reflect the majority of the changes occurred in the chondrocyte development process. The five categories include cartilage, collagen, extra cellular matrix, growth factor, and signaling proteins. Selected proteins in these categories are listed in Table 1.

| Category | Proteins |
|---|---|
| Cartilage | CDH11, FARP1, CRTAP, LECT1, CHI3L1, COMP, HAPLN1, CSPG2, CSPG4, AGC1, NID2, CHSY1, CHRDL2, CSS3, OMD, CHST11, CRTAC1, CHST12, ChGn, BMP8B, SPOCK1, CHPF, MIA, CILP, CHST3, MTFR1 |
| Collagen | CTHRC1, COL1A1, COL2A1, COL3A1, COL4A1, COL4A2, COL4A4, COL5A1, COL5A2, COL6A1, COL6A2, COL6A3, COL7A1, COL8A1, COL8A2, COL9A3, COL11A1, COL11A2, COL12A1, COL15A1, COL16A1, MMP1, MMP2, MMP13, COL5A3, PCOLCE |
| Extra Cellular Matrix | PRG4, CIB2, CIB1, LMAN2, ADRM1, ECM2, ECM1, ALCAM, ITGA11, ITGB3BP, ITGB1BP2, ITGA5, ITGA7, ITGAE, ITGB1, ITGB2, ITGB4BP, ITGB5, L1CAM, LOC391783, LGALS1, LGALS3BP, LGALS8, MCAM, CLEC4A, CEECAM1, PRG1, MASP1, JAM2, HAPLN2, CLEC11A, VCAM1, ASAM, ILKAP, LMAN2L, JAM3, MBL1P1, PAPLN, ITGB1BP1, ITGBL1, CD47, CD302, CLEC2B |
| Growth Factor | FRS2, OGFR, CTGF, HBEGF, EPS8, FGF1, FGF2, FGF7, FGF13, FGFR1, FGFR2, GFER, GRB2, GRB10, GRB14, HDGF, IGF1, IGF2, IGF2R, IGFBP1, IGFBP2, IGFBP3, IGFBP4, IGFBP5, IGFBP6, IGFBP7, LTBP1, LTBP2, LTBP3, NGFB, HDGFRP3, PDGFA, PDGFRA, PDGFRL, PDGFRB, FGFRL1, PDGFC, EPS15L1, TGFB1I1, TGFB2, TGFB3, TGFBI, VEGF, VEGFB, VEGFC, OGFRL1, HDGF2, HGS, FIBP, TBRG4, FGFBP1 |
| Signaling Protein | CMTM7, CLC, SOCS4, CMTM3, ATF2, CMTM4, DOCK1, DOCK9, CXCL1, HSPG2, IK, ATF1, CKLF, CMTM6, CMKOR1, CXCL16, CCL2, CCL20, CXCL12, SOX9, TIMP3, N-PAC, SOCS2, CBFA2T2, CRLF1, SCYE1, SOCS5 |

**Table 1. The selected proteins related to chondrocyte differentiation.**

The expert selected sub-network is generated based on these query genes. It includes the five sub-networks defined in previous paragraph, and also includes extra interactions among the five sub-networks. Proteins interacting with these queries are picked and included in this sub-network. We consider this sub-network is critical and may represent most of the changes in chondrocyte differentiation. Five other sub-networks based solely on one of the categories are also constructed.

For the yeast dataset, 45,146 interactions are available in the POINT database. The entire yeast interaction network is referred as 'Yeast Interactome'. The 800 cell cycle regulated genes are used as 'queries' to construct a cell cycle sub-network. This sub-network is referred as 'Cell Cycle Sub-Network'. The list of the 800 cell cycle genes can be found in [7].

## 2.3: EVALUATION OF STATE CHANGE

State changes obtained from microarray data are the change in expression levels. Using a baseline time point, the rest of the microarray states can be compared with this reference state. We use the stage 0 (day 3) data as the reference set. We define the states of nodes as follows:

$$S_i^N = \begin{cases} 1, \text{if the gene is up regulated} \\ 0, \text{if the gene is unchanged} \\ -1, \text{if the gene is down regulated} \end{cases} \quad (1)$$

where $S_i^N$ is the state of node *i*. A gene (node) is considered as up regulated if its expression level is more than 2 times higher than baseline. If the expression level is less than 1/2 of the baseline, the gene is considered as down regulated. A gene is considered as unchanged if the above two conditions are not met.

Similarly, the states of edges are defined as follows:

$$S_{ij}^E = S_i^N + S_j^N \quad (2)$$

where $S_{ij}^E$ is the edge between node *i* and *j*. According to the definition in eq. (1), edge states have five possible values: 2, 1, 0, -1, and -2, respectively. Edges in protein-protein interaction network are interactions between two proteins. The expression levels of mRNAs imply the abundance of the proteins in the cell. Based on Bayesian approaches, the probability of two proteins interact with each other can be estimated from the abundances of these proteins. In Eq. (2), the edge states are defined based on the states of the two nodes forming this edge. Since the node states are actually a logarithm of the expression ratio, summation of the node states approximates the logarithm on the product of expression ratios.

With the definitions of node and edge states, the networks can be seen as a collection of nodes and edges in different states. The distributions of these states can be used estimate the changes between different time points or conditions. Based on the formulation in information theory [4], the entropy or information contents of a information stream can be defined as follows:

$$H = -\sum_i p_i \ln p_i \quad (3)$$

where *H* is the information content (entropy), and $p_i$ is the frequency of observing state *i* in the entire population. From our definitions of node and edge states, we can calculate node and edge entropies accordingly. For example, the node entropy will be calculated as follows:

$$\begin{aligned} H^N &= -\sum_i p_{S_i^N} \ln p_{S_i^N} \\ &= -\left(p_1 \ln p_1 + p_0 \ln p_0 + p_{-1} \ln p_{-1}\right) \end{aligned} \quad (4)$$

## 2.4: RANDOMIZED NETWORKS

Using a bootstrap method, we have generated randomized networks to be compared with the expert selected and global networks. The randomized networks are generated using two schemes, one is to generate a random network with the same number of nodes as the expert selected one, and the other is to generate a random network with the same number of edges. For both node and edge based randomization, 10,000 networks were generated. The node and edge entropies of these random networks were calculated and averaged.

# 3: RESULTS AND DISCUSSIONS

## 3.1: CHONDROCYTE DATA SET

Node and edge entropies are calculated for several network and sub-networks. The summary of these networks are listed in Table 2. The Human Proteome network represents the global condition and the influences of differentiation in the global scope. Expert Curated sub-network is supposed to be the sub-network with significant changes, since the query nodes are highly related to chondrocyte differentiation. The two types of random sub-networks should be close to the

| Network Name | Description |
|---|---|
| Human Proteome | All of the available human protein interactions |
| Expert Curated | Expert selected sub-network, including the five categories of proteins |
| Cartilage | Sub-network based on proteins in Cartilage category |
| Collagen | Sub-network based on proteins in Collagen category |
| ECM | Sub-network based on proteins in Extra Cellular Matrix category |
| Growth Factor | Sub-network based on proteins in Growth Factor category |
| Signal | Sub-network based on proteins in Signaling Protein category |
| Random (node) | Random network with the same number of nodes as Expert Curated |
| Random (edge) | Random network with the same number of edges as Expert Curated |

**Table 2. Various network and sub-networks used in entropy calculation.**

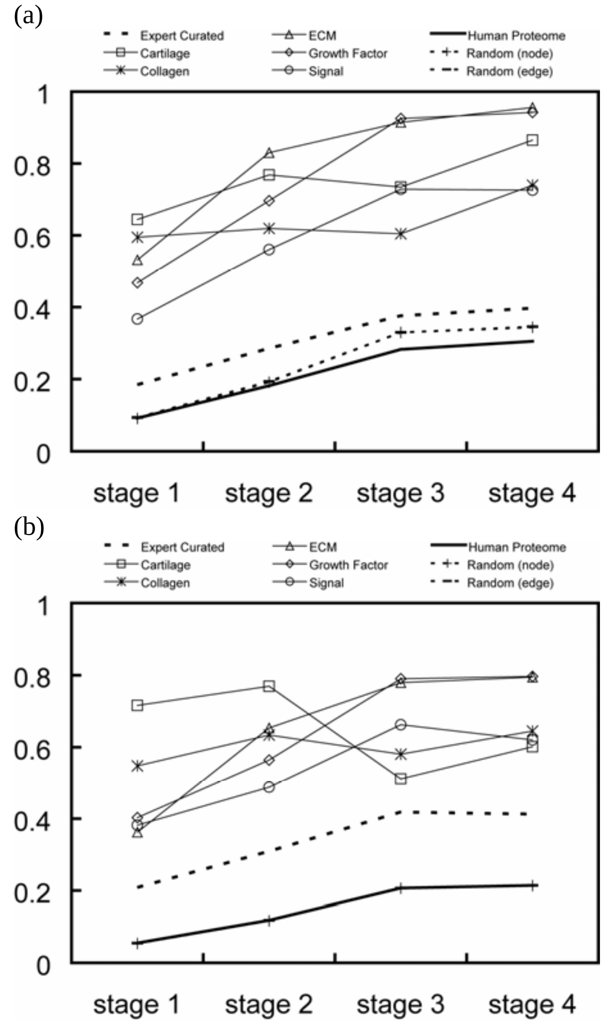

**Figure 1. (a) Node entropy and (b) edge entropy of various networks in different stages.**

global network, since they are sampled randomly from the global network. The node and edge entropies for these network and sub-networks are illustrated in Figure 1. Figure 1a shows the node entropies, whereas Figure 1b shows the edge entropies.

From Figure 1, it is clear that for both node and edge entropies, the entropy values have a trend to increase as chondrocyte differentiation progress. This is not surprising, since differentiation is a process to change toward a condition different from the baseline. Increase in entropy implies that the distribution of states is perturbed and deviate from the original uniform condition. In the baseline time point (stage 0), all of the genes are in unchanged reference state. As stage progress, some genes will become up regulated (expression level higher than those in the baseline) or down regulated (expression level lower than those in the baseline).

There are several differences between node and edge entropies. First of all, the scales of the two entities are slightly different. The maximum value of node entropy is -ln(1/3), where as that of edge entropy is -ln(1/5). The scales shown in Figure 1 are normalized. This enables the comparisons between node and edge entropies. Second, the edge entropy values of the random networks are identical to that of Human Proteome. However, the node entropy values are slightly different for random and global networks. It seems that the distributions of node states are slightly biased. Also, the separation between expert selected sub-network (dashed line) and the global network (solid line) is larger in edge entropy. This suggests that the edge entropy might perform better than node entropy in searching a significant sub-network reflecting maximum changes. Maximizing node entropy should result in a sub-network focused on proteins with modified expression levels, whereas maximizing edge entropy should lead to a network including interactions with changed weights. Edge entropy, in this regard, should reflect more of the functional aspects of the dynamics in biological networks. Most network analyses focus on what nodes (proteins) are critical to the network topology per se [2, 3, 9]; however, we consider the edges (interactions) might provide more important information as what these participators are doing.

Among the five sub-networks from five categories of proteins, the Cartilage related network is most interesting. The node entropy of Cartilage sub-network is increasing as the progress of stages, but the edge entropy is higher at earlier stages and drop later (Figure 1b). Collagen related sub-network also has similar characteristics, but to a lesser extent. Cartilage genes are marker genes in chondrocyte differentiation. Changes in the expression levels of cartilage genes increase the node entropy. It is interesting, however, why the edge entropy drop in later stages. Drop in edge entropy implies the states of interactions involved in this sub-network are moving toward the reference state. That is, functions associated with these cartilage proteins cease to function as the differentiation progress to a certain point. The functional implication of cartilage genes in chondrocyte differentiation is beyond the scope of this paper, and further investigations are required to fully understand the functional roles of cartilage related genes in this particular setting.

## 3.2: YEAST CELL CYCLE DATA SET

Our method has also been applied to the yeast cell cycle dataset. The results are illustrated in Figure 2. In the yeast data set, only two networks are used: the ‘Yeast Interactome’, which contains all available yeast protein-protein interactions, and ‘Cell Cycle Sub-Network’, which contains 800 cell cycle regulated genes identified by Spellman *et al.* [7] Since the cell cycle regulated genes include genes peaked at various phases (M, G1, S, and G2), the periodicity of entropy changes are not clearly observable. However, the distinction between ‘Yeast Interactome’ and ‘Cell Cycle Sub-Network’ is obvious. This result again confirms that our method is able to distinguish sub-networks.

The differences between node and edge entropies, on the other hand, are not obvious. The entropy values are not normalized in the yeast data set, and the scales

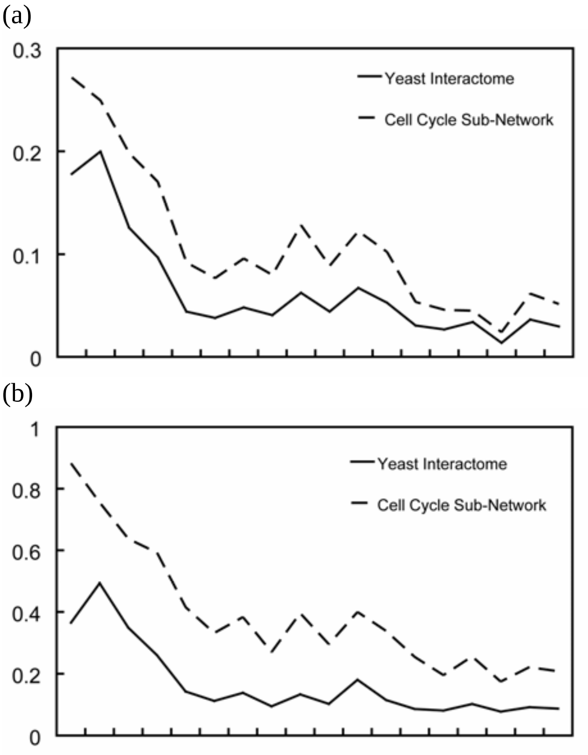

**Figure 2. (a) Node entropy and (b) edge entropy of yeast networks in different time points.**

are not directly comparable for node and edge entropies. However, the overall profiles of node and edge entropies are similar, suggest the two may perform equally well in the yeast data set.

One thing of interest is that the node and edge entropies decreasing along the 18 time points. The yeast cells have been synchronized in this time course experiment, the expression levels are compared to those from the asynchronized culture. Smaller entropy values imply convergence between the two states (synchronized and asynchronized). Decreasing entropy values may suggest that the cell culture may experience 'loss of synchronicity' problem. On the other hand, the synchronization methods may introduce perturbations to the cell, and causes the increase of entropy. As time progresses, the system may stabilize to a steady state. It is amazing that edge and node entropies can catch such subtle changes. Further investigations are required to conclude the actual cause of this phenomenon.

## 4: CONLUSIONS

We have developed novel measures to quantify the dynamics and changes of biological networks under different conditions. Our scores are easy to calculate and applicable to various scenarios. With our scores, new insights might be revealed from older experimental microarray data available in public database like GEO.

We also found that in analyzing biological networks, edges (interactions) might play more important roles than nodes (proteins or genes). More investigation are required to validate this assumption, but we believe this idea will shed new lights to network biology and systems biology.

Node and edge entropy values may also serve as a score to identify sub-networks with biological significances. We are working on a genetic algorithm based method for sub-network identification. Hopefully this method will be helpful to the field of bioinformatics in general.

## REFERENCES

[1] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics,* vol. 18 Suppl 1, pp. S233-40, 2002.

[2] E. Estrada, "Virtual identification of essential proteins within the protein interaction network of yeast," *Proteomics,* vol. 6, pp. 35-40, Jan 2006.

[3] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature,* vol. 411, pp. 41-2, May 3 2001.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Tech. J.,* vol. 27, pp. 379-423, 623-656, 1948.

[5] C. G. James, C. T. Appleton, V. Ulici, T. M. Underhill, and F. Beier, "Microarray analyses of gene expression during chondrocyte differentiation identifies novel regulators of hypertrophy," *Mol Biol Cell,* vol. 16, pp. 5316-33, Nov 2005.

[6] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* vol. 34, pp. D173-80, Jan 1 2006.

[7] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell,* vol. 9, pp. 3273-97, Dec 1998.

[8] T. W. Huang, A. C. Tien, W. S. Huang, Y. C. Lee, C. L. Peng, H. H. Tseng, C. Y. Kao, and C. Y. Huang, "POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome," *Bioinformatics,* vol. 20, pp. 3273-6, Nov 22 2004.

[9] T. Manke, L. Demetrius, and M. Vingron, "Lethality and entropy of protein interaction networks," *Genome Inform Ser Workshop Genome Inform,* vol. 16, pp. 159-63, 2005.