Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Association between pathways in regulatory networks

Yuval Kluger, Harriet Kluger, and David Tuck

Abstract— **During cell progression from one state to another, such as transformation from benign to malignant conditions, cells undergo changes in gene regulation. To reveal state-dependent circuitries in human regulatory networks, we employed drafts of normal and malignant cell networks. Using these condition specific networks, gene profiles and annotated pathways we studied: a) the capacity to separate samples or cell states based on the collective expression of all the genes in each pathway rather than individual genes, b) the degree of regulatory network connectivity within and between pathways. Distinct cell types reveal notable differences in transcriptional activity in numerous pathways. On the other hand, in datasets from breast cancer patients with variable outcome the capacity of single pathway expression signatures to predict disease outcome is very limited, though this can be somewhat improved by combining multiple pathways. Remarkable connectivity between pathways on the transcriptional regulatory level revealed a non-modular network structure. Overall, network blueprints enable us to quantify the degree of interaction between condition specific co-regulated pathways. This can contribute to understanding deregulated processes associated with cancer.**

I. Introduction

The study of mammalian transcriptional regulatory networks based on high throughput gene expression data has primarily focused on the identification of individual differentially expressed genes, co-regulated gene sets and genes with inferred functional similarity. Investigators have identified functional modules from gene expression data using a reverse-engineering approach to reveal regulatory subunits, based on probabilistic graphical models [1], singular value decomposition [2-4] and network component analysis [5]. Methodologies for reconstructing and inferring elements of genetic and metabolic networks [1, 6-11], identifying new gene modules [12-14] and statistically characterizing topological network features are the focus of much research [15-19], especially for model organisms. Recent advances in network analysis have focused on topological changes and static and dynamical network properties in yeast and *E coli* [20-22].

Manuscript received April 24, 2006. This work was supported in part by the Susan G.Komen Breast Cancer Foundation.

Yuval Kluger is with the Department of Cell Biology, Skirball Institute, New York University School of Medicine, New York, NY 10016 USA (corresponding author to provide phone: 212-263-8289; fax: 212-263-5555; e-mail: kluger@ saturn.med.nyu.edu).

Harriet M. Kluger is with the Department of Internal Medicine,Yale University, New Haven, CT 06520 U.S.A. Her effort is supported by the Susan G.Komen Breast Cancer Foundation and the Breast Cancer Alliance of Greenwich, CT.

David. P. Tuck., is with the Department of Pathology, Yale University, New Haven, CT 06520 U.S.A.

In previous work [23] we constructed initial drafts of human transcriptional regulation networks of normal and cancer cells in order to study:
a) the dispersion or localization in the network either of gene sets that yield near optimal classification; if these genes are concentrated in focal regions rather than spread over the entire network, this may suggest strategies for intervention not apparent from study outside of the network context.
b) the use of network features to identify sub-networks that exhibit differential regulatory activity. We examined whether the combined expression levels of central genes are sufficient to characterize the cell state. Additionally, we introduced an approach to utilizing network structure to classify samples, in terms of transcription factor-target gene activities rather than gene expression profiles, which allows the identification of key differential activity of transcription factor-gene pairs. In this work we first explore the extent to which the composite expression signatures of gene pathways distinguish between different cell types or cancer cells from patients with different clinical outcome as an extension to work done by Stolovitzky et al [24] and us [25]. We then examine how gene sets defined by pathways spread or cluster in terms of their geodesic distances in the regulatory network. Finally, we define condition specific transcriptional interaction between each pair of pathways by identifying transcription factors (TFs) that activate or suppress members in both pathways in this specific state.

II. Materials and Methods

A. Microarray datasets

1. Distinct normal cell types: Affymetrix U133 chip mRNA expression data of 10 resting neutrophil and 19 resting monocyte samples obtained from normal individuals [25].
2. Distinct tumor types: comparing acute lymphoblastic leukemia (ALL) to acute myeloid leukemia (AML) [26].
3. Tumors versus normal tissues: Affymetrix U133A and U133B chips of 9 renal cell carcinoma samples and matched normal samples [27].
4. Tumors with variable clinical outcomes: Hu25K oligonucleotide and U95Av2 Affymetrix datasets from studies of predictors of clinical outcome in breast cancer [28, 29].
5. Ectopic production of oncogenes: Affymetrix Mu11K arrays of 55 quiescent mouse embryo cell sample and cells over-expressing Myc, Ras, E2F1, E2F2 or E2F3 **[30].**

B. Simultaneous array and gene normalization

We applied a bi-normalization procedure which has been extensively described in previous work, and has led to clean separation between distinct cell types [25, 31].

C. Connectivity network

The human and mouse connectivity networks are derived from a combination of high quality literature references (documented in the professional version of the TRANSFAC Database) and predictions based on matching known and putative transcription factors consensus binding sites sequences with the 1kb (or 5kb) upstream promoter regions of all human and mouse genes stored in Goldenpath (http://genome.ucsc.edu/). We used the default parameters of the MATCH algorithm (provided by TRANSFAC) and a minimal score of 0.85 as a threshold to define direct regulation connectivity between a transcription factor (TF) and a gene. The information is stored in a rectangular adjacency matrix, in which regulating TFs are represented by column indices and regulated genes by row indices. The elements of the matrix C_{ij} are assigned a value of one if transcription factor j directly regulates gene i. If not, the elements are assigned a value of zero. We arranged the rows of this matrix such that the regulated genes in the first rows are TF genes. Furthermore, the order of the TFs across these rows is equivalent to their order across the columns. Thus, the upper square block of the matrix C consists of the TF regulatory network, and its non-zero diagonal terms C_{ii} correspond to self-regulation.

D. Co-expression networks

Some of the "noise" in the connectivity regulatory network is due to false positive predictions of the putative TF-gene relationships. To reduce the noise we constructed two types of co-expression networks for each dataset and intersected each of them with the connectivity network. The correlation across conditions co-expression network was constructed by computing the correlation across all samples between each TF-gene pair in the connectivity network. We then transformed this matrix to a binary matrix e_{ij} by substituting one for any entry with an absolute correlation coefficient larger than 0.3 and zero otherwise. The condition specific co-expression network was constructed as follows: for each condition (column) of the bi-normalized matrix we computed the outer product between the absolute value of this vector multiplied by the sub-vector corresponding to the transcription factors. This outer-product gives rise to a matrix denoted by E_{ij} that has the same dimensions as the connectivity matrix C_{ij}. Each element of E_{ij} is the product of the absolute value of the bi-normalized gene expression *i* with the bi-normalized TF expression *j*. The elements of E_{ij} are then binarized using 1 if they are larger by a standard deviation from the mean product and 0 otherwise. A nonzero element in this binary matrix indicates that the expression level of the TF is up-regulated and the absolute value expression level of the gene (potentially activated or suppressed by this TF) is up-regulated or down-regulated. For each sample in each class we calculated these binary co-expression matrices (which can be individual co-expression ma. We then added up these binary matrices for each specific condition and replaced each entry in the resultant matrix by one if it is nonzero in at least 1/3 of the binary co-expression matrices of this condition and by zero otherwise. Finally, we used the intersection between the condition specific co-expression matrix and the connectivity matrix to form the condition specific (CS) regulatory network represented by the adjacency matrix $A_{ij}= C_{ij} E_{ij}$. Similarly, the intersection between the correlation across conditions co-expression matrix and the connectivity matrix, represented by the CAC adjacency matrix $a_{ij}= C_{ij} e_{ij,}$ gives rise to the links of the regulatory network that vary in coordinated fashion across more than one biological condition.

E. Geodesic gene-gene distance

The proximity of groups of genes (corresponding to a pathway, a near optimal classification multivariate or the central genes in the regulation network (hubs)) is determined by the particular distance measures we used. The most straightforward distance measure we used on the directed graph representing the regulatory network is the geodesic distance. The geodesic distance between a gene-TF pair ($gene_i$-TF_j) is 1 if the corresponding entry in the adjacency matrix is one (A_{ij}=1). A pair of $gene_i$-TF_j whose A_{ij}=0 could be indirectly connected in the regulatory network via other transcription factors regulated by TF_j. To find these indirect connections we reorganized the matrix A_{ij} such that its upper square block, defined by $T_{ij,}$ consists of transcription factor pairs (TF_i-TF_j) only. If an entry of the adjacency matrix A_{ij} is zero but the same entry of $(AT)_{ij}$ is nonzero the geodesic distance between $gene_i$ and TF_j is 2. Similarly, if the entries of $(AT^m)_{ij}$ for all m=0,..,n-1 is zero and the corresponding entry $(AT^n)_{ij}$ is nonzero the $gene_i$-TF_j geodesic distance is n+1. Thus, the geodesic distance between a given transcription factor and a gene is the shortest directed path between them, i.e., the smallest number of arcs connecting them. A pair of non-regulating genes has an infinite geodesic distance, because there is no directed path in the regulating network, which connects these genes. To define a distance between any pair of genes, whether any of these genes is a transcription factor or not, we identify an "ancestor" transcription factor in the regulatory network, whose sum of geodesic distances to both genes is minimal. If one or two of the genes of a given pair is a transcription factor, the gene-gene distance of this pair is defined by the either the shortest directed path between them or the gene-gene definition above, whichever is smaller.

F. Annotated Pathway Data

The Biocarta database was originally extracted from the caBIG repository (http://cgap.nci.nih.gov/Pathways) and data from the KEGG database (http://www.genome.ad.jp/kegg) extracted from our internal database management system, GeneCube. Linking of mouse–human pathways were provided by linking data from HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene).

G. Combining Transcription Factor-Target Gene Binding Network and Gene Co-expression Networks

We constructed networks by intersecting a predicted and literature-based connectivity binding network (using TRANSFAC [32, 33] and in specific examples additional array based protein-DNA binding location analysis (ChIP-on-chip)) with two types of co-expression networks:

condition specific (CS) network and correlation across conditions (CAC) networks. These co-expression networks were drawn from three types of microarray datasets, using publicly available data, including those derived from normal cell lineages [25], tumors versus normal tissues [27], and disease-specific tumors associated with variable clinical outcomes [28, 29]. The links in the CAC networks represent TF-gene pairs whose expression profiles are correlated across all states considered, i.e., if a TF is over-expressed its target genes are also over-expressed and vice versa, if inversely correlated. CS networks represent TF-gene pairs that are co-expressed in one state or more. Transmission of information or transcriptional control within these networks is state-dependent [20, 21]. This is partially reflected by changes in the stimulatory or inhibitory activity of specific links in the regulatory networks, which can lead to altered distances between genes.

B. Scattering of pathway genes in the regulatory network

We examined distributions of distances (the number of links needed to traverse the path between each pair of genes) in the CS and CAC networks for gene subsets that belong to the same pathway. To explore these distance maps in a simple mammalian system caused by single oncogene perturbations, we used available datasets for ectopic over-expression of RAS, Myc, E2F1, E2F2 or E2F3 in mouse embryo fibroblasts [30, 34]. We examined all pathways in which these genes were known to participate. We found that the collective expression patterns of each E2F-related pathway clearly separated the samples. Similar, though less pronounced, results were obtained with Myc and Ras pathways. In addition, genes separating parental from Ras perturbed cell lines identified by supervised machine learning approach are located in close proximity on the CAC or CS networks (Fig. 1).

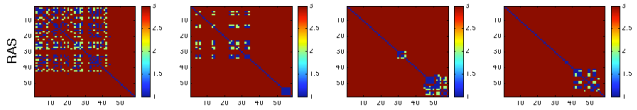

Fig. 1. Distance map of the sub-regulatory network containing genes separating normal fibroblasts from fibroblasts perturbed with ectopic RAS expression. The geodesic distances between a pair of genes in the sub-network containing these separating genes and all the TFs connected to them directly or indirectly are displayed in the left distance map (panel 1). The distances are color coded in the bars to the right of these maps (values greater than 6 represent disconnection, and in rare cases, long geodesic distances). These separating genes are much closer to each other than random sets of genes as shown in panels 2-4. Thus, genes that separate between normal and RAS perturbed fibroblasts tend to localize in the CAC regulatory network when compared with a random, same-size gene sets.

C. Pathway based classification

Next, we assessed whether the collective gene expression patterns of individual pathways could be used to classify samples, and studied the proximity of genes in these pathways in the regulatory network. The KEGG and Biocarta databases present catalogs of groups of genes, classified according to their linkage in known pathways including those for small molecules. We utilized each of the 137 pathways described in KEGG and the 245 networks listed in Biocarta. We employed a partial least squares (PLS) analysis to compare distinct cell types such as neutrophils to monocytes [25] and normal to malignant kidney cells [27]. The combined expression profile for many pathways reveals perfect separation. A small number of pathways (including complement and coagulation cascades, MAPK and integrin pathways) classified AML from ALL with a cross-validated error rate of 8%-12%. We tested the effect of combining the top 5 separating pathways and observed a reduction of the error rate to 4%. For datasets that are more difficult to classify, such as breast cancer cohorts with variable clinical outcome [28, 29], linear discriminant analysis, PLS and k-nearest neighbor classifiers of the combined expression patterns for pathways in these datasets, led to, at best, classification errors of 21 % for the Duke breast data and 40% for the Dutch breast data.

Here we analyzed data with variable degrees of difficulty in classification, the extreme being breast cancer patients with different clinical outcomes. For these data there is a corresponding difficulty in identifying distinguishing pathways. However, specification of the pathways activated in different cell types or different tumors is useful for understanding tumor biology, as shown by Segal et al [1], who characterized entirely different cancers by aggregates of pathways. The combined effect of these "differentially expressed" pathways is a statistical observation.

D. Transcriptional interaction between pathways

To understand transcriptional communication between pathways and identify their common regulators, we explored inter- and intra-pathway regulatory connections. We first prepared a pathway-by-gene matrix that indicates which genes belong to each pathway and multiplied it by the condition-specific regulatory network (TF-target gene matrix) to construct a TF-pathway matrix (Fig. 2a). Each element in this TF-pathway matrix counts the number of links between the corresponding TF and pathway (Fig 2b). We defined an inter-pathway interaction matrix (Fig. 2c), taking into account coordinated expression in a specific condition as well as the connectivity, whose entries represent estimates of interaction strength between any pair of pathways: the higher the score, the stronger the shared regulatory relationships between the two pathways (Fig. 2d). Examples of pathway pairs with strong interactions (in red, Fig. 3) in the Dutch breast dataset were the p53 and Rb pathways. This is primarily mediated by the fact that E2F1 and EGR1 regulate different genes in both pathways. For these data we show (histogram, Fig. 4) that a substantial proportion of pathway pairs share common network regulators. We also studied intra-pathway connectivity. Not surprisingly, we found that the effective gene–gene regulatory distance between members of numerous individual pathways is smaller than the effective distance of random equal-sized groups of genes for both CS and CAC in all categories of microarray experiments. For instance, the shortest gene-gene network distances for 83 out of the 369 pathways of the neutrophil/monocyte dataset are found significantly more frequently than in random gene sets.

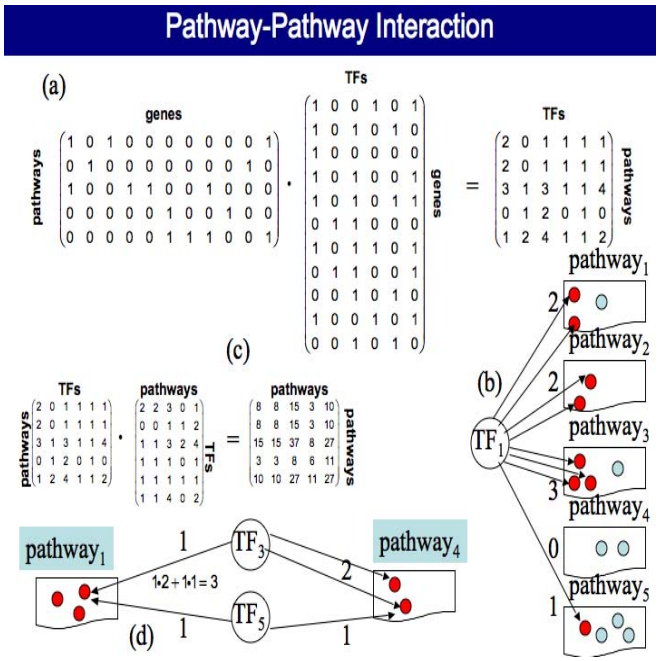

Fig. 2. Pathway-pathway interaction map. We define the strength of a regulatory pathway-pathway interaction by the number of links between a pair of pathways via common direct regulators of the pair. To derive the pathway-pathway interaction map on the right hand side of equation (c), we first construct the TF-pathway matrix (equation (a), right side) obtained from multiplying the pathway-gene matrix by the adjacency gene -TF matrix. Each row in the pathway-TF matrix represents a different pathway and each column represents a TF. The first column of this matrix represents the number of direct targets of the first TF to each of the pathways, as illustrated in (b). In the second step we multiply the TF-pathway matrix by its transposed matrix to obtain the pathway-pathway interaction map (c). For example the interaction strength between pathway 1 (1st row or column) and pathway 4 (4th column or row) is equal to three, as illustrated in (d). The entries in the pathway-pathway map are the number of trajectories via their common regulators. To account for pathway size we normalize each of these scores by dividing them by the corresponding product of the number of genes in each pair of pathways. Because some pathways share common genes, the scores for these pathways excluded the contribution from these shared genes.

Analysis of genome-wide expression data in the context of networks is not merely descriptive, but can offer insight into biological processes. This view allows us to inspect how sub-networks associated with various biological pathways differ between normal and cancer cells or between good and poor prognosis patients. For instance, many of the genes implicated in recurrence in breast cancer in the Dutch study [29] are regulated directly by E2F1, a transcription factor that plays a well established role in cell cycle progression [35], or by other TFs that are targets of E2F1.

We found that distinct cell types can be easily separated by collective expression profiles of pathways or by genes identified by supervised learning. We conclude that differences between these distinct cell types are so pervasive that it is hard to implicate a small number of characteristic biological processes. Mathematically, this can be explained by the fact that the rank of the data matrix is very low – on the order of magnitude of the number of cell types we are trying to partition. We note that in metabolic networks, flux analysis in E. coli reveals that changing conditions led to altered flux only in a few metabolic pathways [9]. In the mammalian cell datasets we studied, many pathways have distinct levels of activity between the cell types or states. Whether this represents a general effect in multi-cellular organisms or whether it is particular to the type of perturbation or change in state warrants further examination.

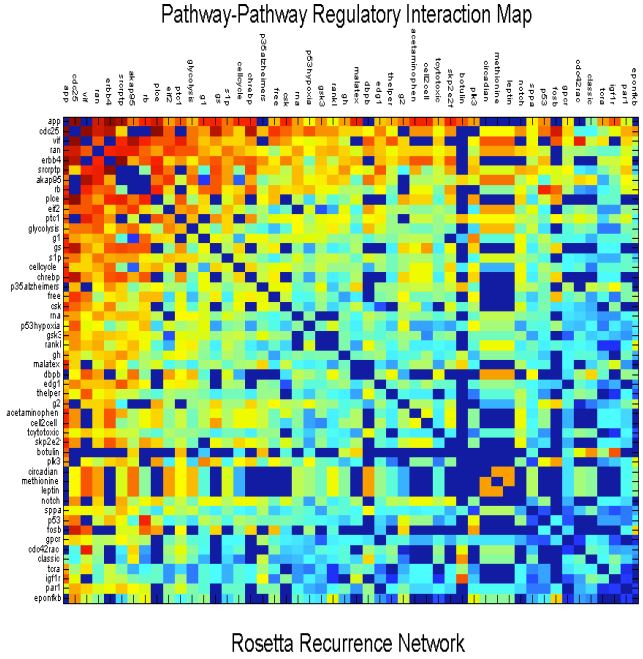

Fig. 3. Pathway-pathway interaction map for the Dutch breast cancer data. Red spots represent pathways with high regulation commonality, and blue represents low regulation commonality. For example: the Erbb4 and Rb pathways share common regulators in the CS recurrence network, as do the Erbb4 and cell cycle pathways, and the map indicates that their pathway-pathway interactions scores are relatively high.

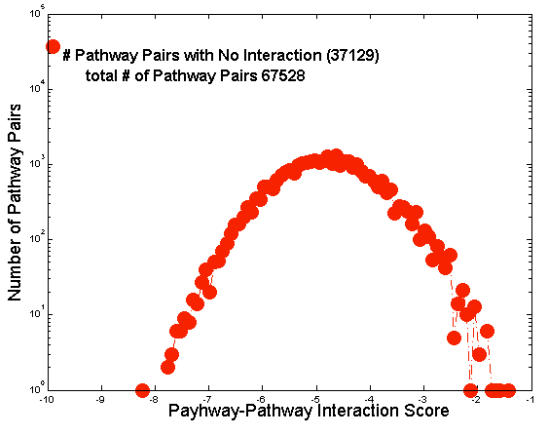

Fig. 4. Pathway-pathway log interaction scores in the Dutch recurrence network reveal strongly connected networks of pathways. Many pairs of pathways share common regulators.

III. Conclusion

In this work we showed that distinct cell types reveal notable differences in transcriptional activity in numerous pathways. On the other hand, correlation (or classification) of disease outcome with single pathway expression signatures, which in principle allows biological interpretation, is limited. The remarkable connectivity between pathways on the transcriptional regulatory level can have important practical implications for further research on cell modularity and interventions for cancer control.

Despite the limitations of the currently available data, these initial networks already enable us to find key regulators of deregulation within the regulatory network.

REFERENCES

[1] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet*, vol. 34, pp. 166-76, 2003.

[2] M. K. Yeung, J. Tegner, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proc Natl Acad Sci U S A*, vol. 99, pp. 6163-8, 2002.

[3] J. Tegner, M. K. Yeung, J. Hasty, and J. J. Collins, "Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling," *Proc Natl Acad Sci U S A*, vol. 100, pp. 5944-9, 2003.

[4] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, pp. 102-5, 2003.

[5] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury, "Network component analysis: reconstruction of regulatory signals in biological systems," *Proc Natl Acad Sci U S A*, vol. 100, pp. 15522-7, 2003.

[6] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, pp. 799-805, 2004.

[7] J. S. Edwards and B. O. Palsson, "Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions," *BMC Bioinformatics*, vol. 1, pp. 1, 2000.

[8] C. H. Schilling, J. S. Edwards, D. Letscher, and B. O. Palsson, "Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems," *Biotechnol Bioeng*, vol. 71, pp. 286-306, 2000.

[9] E. Almaas, B. Kovacs, T. Vicsek, Z. N. Oltvai, and A. L. Barabasi, "Global organization of metabolic fluxes in the bacterium Escherichia coli," *Nature*, vol. 427, pp. 839-43, 2004.

[10] J. J. Rice, Y. Tu, and G. Stolovitzky, "Reconstructing biological networks using conditional correlation analysis," *Bioinformatics*, vol. 21, pp. 765-73, 2005.

[11] P. M. Magwene and J. Kim, "Estimating genomic coexpression networks using first-order conditional independence," *Genome Biol*, vol. 5, pp. R100, 2004.

[12] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford, "Computational discovery of gene modules and regulatory networks," *Nat Biotechnol*, vol. 21, pp. 1337-42, 2003.

[13] J. Ihmels, S. Bergmann, and N. Barkai, "Defining transcription modules using large-scale gene expression data," *Bioinformatics*, vol. 20, pp. 1993-2003, 2004.

[14] D. Steinhauser, B. H. Junker, A. Luedemann, J. Selbig, and J. Kopka, "Hypothesis-driven approach to predict transcriptional units from gene expression data," *Bioinformatics*, vol. 20, pp. 1928-39, 2004.

[15] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, pp. 1538-42, 2004.

[16] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101-13, 2004.

[17] M. J. Herrgard, M. W. Covert, and B. O. Palsson, "Reconciling gene expression data with known genome-scale regulatory network structures," *Genome Res*, vol. 13, pp. 2423-34, 2003.

[18] S. Maslov and K. Sneppen, "Detection of topological patterns in protein networks," *Genet Eng (N Y)*, vol. 26, pp. 33-47, 2004.

[19] A. H. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A. M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone, "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, pp. 808-13, 2004.

[20] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, pp. 308-12, 2004.

[21] J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, pp. 88-93, 2004.

[22] H. Kim, W. Hu, and Y. Kluger, "Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae," *BMC Bioinformatics*, vol. 7, pp. 165, 2006.

[23] D. Tuck, H. Kluger, and Y. Kluger, "Characterizing disease states from topological properties of transcriptional regulatory networks," *BMC Bioinformatics (in press)*.

[24] N. Shah, J. Lepre, Y. Tu , and G. Stolovitzky, "Can we identify cellular pathways implicated in cancer using gene expression data?," presented at The 2nd IEEE Computational Systems Bioinformatics, 2003.

[25] Y. Kluger, D. P. Tuck, J. T. Chang, Y. Nakayama, R. Poddar, N. Kohya, Z. Lian, A. Ben Nasr, H. R. Halaban, D. S. Krause, X. Zhang, P. E. Newburger, and S. M. Weissman, "Lineage specificity of gene expression patterns," *Proc Natl Acad Sci U S A*, vol. 101, pp. 6508-13, 2004.

[26] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, 1999.

[27] M. E. Lenburg, L. S. Liou, N. P. Gerry, G. M. Frampton, H. T. Cohen, and M. F. Christman, "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data," *BMC Cancer*, vol. 3, pp. 31, 2003.

[28] E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang, "Gene expression predictors of breast cancer outcomes," *Lancet*, vol. 361, pp. 1590-6, 2003.

[29] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-6, 2002.

[30] E. Huang, S. Ishida, J. Pittman, H. Dressman, A. Bild, M. Kloos, M. D'Amico, R. G. Pestell, M. West, and J. R. Nevins, "Gene expression phenotypic models that predict the activity of oncogenic pathways," *Nat Genet*, vol. 34, pp. 226-30, 2003.

[31] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome Res*, vol. 13, pp. 703-16, 2003.

[32] E. Wingender, "TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks," *In Silico Biol*, vol. 4, pp. 55-61, 2004.

[33] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374-8, 2003.

[34] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, Jr., J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, pp. 353-7, 2006.

[35] H. Cam, E. Balciunaite, A. Blais, A. Spektor, R. C. Scarpulla, R. Young, Y. Kluger, and B. D. Dynlacht, "A common set of gene regulatory networks links metabolism and growth inhibition," *Mol Cell*, vol. 16, pp. 399-411, 2004.