Detecting Cancer Pathway Crosstalk with Distance Correlation

Michael F. Sharpnack, Kun Huang, PhD
The Ohio State University, Columbus, OH

Abstract

Biological pathway regulation is complex, yet it underlies the functional coordination in a cell. Cancer is a disease that is characterized by unregulated growth, driven by underlying pathway deregulation. This pathway deregulation is both within pathways and between pathways. Here, we propose a method to detect inter-pathway coordination using distance correlation. Utilizing data generated from microarray experiments, we separate the genes into pathways and calculate the pairwise distance correlation between them. The result is intuitively viewed as a network of differentially dependent pathways. We find intuitive, yet surprising significant hub pathways, including glycophosphatidylinositol anchor synthesis in lung cancer.

Background

Biological pathways in the human cell function together in a highly orchestrated fashion. This coordination results from several mechanisms, including the common occurrence of two metabolic pathways sharing a common substrate. Timing is crucial to the development of the cell and deregulation of the interactions among pathways can have disastrous consequences, such as tumorigenesis. We present a method to detect interactions among pathways from gene expression data of multiple samples, and apply it to identify changes of interactions between pathways in cancers versus normal tissues. We hypothesize that phenotypic changes between two conditions, such as tumor and normal, are associated with changes in pathway dependencies, and further that hub pathways are of special importance to these phenotypic changes. This hypothesis has an advantage that it focuses on the collective behavior of genes in pathways instead of individual genes and therefore does not require correlation between expression profiles of individual genes.

In order to mathematically characterize the functional relationship between two gene lists given their expression profiles over a collection of samples, we implement a relatively new similarity metric called distance correlation[1]. Distance correlation is a type of correlation metric which can detect nonlinear relationships between two vectors or matrices. Given two matrices with the same number of columns, for each matrix, we can consider their columns to be feature vectors for a set of samples. Therefore the distance correlation first calculates the distances between the samples. Then the Pearson correlation coefficient (after a normalization process) between the two sets of distances is computed as the relationship measurement between the two matrices. Geometrically, this is equivalent to compare two weighted networks for the samples and thus exactly matches the notion of our hypothesis.

To test our hypothesis, we develop a two stage workflow. The first stage is to establish a pathway network for samples in different conditions such as cancer versus normal tissues using whole genome transcriptome data from microarray experiments. The second stage is to identify interacting pathways and pathway clusters in specific conditions such as cancers. Our results in multiple cancer studies show that we are able to identify specific pathway interaction in cancers, which supports the notion on altered metabolism processes in cancers. These results suggest that our approach will lead to wide applications as a translational bioinformatics tool for studying diseases at the pathway levels.

Pathway regulation is complex and multifactorial. As such, pathways exhibit both linear and nonlinear dependence on each other. Further complicating the situation, different genes in a pathway have varying levels of importance to the overall function of that pathway. It is not clear what constitutes an active pathway. Some methods have used the average gene expression or a threshold for the number of genes needed to be active to say that the entire pathway is active. One pitfall of these assumptions is in pathways with a highly influential rate-limiting reaction that is controlled by a single enzyme, such as in cholesterol synthesis. Cholesterol synthesis begins with Acetyl-CoA and ends at Cholesterol after six reactions; however, the rate-limiting reaction, the reaction that controls the kinetics of cholesterol synthesis, is the conversion of HMG-CoA to Mevalonate by the enzyme HMG-CoA reductase. This reaction is inhibited by HMG-CoA reductase inhibitors.

Much attention has been given to deregulation of genes within pathways, as this intra-pathway deregulation is the hallmark of many cancers. The question of inter-pathway regulation has only recently been posed, perhaps due to the relatively short time that high throughput expression analysis has been available. The effects of one pathway on another could potentially be as important as the effect of one gene in a pathway on another gene in the same pathway. Many genes exert pleiotropic effects on distinct pathways, and transcriptional regulation can be location-specific, rather than function-specific. In other words, single transcription factors can regulate the genes that belong to distinct functional pathways. Because of this, selecting a single pathway to study is likely an incomplete picture of the causes of tumorigenesis. Pathway coordination, or ``crosstalk", has been studied by calculating differential expression of genes or sets of genes within pathways[2,3,4,5]. These methods also frequently incorporate, and therefore rely on protein-protein interaction networks. The nonlinear nature of our pathway also differentiates it from PAthway Network Analysis (PANA), a method proposed by Ponzoni et al., which can only detect linear patterns[4]. PANA also employs dimensionality reduction methods, which result in a loss of information that our method does not suffer. Cho, et al. use a set-wise interaction score that employs the Renyi relative entropy measure to measure pathway crosstalk; however, methods employing information theory techniques are unlikely to be intuitive to biologists[6]. All of these methods address the question of differentially expressed pathways, not differentially correlated pathways. This subtle difference separates two biological questions. We seek to answer the question, how does the dependency between pathways differ between cancer and normal cells? It is a much more general question than asking whether or not the pathways are over or underexpressed together.

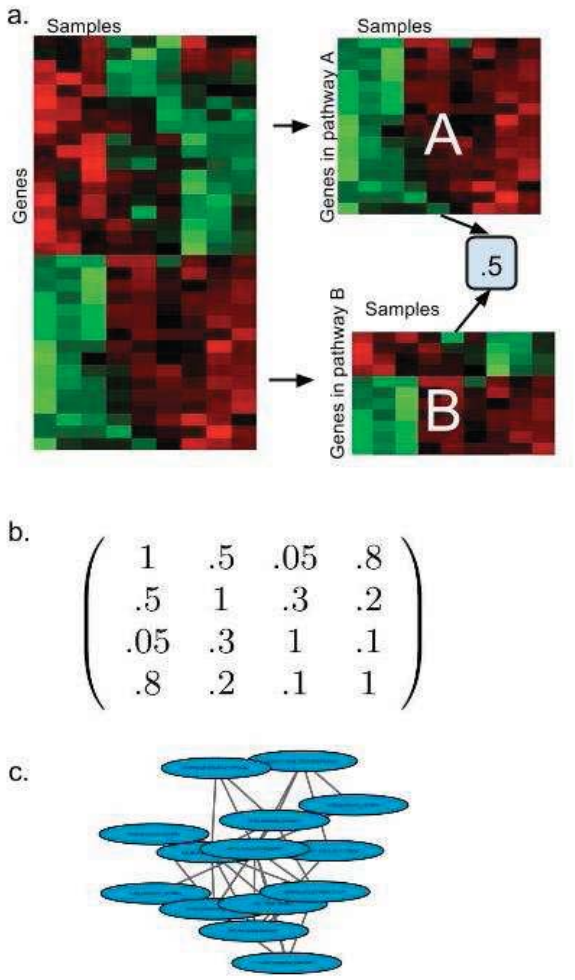

Figure 1 The workflow for establishing pathway dependency networks and compare between different conditions. **(a)** Matrix of expression values for each pathway (A and B) are extracted from the full expression matrix. (**b)** A single distance correlation values represents the dependency on two pathways, such as A and B. These distance correlation values are combined into the symmetric distance correlation matrix shown in **(b)**. **(c)** Two matrices such as that shown in **(b)**, created from two different phenotypes, can be subtracted and their entries used as edge weights to create a differential network. Here we show a representative example of such a network.

Methods

We selected a non-small cell lung cancer paired tumor:normal microarray dataset[7] for study. This dataset was normalized following standard Affymetrix RMA normalization and log transformed. To assign genes to pathways, we adopted the Kyoto Encyclopedia of Genes and Genomes (KEGG)[8] pathways to find the intersection of our microarray genes and genes known to be involved in pathways. In total, we identified 186 diverse pathways from KEGG ranging from cancer-related pathways to signaling and metabolic pathways. Many genes are present in more than one of the 186 pathways, which could prevent a bias in pathway dependence. To minimize the bias between pathways with many genes in common, we simply combine the gene lists of pathways with similar functions and high gene overlap. For instance, 11 pathways related to autoimmunity exhibited high gene overlap, so they were

condensed into a single pathway which is the union of all genes contained in each of 11 pathways. These condensed pathways were relabeled to reflect the overall functional theme. Once the 186 original pathways were condensed, we were left with 116 pathways with low gene overlap. We were able to reduce the number of pathway pairs with >20% genes present in both pathways from 220 to 30. While the remaining 30 pathway pairs had high gene overlap, they were not clearly functionally related, which may be a result of the pleiotropic nature of many proteins. As a further safegaurd against the bias created by pathway overlap, our method considers the differences in dependence, which may be minimal in pathway pairs in which the dependence is high under any experimental conditions due to gene overlap.

To compare the expression of genes within pathways, we employ distance correlation, a measure that can summarize this interaction into a single value. Distance correlation, R, is a measure of the dependence between two random variables, $(X,Y) = \{(X_k,Y_k) : k = 1, \ldots ,n\}$. Conveniently, $0 \leq R \leq 1$, and $R = 0$ if and only if X and Y are independent. In this paper, we consider X_k in $\mathbb{R}^{p,k}$ and Y_k in $\mathbb{R}^{q,k}$ to be microarray expression values for patient k where p and q are the number of genes measured for a given pathways X and Y, respectively. Our expression measurements are therefore organized into 116 p_i x k matrices, where $i = \{1, \ldots 116\}$ and p_i is the number of genes in each pathway. A single distance correlation value is calculated for each pair of pathways, which creates a 116x116 matrix with values between 0 and 1. For our purposes, distance correlation has several advantages when compared to Pearson Correlation: it can detect nonlinear relationships; it can be used to compare two matrices with the same patient sample size but different gene sample size; and it is between 0 and 1. In addition, when a dataset exhibits a bivariate normal distribution, R is a linear function of Pearson Correlation, ρ, with a slope of approximately .9 and $R = |\rho|$ when $|\rho| = 1$. Distance correlation is empirically calculated as described in Szekely et. al[1].

After calculating distance correlation, we arrive at a 116x116 weighted adjacency matrix, where each entry in this matrix is the pairwise distance correlation between two pathways. We have two such matrices, one for tumor samples and one for normal samples. To compare tumor and normal samples, we subtract the normal adjacency matrix from the tumor. We now have the pieces we need to construct a graph representing the change in dependency between each pair of pathways.

To visualize these distance correlation networks, we employed the Clustergram function in Matlab and the widely used network visualization tool, Cytoscape[9]. The Matlab clustergram function uses average linkage for the hierarchical clustering. We then calculated the difference of distance correlation values for every pathway pair between cancer and normal samples for each dataset. The networks were imported into Cytoscape as tables, containing the top .5% of pairwise differential dependencies (based on absolute values of the differences in distance correlation), roughly corresponding to an absolute value of $\geq .15$.

Results

Our method found that the coordination between GPI-anchor biosynthesis and several other pathways, including metabolic pathways, was significantly lower in lung tumor samples than in normal samples as shown **Figure 3** with the clustering of the pathways based on the difference of distance correlation values in **Figure 3.a** and network diagram in **Figure 3.b**. The GPI anchor synthesis, thyroid cancer and circadian rhythm pathways are the three leftmost columns and bottommost rows of **Figure 3.a**, while they are highlighted as yellow nodes in **Figure 3.b**.

Cancer cells are characterized by changes to surface marker proteins, such as glycophosphatidlinositol (GPI)-anchored membrane-bound proteins. For example, carcinoembryonic antigen, a GPI-anchored protein that is usually only expressed in the developing fetus, has been used as a biomarker for colorectal adenoma progression and recurrence[10]. In addition, GPI biosynthetic enzymes have been shown to be elevated in cancer cells [11,12]. We also found that the thyroid cancer pathway was relatively out-of-sync in the cancer samples. Lung and thyroid cancers are diverse, although they may share common pathways. For instance, thyroid transcription factor 1 is active in both lung and thyroid cancers, and its detection is a principal way in which lung adenocarcinomas and large cell carcinomas are differentiated from other lung cancers[13]. We also found that the circadian rhythm pathway was deregulated with several metabolic pathways. Cancer patients are known to have altered circadian rhythms, which is important when considering the timely administration of chemotherapy[14].

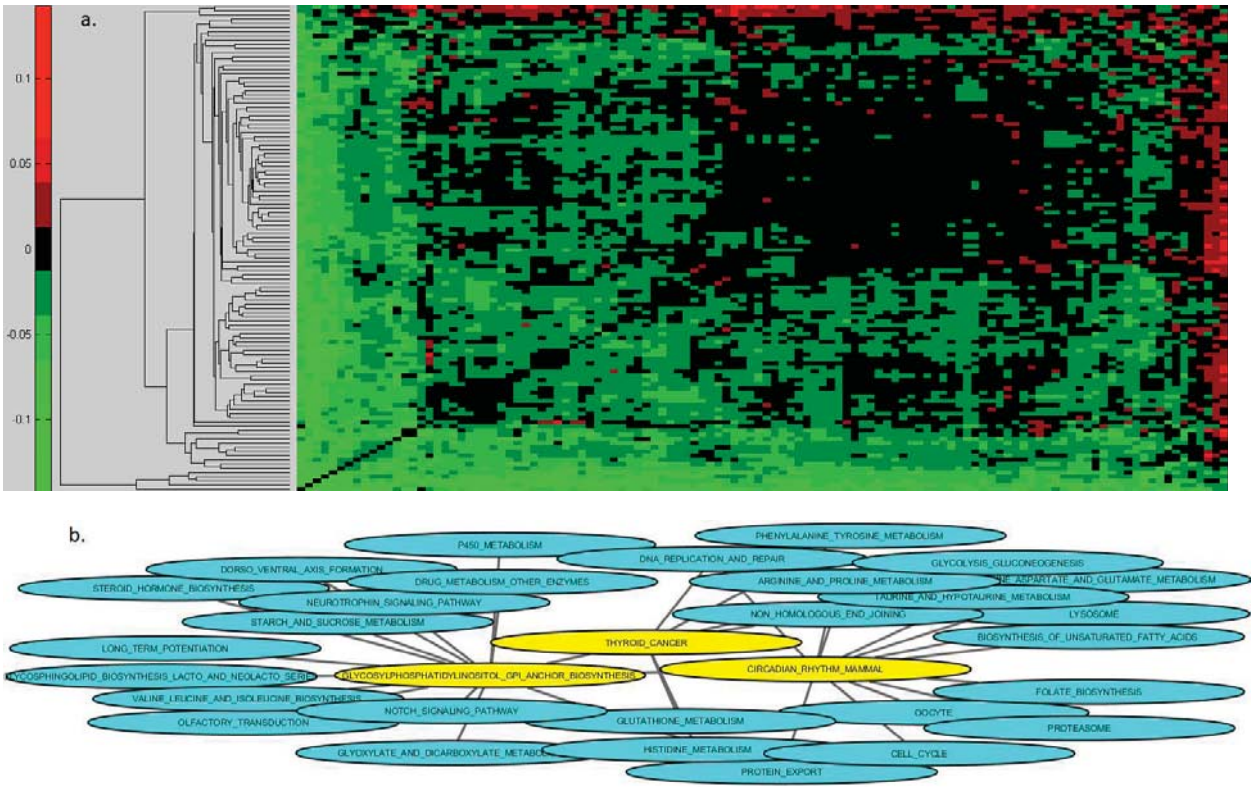

Figure 1. (a) Clustergram of the Lung Cancer pairwise pathway distance correlation with glycophosphatidyl (GPI) anchor synthesis, circadian rhythm, and thyroid cancer pathways in the three leftmost columns and bottommost rows. **(b)** The network diagram of the pathway pairs with high differential dependence, with GPI anchor synthesis, circadian rhythm, thyroid cancer (in yellow) as the clear hub nodes of this network

Discussion

The behavior and fate of a cell can only be understood in the context as an interlocking machine, not as a set of disconnected parts. Pathways are in many ways artificial groupings set up to help humans organize the functions of genes. Enzymes of distinct pathways share cofactors, and the product of one pathway may be the substrate of another. Understanding how these pathways interact is key to identifying the effect on the cell that is created by altering a gene. In this paper, we focus at the interactions among pathways, which provides complementary insights with reduced number of elements in the system. Our method is based on widely available gene expression data. Our use of distance correlations enables multivariate analysis without the need to identify correlated genes. Combining distance correlation and networks also breaks from a long tradition of using differential gene expression to identify important pathways. Our method is simple, requires no parameter input from the user, and it seeks to answer a fundamental question of biology: are two pathways dependent? Using more extensive pathway databases and predesigned datasets, we could explore pathway dependency in greater detail, and perhaps even elucidate the underlying genes responsible for the pathway dependency. The results on a large lung cancer demonstrated the effectiveness of our method in generating new insights on pathway interactions during the disease process.

References

1. Szekely G, Rizzo M, Bakirov N. Measuring and Testing Dependence by Correlation of Distances. The Annals of Statistics 35;6: 2769-94 2007
2. Wang T., Gu J., Yuan J., Tao R., Li Y., Li S. Inferring Pathway Crosstalk Networks Using Gene Set Co-Expression. Molecular BioSystems 9;7: 1822-8 2013
3. Huang Y., Li S. Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network. BMC Bioinformatics 11 Suppl. 1:S32, Jan. 2010

4. Ponzoni I, Nueda M, Tarazona S, Gotz S, Montaner D, Dussaut J, Dopazo J, Conesa A. Pathway Network Inference From Gene Expression Data. BMC Systems Biology 8(Suppl 2):S7, 2014
5. Dutta B, Wallqvist A, Reifman J. PathNet: a Tool for Pathway Analysis Using Topological Information. Source Code for biology and Medicine, 7(1):10, Jan. 2012
6. Cho SB, Kim J, Kim JH. Identifying Set-Wise Differential Co-expression in Gene Expression Microarray Data. BMC bioinformatics 10:109, Jan. 2009.
7. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Farez-Vidal ME. Gene Expression Profiling Reveals Novel Biomarkers in Nonsmall Cell Lung Cancer. International Journal of Cancer 129(2):355--64, July 2011.
8. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1):27-30, Jan. 2000.
9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome research 13(11):2498--504, Nov. 2003.
10. Hammarstrom S. The carcinoembryonic antigen (CEA) family: Structures, Suggested Functions and Expression in Normal and Malignant Tissues. Semin. Cancer Biol., 9(2):67--81.
11. Dolezal S, Hester S, Kirby PS, Nairn A, Pierce M, Abbott KL. Elevated Levels of Glycosylphosphatidylinositol (GPI) Anchored Proteins in Plasma from Human Cancers Detected by C. Septicum Alpha Toxin. Cancer biomarkers 14(1):55--62, Jan. 2014.
12. Wu G, Guo Z, Chatterjee A, Huang X, Rubin E, Wu F, Mambo E, Chang X, Osada M, Sook Kim M, Moon C, Califano JA, Ratovitski EA, Gollin SM, Sukumar S, Sidransky D, Trink B. Overexpression of Glycosylphosphatidylinositol (GPI) Transamidasesubunits Phosphatidylinositol Glycan Class T and/or GPI Anchor Attachment 1 Induces Tumorigenesis and Contributes to Invasion in Human Breast Cancer. Cancer Research, 66(20):9829--36, Oct. 2006.
13. Kaufmann O and Dietel M. Thyroid Transcription Factor-1 is the Superior Immunohistochemical Marker for Pulmonary Adenocarcinomas and Large Cell Carcinomas Compared to Surfactant Proteins A and B. Histopathology 36(1):8--16, 2000.
14. Mormont MC and Levi F. Circadian-System Alterations During Cancer Processes: A Review. International Journal of Cancer 70(2):241--7, Jan. 1997.