Network Analysis of Genes Regulated in Renal Diseases: Implications for a Molecular-Based Classification

Suresh K. Bhavnani¹,²,⁵ PhD, Felix Eichinger³,⁵, Sebastian Martini³,⁵ MD, Paul Saxman², H.V. Jagadish¹,⁴ PhD, Matthias Kretzler¹,³ MD

¹Center for Computational Medicine & Bioinformatics, ²Michigan Institute for Clinical & Health Research, ³Dept. of Internal Medicine, ⁴Comp. Science & Engineering, Univ. of Michigan, Ann Arbor, MI; ⁵Contributed Equally.

Abstract

Chronic renal diseases are currently classified based on morphological similarities. Such classifications do not reliably predict the course of the disease and its response to therapy. In contrast, recent studies in other diseases suggest that a classification which includes molecular information could lead to more accurate diagnoses and prediction of treatment response. We therefore extracted gene expression profiles from biopsies of patients with chronic renal diseases, and used network visualizations and associated quantitative measures to rapidly analyze similarities and differences between the diseases. The results revealed unexpected combinations of renal diseases that share relatively large numbers of genes. These results demonstrate (1) the utility of network analyses to rapidly understand complex relationships, and (2) the need to define a new molecular-based classification of renal diseases.

Introduction

The rapid development of molecular biology and powerful analytical methods such as network analysis are enabling a shift in our understanding of diseases from a morphological (based on clinical and histological findings) to a molecular basis. This shift in focus has led to improvements in the classification of diseases. For example, gene expression analyses have been shown to improve prediction of treatment response in several diseases such as breast cancer [1].

Unfortunately, relatively little is known about how renal diseases are similar and different at the molecular level. Currently, renal diseases are classified largely on morphological similarities. For example, systemic lupus erythematosus (SLE) is classified as a “predominant” inflammatory disease based on clinical findings, whereas focal and segmental glomerulosclerosis (FSGS) is classified as a “predominant” non-inflammatory disease based on histology. Several studies suggest that such morphology-based classifications could be significantly improved through the analysis of similarities and differences in gene expression, leading to more accurate diagnosis and targeted treatment options [2].

The analysis of gene expressions in chronic renal disease has either been studied at the level of a single renal disease [e.g., 3], or by studying gene expressions across all known Mendelian disorders in the OMIM database [4]. The former obviously cannot reveal gene expressions that are common across renal diseases; the latter analyzed renal genes based on limited data, and at a high level (glomerular versus tubular), and therefore excluded important disease subcategories such as SLE. This paper attempts to directly address the lack of understanding about gene expressions in chronic renal diseases. By using new data at the appropriate level of granularity, our goal was to evaluate the current classification of renal diseases, and generate hypotheses about the mechanisms underlying those diseases.

We begin by describing how we assembled a dataset of renal diseases and implicated genes, why and how we represented it using networks, and how we analyzed the networks using visualizations and quantitative verifications. We then discuss how the network analysis rapidly revealed unexpected overlaps of genes across the diseases. We conclude with discussing the utility of the network analysis approach to rapidly understand complex relationships, and the need to define a molecular-based classification of chronic renal diseases.

Method

Our research began with the question: What are the molecular similarities and differences between chronic renal diseases? If gene expressions occur in patterns that match the current classification of renal disease, then we can infer that the current classification is sufficient. However, if diseases have unexpected gene expression similarities, then we can infer that the current classification of renal diseases needs re-evaluation. To address our research question, we made critical decisions regarding data selection, data representation and data analysis as discussed below:

Data Selection. Gene expression data were obtained from 106 patients with one of seven chronic renal diseases (classified in three categories, see Table 1) seen in one of 28 medical centers who received a
renal biopsy in conjunction with diagnostics of their disease. Gene expression data were compared with biopsies from healthy kidney donors (control). Due to the rarity of three diseases (MCD, TMD, and DN, colored gray in Table 1) they currently have very small sample sizes (less than five) in the experimental and/or control conditions.

Microdissected renal tubuli in the biopsies underwent gene expression analysis of 12029 genes in each sample using Affymetrix HG-U133A microarrays. This analysis was done to identify the significantly regulated genes compared to pre-transplant living donor kidney tissues (controls) in each disease. A gene was considered to be significantly regulated in a disease if: (1) the difference of the normalized expression values between control and disease samples was significant at the 0.05 level (after correcting for multiple testing with the false discovery method), and (2) the regulation effect size (as defined by the log2 fold change standardized by a pooled standard deviation) of that gene exceeded +0.3 (for up-regulated) or -0.3 (for down-regulated) when compared to controls. These statistical comparisons between experimental and controls were made within the same expression analysis batch to control for variations in equipment and context. The rigorous controls and tests resulted in a dataset of 747 genes significantly implicated in 7 renal diseases.

**Data Representation.** Networks are increasingly being used to analyze a wide range of molecular phenomena, such as gene regulation and protein-protein interactions, and their relationships to diseases, symptoms and syndromes [4-8]. A network is a graph consisting of nodes and edges; nodes represent one or more types of entities (e.g., diseases or genes), and edges between the nodes represent a specific relationship between the entities (e.g., a disease is significantly correlated with a gene). Figure 1 shows a bipartite network (where edges exist only between two different types of entities) of diseases and their implicated genes.

Networks have three advantages for analyzing complex relationships. (1) They do not require *a priori* assumptions about the data, such as whether the data are hierarchically clustered or contain fuzzy clusters. Instead, by using a simple pair-wise representation of nodes and edges, networks enable rapid discovery of complex relationships using a single representation. (2) The specificity provided by the pair-wise representation between nodes can reveal details of relationships, for example how specific diseases are connected to specific genes. (3) They can be rapidly visualized and analyzed using a set of network algorithms to reveal global regularities in the data. For example, Figure 1 shows how a *force-directed* layout algorithm [7] helps to visualize the relationship between diseases and genes. The algorithm pulls together nodes that are tightly connected, and pushes apart nodes that are not. As shown, the result is that diseases that share genes (e.g., FSGS and SLE in the center of Figure 1) are placed close to each other, and close to their implicated genes. Given these advantages, we used networks to explore the relationship between renal diseases and their implicated genes.

**Data Analysis.** The insights from the network visualizations were verified using two standard quantitative network measures. (1) To provide a quantitative verification of the overall network topology observed in the visualization, we calculated the *degree* of each gene (number of diseases implicated with a gene) and plotted the *degree distribution* of the genes. (2) To understand more clearly how diseases had common gene expressions, we transformed the bipartite network to inspect only how the diseases shared genes in a method called a *one-mode projection* [7]. Here, all gene nodes were removed, and an edge was placed between two diseases if they shared one or more genes as shown in Figure 3. The resulting network visually represented how pairs of diseases shared gene expressions, revealing how two diseases share more or less genes than expected in comparison to the current classification of renal diseases. The networks were created using *Pajek* (version 1.23).

**Results**

As shown in Figure 1, the bipartite network visually represents the explicit relationships between the 7 renal diseases and the 747 expressed genes. Furthermore, the size of a node is proportional to its *degree* (number of edges that connect to that node). Therefore, the larger a node, the more edges it shares with other nodes. Finally, the light edges (orange...
colored) and dark edges (blue colored) between diseases and genes are significantly up and down regulated respectively. The bipartite network visualization revealed three critical patterns related to renal diseases and genes:

1. **Many Specific Genes, Few Non-Specific Genes.** There are a large number of specific genes in the periphery of the network that are connected to a single disease. These low degree nodes have been pushed out to the periphery by the force-directed layout algorithm due to their low connectivity with many diseases. In contrast, there are relatively fewer non-specific genes that are in the center of the network due to their high connectivity to many diseases. The degree distribution of genes (Figure 2) verifies this observation: more than half (54%) of the total 747 genes are implicated in just one disease, only two genes (0.3%) are implicated in 5 diseases, and no genes are implicated in all diseases. This result provided the first glimpse into the pattern of overlap between diseases and implicated genes. The network visualization therefore revealed the multidimensional relationship between diseases, genes, and regulation type. The gene expression overlap is verified by a one-mode projection on diseases. As shown in Figure 3, the one-mode projection removed the gene nodes, and placed edges between the diseases to correspond with how many genes each pair shares. The one-mode projection is not designed to reveal whether more than two diseases share the same genes, the graph clearly shows the dominant relationship between SLE, FGS, MGN, and a less dominant relationship with IgAN. Future studies with larger samples of TMD, DN, and MCD should reveal how they relate to the other renal diseases.

2. **Dominant Disease Sets.** To understand how genes overlapped across diseases, we analyzed different sizes of disease sets. There are three disease sets which share a disproportionately large number of genes: (a) the four dominant diseases (SLE, FSGS, MGN, IgAN) on the right hand side of the network share 52 (88%) of the total 4-degree genes. These genes are mainly down-regulated. (b) A proper subset of the above disease set (SLE, FSGS, MGN) share 88 (79%) of the total 3-degree genes. These genes are mainly up-regulated. (c) A pair of diseases (SLE, FSGS) which overlap with the above sets share 130 (72%) of the total 2-degree genes. These genes are mainly up-regulated.

Figure 1. A bipartite network on the left (automatically generated by the Fruchterman Rheingold algorithm [7]) showing the relationship between 7 renal diseases (white nodes), and 747 genes (black nodes). The size of the nodes is proportional to the edges that connect to them. Therefore diseases with many genes have large nodes, whereas diseases with few genes have smaller nodes. The light edges (orange colored) represent up-regulated genes, the dark edges (blue colored) represent down-regulated genes, and gene node labels are the numerical identifiers for each gene from the Entrez Gene database. The inset shows that the genes regulated by all four of the high degree diseases (SLE, FSGS, MGN, and IgAN) are mostly down-regulated (shown as having mostly dark edges), whereas those regulated by a subset of the diseases shown (SLE, MGN, and FSGS) are mostly up-regulated (shown as having mostly light edges).
3. Concordance in Gene Regulation. All the genes in the network, regardless of degree, are either up or down regulated. In other words, no gene was up-regulated in one disease, and down-regulated in another. This uniform concordance in gene regulation can be seen by the large areas of uniform color for edges connecting to high degree genes. Given the 100% uniformity of gene concordance, we re-examined the data to check for programmatic and bias errors, and found none. Furthermore, we examined another dataset containing biopsies from patients with acute renal failure. When the two data sets were merged, we found two genes that were discordantly regulated. This suggests that the uniformity in gene regulation within chronic renal diseases (presented here) is most probably the result of similarity in biological mechanisms across chronic renal diseases, rather than a selection bias or error. Future research should further verify this conclusion.

Discussion

Given that the data consisted of only renal biopsies, we expected to find a large number of non-specific (shared) genes. Instead, we found relatively few non-specific renal genes. This relationship held even when we removed diseases with a low sample size. This unexpected result was possible to observe because we used networks to analyze a set of renal diseases simultaneously. It is important to note that the many specific genes in our dataset could be implicated in other renal diseases not included in our study, and therefore could be non-specific with respect to a wider scope of diseases.

Besides the distribution of specific and non-specific genes, the network analysis also revealed patterns of molecular similarity between diseases which do not match the current morphology-based classification of renal diseases. As shown in the first column of Table 1, SLE and IgAN belong to the class of inflammatory diseases. However, the network analysis revealed that SLE shares many more genes with FSGS and MGN (from the non-inflammatory class), compared to IgAN (from its own class). While similarities between non-inflammatory glomerular and inflammatory diseases have been previously reported, the unexpected finding was the strength of the association with members outside its class. Similarly, IgAN shares an equal number of genes with SLE (from its own class) as it does with FSGS, and MGN (from the non-inflammatory class).

Overall, the above relationships suggest that the current morphology-based classification of renal diseases does not match the pattern of shared tubulo-interstitial gene expression, and therefore motivates future research to define a molecular-based classification of renal diseases. Hypotheses to explain this result include: (1) a yet-to-be identified pathway that is common to all progressive renal diseases, (2) tissue compartment-specific pathology activated in a subset of the diseases, (3) transcriptional pathways as elements of a primary disease pathogenesis which might be shared between current disease categories.

Genes common to disease sets can also be used to identify existing gene regulatory pathways [8]. For example, we used canonical pathways (developed by Ingenuity® Systems) to search for existing regulatory pathways that best matched the genes shared by the three dominant disease sets. For the 59 genes shared
by the disease set FSGS, MGN and IgAN, the search retrieved 45 canonical pathways with a p<0.01 many of them experimentally-verified (TGF-β, JAK/STAT, NF-κB and VEGF pathways).

The shared repression of the VEGF pathways suggest vascular rarefaction as an underlying driving force for ischemic damage in renal failure [9], and a potential biomarker for progressive renal disease. However, it is possible that shared genes do not match known pathways, which would suggest the existence of new pathways yet to be discovered. These new pathways could be important in the pathophysiology of the diseases. The network analysis has therefore led to testable hypotheses about underlying pathophysiological mechanisms.

Finally, the concordance in gene regulation and the fact that genes in the three dominant disease sets are either mostly up or down-regulated require further investigation by analyzing the properties of the shared genes using categories from the Gene Ontology database. Patterns in how the gene categories relate to different disease sets should lead to an understanding of this and other phenomena related to the type of gene regulation.

Conclusion and Future Research

While networks have been used to analyze a wide range of molecular phenomena, they have not been used to analyze how genes are implicated by multiple renal diseases. Our analysis has enabled us to question the adequacy of the existing morphological based taxonomy of renal diseases. Furthermore, the analysis rapidly revealed useful biological insights, without requiring additional filtering to reveal complex but understandable relationships. This could be because the network was of medium size and density compared to many large networks. In addition, the resulting network quickly revealed overlapping, nested, and subset groups from the same representation, a result that would be difficult using traditional data mining techniques. However, it is important to note that like most data mining techniques, network analysis is essentially an exploratory tool, and most useful for generating hypotheses, which need rigorous testing using other techniques to arrive at definitive answers.

The limitation of the current analysis is the small sample sizes for three diseases, which new data will soon address. However, similar to the Diseasome project [4], studies that attempt to analyze gene expressions of many diseases simultaneously often have to deal with incomplete data. Networks are surprisingly useful for incomplete data because they enable us to visually inspect the data, and make appropriate choices for filtering and interpretation.

Our future research includes: (1) Using categories from the Gene Ontology database to annotate the genes in the bipartite network, with the goal of understanding why sets of shared genes have similar regulation type. (2) Comparing the current results with those generated from other methods such as fuzzy clusters. (3) Analyzing a network consisting of individual patients and expressed genes. The goal of analyzing individual patients is to construct a new classification of renal diseases using a bottom-up approach without the use of a priori disease classifications as was done in the current study. As previous research on biomarkers in renal diseases have stated, gene expression data should lead to a systematically constructed molecular-based classification, resulting in the identification of more targeted treatments for patients with chronic renal disease.

Acknowledgements

This study is funded by NIH grants # UL1RR024986, # U54 DA021519, and R01 DK079912-01. We thank J. Kurlander and A. Ganesan for their contributions.

References

1. Wulfkuhle JD, Speer R, Pierobon M, Laird J, et al. Multiplexed cell signaling analysis of human breast cancer applications for personalized therapy. J Proteome Res 2008; 7:1508-17.
2. Loscalzo J, Kohane I, Barabasi A-L. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 2007; 3:124.
3. Martini S, Eichinger F, Nair V, Kretzler M. Defining human diabetic nephropathy on the molecular level: Integration of transcriptomic profiles with biological knowledge. Rev Endocr Metab Disord 2008; 9:267-74.
4. Goh KL, Cusick ME, Valle D, Childs B, et al. The human disease network. Proc Natl Acad Sci USA 2007; 104:8685-8690.
5. Oti M, Brunner HG. The modular nature of genetic diseases. Clin Genet 2007; 71:1-11.
6. Ideker T, Sharan R. Protein networks in disease. Genome Res 2008; 18:644-652.
7. Newman, M. The structure and function of complex networks. SIAM Review 2003; 45(2):167-256.
8. Sam L, Liu Y, Li J, Friedman C, et al. Discovery of Protein Interaction Networks Shared by Diseases. Pacific Symposium on Biocomputing. 2007; 12:76-87.
9. Lindenmeyer MT. Interstitial vascular rarefaction and reduced VEGF-A expression in human diabetic nephropathy. J Am Soc Nephrol. 2007; 18:1765-7.