

BY KARTIK MANI, PhD

Network-based Approaches to Prediction of Disease Genes



The recent surge of high-throughput experimental data, such as gene expression microarrays, offers a profound opportunity to gain a more detailed understanding of the genes involved in the progression of disease. While initial analyses of these data used statistical techniques to identify genes capable of distinguishing disease tissue from normal (biomarkers), researchers are now turning to the analysis of gene interaction networks to address this problem.

Gene interaction networks may be developed from several sources including manual curation, high-throughput experiments (such as yeast 2-hybrid), literature mining and reverse engineer-

ing, or other), which control a large set of genes differentially expressed in the disease state. The third, the focus here, relies on the fact that interaction networks are themselves dynamic and may change from a normal to disease state. Thus, if one identifies interactions that have actually changed between phenotypes, one might then work backwards to identify genes that could prove promising for further investigation.

We will detail two examples of the third category, both of which incidentally use an information-theoretic approach. The first defines a concept called synergy, which measures the cooperative effect of two variables on the state of a third. The two variables in this

one particular disease phenotype (P). Formulaically, this test is represented as the difference (ΔI) between $I_{all}(G1;G2)$ and $I_{all-P}(G1;G2)$, where I_{all} includes all sample points, and I_{all-P} excludes the phenotype P. Biologically, a positive or negative ΔI implies that these two genes have gained or lost an interaction in the phenotype P respectively (e.g., an oncogene “loses” its ability to be regulated in cancer). The genes participating in a statistically significant number of these interactions are then selected. When applied to data from three primary B cell lymphomas, IDEA correctly predicted the known oncogenes reported in the litera-

If one identifies interactions that have actually changed between phenotypes, one might then work backwards to identify genes that could prove promising for further investigation.

ing algorithms. They can include many different types of interactions as well (complexes, regulatory, signaling, etc). Integrating and analyzing all of this information to discover genes relevant to disease requires network-based algorithms. Thus far, such algorithms fall into three general (though not necessarily mutually exclusive) categories. The first predicts protein complexes, rather than individual genes, associated with the disease phenotype. The second identifies key regulators (transcriptional, sig-

case are genes (G1 and G2), and the third is a binary state variable representing disease or normal (D). Formulaically, this can be represented as the difference between $I(G1,G2;D)$ (the cooperative effect) and the sum $I(G1;D) + I(G2;D)$ (the individual effects), where I is mutual information. Biologically, synergistic interactions imply that the combined state of the two genes affects disease, while individually the genes have a far lesser or no effect. This algorithm computes this quantity across all gene pairs represented on the input microarray data, and a “synergy network” is generated from the highest scoring interactions. When applied to publicly available prostate cancer data, this approach showed the *RBP11* gene participating in a large number of synergistic interactions. This finding along with others indicated that the progression of prostate cancer is linked with oxidative stress and inhibition of the apoptosis pathway, consistent with previous hypotheses.

The second algorithm, Interactome Dysregulation Enrichment Analysis (IDEA), computes the mutual information between two genes across a large, diverse dataset, including or excluding

ture (e.g., MYC in Burkitt’s Lymphoma), as well as effector genes not identified by differential expression analysis.

These network-based approaches, along with others, have shown promise in more accurately delineating the mechanisms of disease progression. Like any new class of methods, however, there are drawbacks. First and foremost, there is no “gold standard” of gene interactions that can be used, although the knowledge base is growing rapidly. They often require large training sets or sample diversity to be effective, which may not always be available. Lastly, computational complexity may limit their applicability.

Nevertheless, the application of networks and these algorithms to the identification of disease-causing genes remains an exciting new area of computational biology. Expect to see several new network-based approaches emerge as the body of high-throughput and interaction-based data continues to grow.

REFERENCES

1. Watkinson, J., X. Wang, et al. (2008). *BMC Syst Biol* 2: 10.
2. Mani, K. M., C. Lefebvre, et al. (2008). *Mol Syst Biol* 4: 169. □

DETAILS

Kartik Mani received his PhD in Biomedical informatics at Columbia University, working in the Multi-Scale Analysis of Genomic and Cellular Networks (MAGNet) Center under the direction of Dr. Andrea Califano. His research focused on the application of interaction networks to gene-disease association, and culminated in the development of the IDEA algorithm described above. He is currently pursuing his MD at the Albert Einstein College of Medicine in Bronx, NY.