EDITORIAL PREFACE

Network-Based Approaches to Personalized Medicine: Challenges and Opportunities

Aryya Gangopadhyay, Department of Information Systems, University of Maryland Baltimore County (UMBC), Baltimore MD, USA

INTRODUCTION

Personalized medicine is a new approach to medicine that takes into account the pathophysiology of the patient, molecular information such as genomic, proteomic, and metabolomic processes and interactions, and pharmacodiagnostic tests to prescribe the "right drug(s) for the right patient". This approach is radically different from the traditional "blockbuster" approach for drug development, marketing, and prescription. Personalized medicine brings tremendous potential for the treatment of diseases by avoiding roadblocks such as adverse side effects and ineffective responses to medication. Personalized medicine also brings tremendous challenges in that it can revolutionize the entire pharmaceutical industry by changing it from its current "big pharma" model to the personalization of drugs and treatment outcomes. Personalized medicine will require technological advances such as next generation sequencing, understanding the interactions among intra and inter-cellular components, and pharmacodiagnostic tests to individualize drug prescription for each individual patient. This will require a radical shift from large-scale randomized clinical trials prevalent in current practice for drug development to more targeted, homogeneous subgroups of patients that are more likely to respond positively to drugs due to their similarities at the molecular level (Jørgensen, 2008). Hence, personalized medicine is one of the grand challenges in science, technology, and the healthcare industry of the twenty-first century.

EXAMPLE APPLICATIONS OF PERSONALIZED MEDICINE

Some aspects of personalized medicine are already in use for the treatment of diseases such as cancer. For example, in treating breast cancer a pharmacodiagnostic test is performed to determine whether the human epidermal receptor (HER2) protein is over-expressed in the tumor cells, which happens in 20-25% of the

cases. In such cases the monoclonal antibody trastuzumab (Herceptin®, Genentech, CA, USA) has been found to be an effective adjuvant to chemotherapy (Slamon et al 2001). A number of anticancer drugs that target specific gene products resulting from generic variations exist for other cancers including colorectal cancer, gastrointestinal stromal tumors, and head and neck cancer. In all of these cases pharmacodiagnostic tests are performed to identify patients that are likely to respond positively to the treatment.

Genetic factors can also affect treatments in other ways. For example, patients with mutations of the melanocortin-1 receptor gene require higher doses of anesthetics as compared to those without such mutations (Liem et al., 2005). Another example of the importance of genetic factors is for drugs that are metabolized through the CYP enzymes. Patients with polymorphisms of the CYP2D6 and CYP2C19 genes can experience adverse side effects to many drugs such as β-blockers, antidepressants, antipsychotics, and proton pump inhibitors. An example of a phramacodiagnostic test is the AmpliChip CYP450 test[1] that can detect polymorphisms of CYP2D6 and CYP2C19 genes that play a significant role in the metabolism of a large number of prescription drugs.

BIOLOGICAL NETWORKS

Recent advances in systems biology have given rise to the notion of the human "interactome" that can be modeled as graphs of the biomolecular system and analyzed to gain insights into the pathogenesis of diseases and responses to treatments. This is a radical departure from the approach of one-gene, one-disease, one-drug model to interactions within and between elements of multimodal networks such as protein-protein interaction networks, metabolic networks, regulatory networks, and genetic networks. Interesting similarities have been found between the properties of real world networks such as social networks and biological networks such as protein interaction networks. In a protein-protein interaction network a node represents a protein and an edge represents physical or binding interactions (Barabási et al., 2011).

A biological network is represented as a graph that contains a set of nodes and a set of edges connecting the nodes. The degree of a node is the number of other nodes it is directly connected to. If two nodes are not directly connected it might still be possible to traverse from one to the other through the other nodes in the graph. A connected component is a path in which all nodes are either directly or indirectly connected with each other. A fully connected graph will have one connected component, which is unusual in real-life graphs. On the other hand too many connected components indicate that most nodes are not linked to each other. Hence the number of connected components provides a measure of connectivity of the network. The diameter of a graph is length of the path that connects two of the furthest nodes in the graph, without backtracking, taking detours, or going around loops. Hence the diameter is sometimes referred to as the "longest shortest path". The degree distribution of a graph, $P(k)$ describes the probability that a node will have a degree k . Typically, the degree distribution of real world networks exhibit a power law distribution, which is an L-shaped graph where the x-axis represents the degree and the y-axis represents the number of nodes with the corresponding degree. This indicates that very few nodes have very large degrees whereas most nodes have very small degrees. The web pages on the Internet exhibit this clearly. For example websites such as google.com have very high in-degrees (number of sites that link to them) whereas most of the web pages have much smaller in-degrees. In contrast, random networks have Gaussian degree distributions. The degree distribution of protein-protein interaction networks have been shown to obey the power law distribution described by the equation below:

$$P(k) \approx k^{-\gamma} \quad (1)$$

where $P(k)$ is the probability that a protein will have a degree k, and γ is a constant whose value is reported to gave the value 1.57 in the human protein-protein interaction network (Bader et al 2007). Important information can be gleaned from the properties of such networks including the number of connected components, average component size, diameter, and average path length. Furthermore, centrality measures of nodes can provide important information about the significance of proteins that are hubs connecting sub-networks. The network topology can also provide a way to compare the networks of different organisms.

Another network property is the clustering coefficient, which is a measure of the connectivity of the nodes in a graph. The neighborhood of a node consists of all the nodes it is directly connected with. The clustering coefficient for v is given by equation (2) below, where n_v is the actual number of edges connecting the neighboring nodes of v:

$$c(v) = \frac{2.\, n_v}{d.(d-1)} \tag{2}$$

In other words, the clustering coefficient of a node is the ratio of the actual to the maximum possible number of connections in its neighborhood. The distribution of the clustering coefficient tends to be uniform in both random and scale-free networks, which indicates that neighborhoods are roughly equally connected for nodes with all degrees.

NETWORK PROPERTIES

In this section we discuss some implications of network properties that can be used to characterize its behavior:

- **Network Connectivity:** Networks can vary in their degrees of connectivity, from tightly to sparsely connected. As expected, nodes tend to affect each other more quickly in tightly connected networks than in sparsely connected networks. The rate at which a viral infection may die out depends on the connectivity of the network and when the connectivity is above a threshold a viral infection can take on an epidemic proportion (Chakrabarti, 2005). It is well known from protein-protein interaction networks that proteins that have high levels of interactions are also involved in the same cellular processes, and mutations in causative genes may result in the same disease phenotypes (Barabási et al., 2011, Oti et al., 2006). Thus, highly interlinked local regions in a biological network may suggest the existence of topological modules that might be correlated with disease networks (Albert 2005);
- **Scale Free Networks:** The degree distribution of scale free networks follows the power law shown in equation (1). The term "scale free" refers to the fact the functional form does not change with network size. Many biological networks are scale free networks. For example, the flux distribution in the central metabolism of *Escherichia coli* follows the power law, which implies that most reactions have small metabolic fluxes while a few reactions with high fluxes carry most of the metabolic activities. Jeong et al (2000) studied the metabolic networks of 43 different organisms from all three domains of life: eukaryotes, prokaryotes, and archaea, and found that all of the metabolic networks they studied are scale free networks. Other biological networks such as gene regulatory networks where the nodes represent genes and the edges represent expression correlations also exhibit scale free properties (Vogelstein et al., 2000). Transcription factors are molecules that control the activities of genes, and the number of number of genes that are controlled by a transcription factor also exhibit scale free behavior: most transcription factors regulate a few genes whereas a few transcription factors

regulate a large number of genes. However, only a few transcription factors regulate a given gene;

- **Small World Property:** The small world property of a network refers to the fact that most nodes in the network can be connected via a small number of links. Random networks exhibit small world property, while scale free networks exhibit ultra small world effects. In metabolic networks most metabolites are within 3-4 links from each other, which implies that local perturbation can affect the whole network very rapidly. The diameter of the metabolic networks of simple organisms such as the parasitic bacterium have been found to be similar to that of larger and more complex multicellular organisms, which is an interesting fact from the evolutionary standpoint;
- **Disassortativity:** an interesting network property, assortativity, is the correlation of nodes that are directly connected with each other. A highly assortative network is one in which directly connected nodes have a high correlation, which is often measured in terms of their degrees. Biological networks such as protein-protein interaction networks are disassortative, which implies that hubs, or nodes with larger degrees, are less likely to be directly connected. On the other hand, nodes with smaller degrees are more likely to be directly connected with hubs. Social networks, on the other hand, tend to be assortative where individuals with many "friends" also tend to connect with each other. While the reasons behind the disassortative nature of biological networks remain unclear, some interesting inferences are possible. For example, such networks are more vulnerable to selective perturbation such as removal of the hubs. At the same time, such networks are resilient to random perturbations since most of the nodes are non-hubs and hence removal of such nodes will not impact the connectivity of the entire network;
- **Hierarchy:** It has been shown that cellular networks exhibit hierarchical topologies. However, the topologies of biological networks are not tree structures because they contain hubs with very high degrees. Structurally, cellular networks consist of highly connected sub-graphs or modules that are associated with specific tasks or functions. Individual nodes are connected to one or more modules. A node at the lowest level of a hierarchy is part of one module and such a node participates in a single functional task. Bridge nodes connecting several modules are at a higher level of the hierarchy. These nodes are part of two or more modules and hence participate in multiple functions. Nodes at the highest level of the hierarchy act as hubs that do not belong to any specific module but are the sole connections between sub-graphs. The hierarchy of a node can be determined by its clustering coefficient, as explained in Equation (2). Nodes at higher levels have lower clustering coefficients that those at lower levels in the hierarchy;
- **Motifs:** Motifs are significant patterns or sub-graphs that occur much more frequently than are expected to occur in randomized networks (Milo et al., 2002). Given the topological hierarchy and the correspondence between modules and their biological functions, it is important to identify the important motifs in cellular networks. A motif is deemed important if it is over-represented in the network, because natural selection will result in more abundance of Motifs that carry important biological functions rather than those occurring randomly. The first systematic study of network motifs was done *Escherichia coli*, and the same motifs were found in subsequent studies in bacteria, yeast, plants, and animals (Aron, 2007). Milo et al (2002) described 13 types of 3-node sub-graphs for networks in general. Three and four node motifs with special

significance include bi-fan, feed-forward loop, bi-parallel, and feedback loops. Alon (2007) provides a functional interpretation of motifs that are characterized as simple regulation, feed-forward loops (FFL), single input modules (SIM) and dense overlapping regulons (DOR). Motifs can help in understanding biological interactions such as transcriptional regulations of genes, co-expressions of genes, and protein-protein interactions. Motifs, in turn, can form clusters in the integrated network or genes, proteins, and other macromecules.

The algorithm for detecting motifs in a network can be loosely described as follows: (1) identify all n node sub-graphs in the network, (2) randomize the network while keeping the same number of nodes, edges, and degree distribution as the original network, (3) repeat step (1) for the randomized network created in step (2), and (4) identify the sub-graphs that occur with significantly higher frequency in the original network as compared to the randomized network as motifs.

CHALLENGES

In spite of the advances made in recent years to understand the structure and function of biological networks, this area is still in its infancy. For example Pržulj et al (2004) challenged one of the basic characterizations of biological networks by demonstrating that the protein-protein interaction networks of yeast *S. Cerevisiae* and fruitfly *D. Melanogaster* resemble random geometric networks more closely than scale-free networks both in terms of global network properties such as clustering coefficient as well as local measures such as motifs (the term graphlets was used in Pržulj et al 2004). Another limitation of current state of the art is that network motifs have been experimentally studied only in bacteria. Experimental studies of network motifs are needed to understand the functions in eukaryotic organisms. Such studies may discover new motifs and help explain the dynamics of large networks based on the interactions at the motif level.

At a more basic level, biological networks such as protein-protein interaction networks are incomplete and tissue specific as opposed to universal and multi-cellular organism based. Interactions in biological networks have spatial and temporal properties. However, most biological networks have been derived from samples that do not capture the dynamic nature of the interactions.

Perhaps the biggest challenge that is yet unaddressed in biological networks is the analysis of interactome as a whole. While such networks have been studied individually, including protein-protein interaction networks, genetic networks, metabolic networks and disease networks, there is a need to study these multi-modal networks as a whole. Hence it is necessary to develop methods for multi-modal network analysis to get a complete understanding of the entire interactome of the complex, eukaryotic organisms.

Aryya Gangopadhyay
Editor-in-Chief
IJCMAM

REFERENCES

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, *118*, 4947–4957. doi:10.1242/jcs.02714 PMID:16254242.

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews. Genetics*, *8*, 450–461. doi:10.1038/nrg2102 PMID:17510665.

Bader, D. A., & Madduri, K. (2007, March 26). A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms. In *Proceedings of the Sixth IEEE International Workshop on High Performance Computational Biology* (HiCOMB), Long Beach, CA.

Barabási, A.-L., Gulbache, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. *Nature Reviews. Genetics*, *12*(1), 56–68. doi:10.1038/nrg2918 PMID:21164525.

Chakrabarti, D. (2005). *Tools for large graph mining*. Unpublished PhD dissertation, Carnegie Melon University.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*, 651–654. doi:10.1038/35036627 PMID:11034217.

Jørgensen, J. T. (2008). From blockbuster medicine to personalized medicine. *Personalized Medicine*, *5*(1), 55–63. doi:10.2217/17410541.5.1.55.

Liem, E. B., Joiner, T. V., & Tsueda, K. et al. (2005). Increased sensitivity to thermal pain and reduced subcutaneous lidocaine efficacy in redheads. *Anesthesiology*, *102*, 509–514. doi:10.1097/00000542-200503000-00006 PMID:15731586.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, *298*, 824–827. doi:10.1126/science.298.5594.824 PMID:12399590.

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, *43*(8), 691–698. doi:10.1136/jmg.2006.041376 PMID:16611749.

Pržulj, N., Corneil, D. G., & Jurisica, I. (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics (Oxford, England)*, *20*(18), 3508–3515. doi:10.1093/bioinformatics/bth436 PMID:15284103.

Slamon, D., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., & Bajamonde, A. et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *The New England Journal of Medicine*, *344*, 783–792. doi:10.1056/NEJM200103153441101 PMID:11248153.

Vogelstein, B., Lane, D., & Levine, A. J. (2000). Surfing the P53 network. *Nature*, *408*, 307–310. doi:10.1038/35042675 PMID:11099028.

Zhang, L. V., King, O. D., Wong, S. L., Goldberg, D. S., Tong, A. H. Y., & Lesage, G. et al. (2005). Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *Journal of Biology*, *4*(6). PMID:15982408.

ENDNOTES

[1] http://molecular.roche.com/assays/Pages/AmpliChipCYP450Test.aspx