IDENTIFICATION OF ABERRANT PATHWAY AND NETWORK ACTIVITY FROM HIGH-THROUGHPUT DATA

RACHEL KARCHIN
Department of Biomedical Engineering and Institute for Computational Medicine
Johns Hopkins University
Baltimore, MD 21218, USA
Email: karchin{at}jhu.edu

MICHAEL F. OCHS
Department of Oncology and Division of Oncology, Biostatistics and Bioinformatics
Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University
Baltimore, MD 21205, USA
Email: mfo{at}jhu.edu

JOSHUA M. STUART
Biomolecular Engineering
University of California Santa Cruz
Santa Cruz, CA 95064, USA
Email: jstuart{at}soe.ucsc.edu

JOEL S. BADER
Department of Biomedical Engineering and High-Throughput Biology Center
Johns Hopkins University
Baltimore, MD 21218, USA
Email: joel.bader{at}jhu.edu

Introduction

Biological functions are often explained in terms of networks and pathways. Innumerable college students have memorized canonical metabolic pathways, and signal transduction pathways such as MAPK, JAK/STAT and WNT have entered the biological lexicon. Inasmuch as these pathways provide a useful abstraction of biology, they can be used as a framework for understanding how mutations affect life processes and potentially cause disease.

Much work has been invested in mapping these and other pathways in human and simpler model organisms. Data sets are not pathways *per se*, but rather measurements of the individual interactions between proteins, genes, metabolites, drugs, and other species that define a network. In model organisms these networks can be perturbed by directed experiments. The first series of papers in this session explores how the experimental data sets can be analyzed and explained.

Directed perturbations of human networks *in vivo* are not possible, but genotype and phenotype data sets from individuals present a rich spectrum of information. The session concludes with analyses of the effects of primarily somatic mutations in cancer, focusing on the identification of subnetworks that may be responsible for disease phenotypes.

Genetic interaction networks in model organisms

Genetic epistatic interactions occur when mutations in two genes simultaneously create a phenotype different from the expectation from the two individual mutations. Genetic interactions are distinct from gene regulatory interactions, which refer to physical interactions between transcription factors and promoters. Large data sets of genetic interactions have been generated for yeast using a variety of experimental methods, which has motivated several contributions investigating the properties of these networks.

Bandyopadhyay *et al.* use gene expression data to assist in the prediction of genetic epistatic interactions in "SSLPred : Predicting Synthetic Sickness Lethality." The concept explored in this work is the feasibility of using network information to guide the generation of features to use in a predictive regression model. This work illustrates a recent theme in regression in a large feature space, the use of L1 regularization known as the lasso [1]. Lasso regression adds a penalty term that is related to the absolute value of parameter estimates. Quadratic programming and other methods permit fast identification of the global optimum, and, unlike ridge regression using L2 regularization, features can have regression coefficients that are set strongly to zero.

While Bandyopadhyay *et al.* provide global predictions for a genome-scale network, Carter *et al.* focus on quantitative understanding of a single network in "Predicting The Effects Of Copy-Number Variation In Double And Triple Mutant Combinations," investigate the contribution of genetic interactions to quantitative traits. The problem is important because of current interest in what has been called "missing heritability" in the context of human genome-wide association studies, where the contributions of known genetic factors remain below the total genetic contributions, which can be estimated accurately from epidemiology [2]. One possibility is that most analyses consider only additive models, in part because populations are not yet sufficiently large to have power to detect interaction terms. This work by Carter *et al.* is valuable in investigating non-additive contributions to a quantitative trait, defined by gene expression levels within the context of a detailed model of filamentous growth in yeast. The authors build on their earlier model of the gene regulatory network for filamentous growth [3] and use singular-value decomposition to investigate the dominant modes of transcriptional output [4]. The computational models are able to predict transcriptional phenotypes for perturbations including deletions, hypomorph alleles, and copy number variations, all important for human studies. Predictions for novel mutant combinations are in general concordance with experimental results. This manuscript points to the complexity of predicting phenotypes for multi-mutant combinations in yeast and illuminates the challenges ahead for predicting phenotypes from personal human genome data.

The subject of "Role of Synthetic Genetic Interactions in Understanding Functional Interactions Among Pathways" by Mohammadi *et al.* is the interpretation of genetic interactions in the context of other network data. Previous efforts have described genetic interactions in terms of within-pathway models (genetic interaction partners reside within the same pathway) and between-pathway models (genetic interactions occur between pathways with compensating or overlapping function). Comparisons with physical interactions [5, 6] and metabolic pathways [7] have typically favored the between-pathway model, particularly for deletion mutations that destroy the functionality of a linear pathway. The current contribution reexamines this problem using more recent data sets and pathways from KEGG to suggest functional connections between biological pathways in yeast.

Human data and local subnetworks

Data sets revealing network activity in human cancer are becoming available through The Cancer Genome Atlas (TCGA). These data sets provide an important test of the ability of methods developed for model organisms, primarily yeast, to scale to much more complex human systems involving more genes, larger networks, and distinct cell types.

In "Integrative Network Analysis to Identify Aberrant Pathway Networks in Ovarian Cancer," Chen *et al.* investigate the challenge of predicting cancer survival from genome-scale TCGA data involving copy number alterations, gene expression changes, and protein interactions. The rationale of TCGA is the hypothesis that genome-scale data will indeed result in improved methods for detecting, grading, and treating cancers. Demonstrations of effectiveness, including this work, are crucial for making progress and justifying future efforts related to TCGA. The authors approach this problem by identifying subnetworks of genes linked by protein-protein interactions that are effective at classification, then using support vector machines (SVMs) to perform the classification. The subnetwork learning illustrates an important theme in biological network analysis known as the "active subnetwork problem," in which a subset of genes or proteins strongly implicated in a process are linked together parsimoniously [8]. This problem can be expressed as a version of a classic problem in computer science known as the prize-collecting Steiner tree, known to be NP-hard, but for which efficient solvers have been recently invented [9, 10, 11].

Vandin *et al.* also analyze TCGA data in "Discovery of Mutated Subnetworks Associated with Clinical Data in Cancer," but with the goal of identifying subnetworks of genes with high mutation rates whose proteins are known to interact. This method adds to exploratory tools for investigating data from next-generation sequencing of tumors. The methods make use of graph diffusion kernels, which provide robust measures of association or similarity for graphs [12] and are the basis of the Google PageRank algorithm [13]. Graph kernels are a method of choice for extracting features for machine learning from network data, with applications ranging from predictions of protein similarity [14] to genetic epistasis [15].

The local subnetwork problem attacked by Chen *et al.* and Vandin *et al.* in these contributions has an analogous global subnetwork problem, carving a large network into subnetworks that reflect discrete biological processes or functions. Algorithms developed for clustering in other contexts, including powerful spectral clustering techniques [16, 17], have not performed well for biological networks or social networks. Modularity scores that maximize the enrichment of edges within groups are closely related to spectral clustering [18] and suffer from resolution limit problems in which small groups are merged inappropriately into larger groups [19]. Important recent progress has come from probabilistic models for hierarchical networks, essentially hierarchical stochastic block models. These methods can identify groups through Monte Carlo searches for small networks [20] and by approximate greedy algorithms for genome-size networks [21]. These methods have the advantage of extending readily to joint analysis of physical, genetic, gene regulatory, metabolic, and other types of interactions.

Converging problems and challenges

These examples highlight the use of networks as a framework for interpreting genome-scale data and predicting the response to new mutations. Important new challenges are motivated by the growing ability to obtain personal genome sequences, identifying variants that are unique to individuals. The future may involve a synthesis of methods designed for common variants, such as the copy number variants seen across TCGA samples, with methods designed for rare variants, leading to predictions of personal pathway activities and therapeutic options.

References

[1] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[2] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I Mccarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A Mccarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, October 2009.

[3] Gregory W Carter, Susanne Prinz, Christine Neou, J Patrick Shelby, Bruz Marzolf, Vesteinn Thorsson, and Timothy Galitski. Prediction of phenotype and gene expression for combinations of mutations. *Molecular systems biology*, 3:96, 2007.

[4] O Alter, P O Brown, and D Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 97(18):10101–10106, August 2000.

[5] Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5):561–566, May 2005.

[6] Ping Ye, Brian D Peyser, Xuewen Pan, Jef D Boeke, Forrest A Spencer, and Joel S Bader. Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular systems biology*, 1:2005.0026, 2005.

[7] Daniel Segrè, Alexander Deluna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nature genetics*, 37(1):77–83, January 2005.

[8] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40, 2002.

[9] I Ljubić, R Weiskircher, U Pferschy, and G Klau. Solving the prize-collecting Steiner tree problem to optimality. *Proceedings of ALENEX*, 2005.

[10] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–31, July 2008.

[11] M Bailly-Bechet, C Borgs, A Braunstein, J Chayes, A Dagkessamanskaia, J-M François, and R Zecchina. Finding undetected protein associations in cell signaling

by belief propagation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(2):882–887, January 2011.

[12] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. *The Nineteenth International Conference on Machine Learning (ICML-2002)*, 2002.

[13] S Brin and L Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks And Isdn Systems*, 30(1-7):107–117, 1998.

[14] Jason Weston, Andre Elisseeff, Dengyong Zhou, Christina S Leslie, and William Stafford Noble. Protein ranking: from local to global structure in the protein similarity network. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6559–6563, April 2004.

[15] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D Boeke, and Joel S Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome research*, 18(12):1991–2004, December 2008.

[16] Alex Pothen, Horst D. Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452, July 1990.

[17] U Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[18] M E J Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, June 2006.

[19] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, 2007.

[20] Aaron Clauset, Cristopher Moore, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[21] Yongjin Park and Joel S Bader. Resolving the structure of interactomes with hierarchical agglomerative clustering. *BMC Bioinformatics*, 12 Suppl 1:S44, 2011.