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

Trey Ideker
Departments of Medicine and Bioengineering
University of California, San Diego
9500 Gilman Drive, Mail Code 0688
La Jolla, CA 92093-0688

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

January 3-7, 2012

Overview

Biology has become an information science, with an increasing capacity to generate data of great relevance to human disease. An important example is The Cancer Genome Atlas (TCGA) [1], which generates data on well-characterized oncology samples and provides a public portal for linking gene mutation and regulation to cancer therapies and outcomes. These types of well-characterized data sets provide an opportunity for researchers from many fields to contribute new ideas for computational analysis.

One theme represented in the 2013 Proceedings is analysis of such public data sets by algorithms known from computer science but less often applied in computational biology and bioinformatics. Previous types of algorithms have included support vector machines [2], graph diffusion [3, 4, 5], and Steiner trees [6, 7]. Algorithms represented this year include set cover (Przytycka and coworkers), color-coded paths (Kahveci and coworkers), and regularized regression (Gevart and Plevritis).

A second theme is using known biological networks and pathways to organize calculations. Perhaps the most prevalent example is Gene Set Enrichment Analysis (GSEA) [8]. Lussier and co-workers describe extensions of GSEA to data sets from individuals rather than groups, and Ritchie and coworkers use interactions to organize analysis of interaction terms in genome-wide association studies (GWAS).

New algorithms from computer science

Przytycka and coworkers extend a set-cover algorithm from genes [9] to modules. These cover algorithms work on bipartite graphs, here with one set of vertices representing disease cases, a second set of vertices representing features (genes or gene modules), and edges indicating that the gene or module is dysregulated in a specific disease case. The k-cover optimization problem is to identify the smallest number of features so that each case has edges to at least k features. The authors generalize this NP-hard problem by also assigning a cost for each module that is reduced when the genes within the module have concordant expression regulation. A fast, greedy forward selection adds modules incrementally, either from a pre-calculated set or by defining modules on the fly. The method is effective in recovering known subtypes of glioblastoma multiforme. This type of approach, based on support, recalls approaches such as the Apriori algorithm for itemset mining [10] and the Teiresias algorithm for pattern discovery [11].

Kahveci and coworkers investigate an algorithm to identify signaling pathways of defined length. For a pathway desired to have m steps, a possible algorithm explored is to color each vertex one of m colors, and then to search for paths that include one vertex of each color. It remains to be seen whether this method is competitive with other related approaches, such as prize-collecting Steiner trees [7] and flow-based methods [12] that have fast, optimal solvers. The restriction to length m paths is motivated by a requirement that signaling pathways include a membrane-bound receptor, cytoplasmic signaling proteins, and nuclear

transcription factors; constraints based on this biology and directed interactions may also perform better than path length restrictions.

Gevart and Plevritis also describe methods motivated by TCGA data. This approach generally follows successful methods introduced by others that use genetic and epigenetic features (copy number variation, methylation) to suggest driver genes, and then build out downstream pathways using regularized regression [13, 14] or other network-based association tests [15]. While predictions of expression perform better than random for an ovarian cancer data set, the top drivers predicted for a gliobastoma multiforme data set perform no better than a random collection. These results point to the uncertainty of applying established algorithms to new data sets and the importance of randomization tests for unbiased assessment of performance.

Pathways as a guide to analysis

Lussier and coworkers investigate personalized RNA-seq data by generalizing a single-sample method they developed for microarray data [16]. The main idea is to generate pathway scores by comparing expression levels between pathway and non-pathway genes. The authors find that converting raw expression values to ranks improves performance for many tasks. While the method is assessed to be feasible, traditional analysis of sample groups still appears to out-perform single-sample analysis.

Ritchie and coworkers investigate interaction terms in genome-wide association studies. Gene-environment interactions are already addressed by conventional methods, but gene-gene interactions are more challenging for both computational and statistical reasons. Computing all gene-gene interactions, or more accurately SNP-SNP interactions, incurs a large computational cost. Furthermore, the large number of tests requires an interaction term to be large for adequate power. The method proposed by Ritchie and coworkers, and also explored by others previously, is to restrict tests to SNPs to pairs in genes that have prior evidence for participating in a shared biological process or pathway. The threshold for evidence is increased until the candidate pairs are reduced to an acceptably small number, for example equivalent to the number of single-SNP tests. One challenge with including interaction terms is that tests for marginal effects may actually have greater power even when the interaction term is non-zero. For example, dominant and recessive genetic models are equivalent to interaction terms at a single locus, and a one degree-of-freedom test of a linear model for phenotype versus allele dose can have greater power than a two degree-of-freedom test that includes the interaction term. In an application to a cataract phenotype, the authors test 57,376 two-SNP models, requiring a p-value of 8.7×10^{-7} for genome-wide significance. The best p-value is 3.4×10^{-6}, however, typical of other searches for that have failed to identify interactions with statistical significance. While it may be feasible to identify interaction terms with greater power from larger population sizes, the lack of significance sets an upper limit on the magnitude of interaction terms and hence a

possible limit on the biological relevance. Furthermore, it remains unclear whether genes identified through interaction terms would have been missed by conventional marginal tests on individual SNPs.

Future perspective

The contributions to this Proceedings consider two types of network models: on the one hand pre-calculated modules or curated pathways, on the other modules or pathways discovered from biological data. An important future direction may be module searches that use high-throughput data but are biased by existing network models. Generative models, such as stochastic block models, may provide an appropriate framework for network analysis biased by empirical knowledge. These models have received increasing attention for both static module discovery and dynamic network evolution [17, 18, 19, 20].

A critical limitation of network biology is the limited amount of high-quality network data. High-throughput protein-protein interaction data sets are available for human [21] but are incomplete [22, 23, 24]. Interactions between transcription factors to regulated genes provide crucial links between protein signaling and gene regulation, but are even less well mapped for human. Experimental progress here could result in dramatic gains for computational methods that already exist but which have been limited by lack of data.

References

[1] International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Thomas J Hudson, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel A Piris, Rajiv Sarin, Aldo Scarpa, Tatsuhiro Shibata, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Daniela S Gerhard, Sean M Grimmond, Ivo Gut, Martin Hirst, Carlos López-Otín, Partha Majumder, Marco Marra, John D McPherson, Hidewaki Nakagawa, Zemin Ning, Xose S

Puente, Yijun Ruan, Tatsuhiro Shibata, Michael R Stratton, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Peter J Campbell, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard, Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Xose S Puente, Victor Quesada, Benjamin J Raphael, Chris Sander, Tatsuhiro Shibata, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David A Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Lincoln D Stein, Roderic Guigó, Tim J Hubbard, Yann Joly, Steven M Jones, Arek Kasprzyk, Mark Lathrop, Nuria López-Bigas, B F Francis Ouellette, Paul T Spellman, Jon W Teague, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, P Andrew Futreal, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, Arek Kasprzyk, Lincoln D Stein, Junjun Zhang, Syed A Haider, Jianxin Wang, Christina K Yung, Anthony Cros, Anthony Cross, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Martin Bobrow, Don R C Chalmers, Karl W Hasel, Yann Joly, Terry S H Kaan, Karen L Kennedy, Bartha M Knoppers, William W Lowrance, Tohru Masui, Pilar Nicolás, Emmanuelle Rial-Sebbag, Laura Lyman Rodriguez, Catherine Vergely, Teruhiko Yoshida, Sean M Grimmond, Andrew V Biankin, David D L Bowtell, Nicole Cloonan, Anna deFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke B Gardiner, Brooke A Gardiner, James G Kench, Aldo Scarpa, Robert L Sutherland, Margaret A Tempero, Nicola J Waddell, Peter J Wilson, John D McPherson, Steve Gallinger, Ming-Sound Tsao, Patricia A Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Lynda Chin, Ronald A DePinho, Sarah Thayer, Lakshmi Muthuswamy, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Guangyu Zhou, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Huanming Yang, Mark Lathrop, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, and Alvis... Brazma. International network of cancer genome projects. *Nature*, 464(7291):993–998, April 2010.

[2] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines.* And Other Kernel-Based Learning Methods. Cambridge Univ Pr, March 2000.

[3] 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.

[4] 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.

[5] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 55–66, 2012.

[6] I Ljubic, R Weiskircher, U Pferschy, GW Klau, P Mutzel, and M Fischetti. An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming*, 105(2-3):427–449, 2006.

[7] 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.

[8] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[9] Yoo-Ah Kim, Stefan Wuchty, and Teresa M Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Computational Biology*, 7(3):e1001095, March 2011.

[10] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc, September 1994.

[11] I Rigoutsos and A Floratos. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[12] Esti Yeger-Lotem, Laura Riva, Linhui Julie Su, Aaron D Gitler, Anil G Cashikar, Oliver D King, Pavan K Auluck, Melissa L Geddie, Julie S Valastyan, David R Karger, Susan Lindquist, and Ernest Fraenkel. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature genetics*, 41(3):316–323, March 2009.

[13] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe'er. An Integrated Approach to Uncover Drivers of Cancer. *Cell*, 143(6):1005–1017, December 2010.

[14] Su-In Lee, Aimée M Dudley, David Drubin, Pamela A Silver, Nevan J Krogan, Dana Pe'er, and Daphne Koller. Learning a Prior on Regulatory Potential from eQTL Data. *PLoS Genetics*, 5(1):e1000358, January 2009.

[15] Kartik M Mani, Celine Lefebvre, Kai Wang, Wei Keat Lim, Katia Basso, Riccardo Dalla Favera, and Andrea Califano. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular systems biology*, 4:169, 2008.

[16] Xinan Yang, Kelly Regan, Yong Huang, Qingbei Zhang, Jianrong Li, Tanguy Y Seiwert, Ezra E W Cohen, H Rosie Xing, and Yves A Lussier. Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Computational Biology*, 8(1):e1002350, January 2012.

[17] 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.

[18] J Hofman and C Wiggins. Bayesian approach to network modularity. *Physical Review Letters*, 2008.

[19] Joel S Bader and Yongjin Park. How networks change with time. *Bioinformatics*, 28(12):i40–i48, January 2012.

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

[21] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[22] G Traver Hart, Arun K Ramani, and Edward M Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006.

[23] Hailiang Huang, Bruno M Jedynak, and Joel S Bader. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Computational Biology*, 3(11):e214, November 2007.

[24] Hailiang Huang and Joel S Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3):372–378, February 2009.