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 glioblastoma 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 $8.7 \times 10^{-7}$ for genome-wide significance. The best p-value is $3.4 \times 10^{-6}$, 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


PSB 2013 Tutorial: Aberrant Pathway


Annotated Bibliography: Identification of Aberrant Pathway and Network Activity from High-Throughput Data

Background Reading

Foundations


Singular value decomposition explains global gene expression profiles through a concise set of factors.


The improvements in computational power allow the application of Bayesian approaches that provide much greater insight into functional genomics experiments.


The Google PageRank algorithm uses graph diffusion as the basis of web search.


The issue of high-dimensional spaces is addressed from a biological and statistical viewpoint, introducing concepts of the 'curse of dimensionality' and issues with distance metrics to a non-specialist audience.

Describes a hierarchical decomposition of large-scale networks to shed light on observed network properties in many real-world networks.


Spectral clustering is a global method for decomposing a network into subnetworks, but has not worked well for biological and social networks.


The methods used to identify drug targets in signaling pathways underlying disease are presented together with an analysis of outstanding issues for targeted drug development.


Diffusion processes lead to useful kernels defined on graphs.


This work reviews matrix factorization methods in terms of their ability to recover biologically meaningful results and shows that Bayesian approaches are the most effective.


A view of intracellular networks in terms of control networks is presented together with a number of examples of regulatory networks in biological systems.

A mechanistic view of signaling pathways is provided together with a differential equation-based modeling approach. The relationship of the biological system to electronic systems is explored.


L1 regularization for regression is useful for both model selection and parameter estimation.


Cell motility remains one of the most thoroughly studied aspects of physiological response to signals. This review looks at our present understanding of the process and at computational models of cell migration.


Methods based on modularity are closely related to spectral clustering.


A summary of recent developments in knowledge-based and Bayesian analysis methods for high-throughput data is presented.


An introduction to important concepts in cell signaling, including components and effectors of pathways, is presented.
Network Analysis


Belief propagation is used to solve the NP-hard Prize-Collecting Steiner Tree problem, with applications to identifying novel components of cellular signaling pathways.


Differential networks chart a new type of genetic landscape that is invaluable for mapping cellular responses to stimuli.


Describe the largest synthetic genetic network to date and show how to identify pathways using congruency analysis.

Methods for interpreting high-throughput data in terms of coordinated changes and causal relationships are presented.


Functionally related genes share genetic interaction partners, but do not necessarily have genetic interactions with each other.


Genetic interactions in metabolism are dominated by interactions between metabolic pathways.


Describes a probabilistic method for integrating diverse information to predict gene-gene functional couplings.


The active subnetwork problem is described as identifying all the components of a subnetwork in which only a subset are experimentally detected.

May 2005.

Genetic interactions are explained by between-pathway models and within-pathway models.


Describes a method for connecting perturbations from the signaling to the transcriptional levels.


This study uses functional networks to identify novel modulators of a pathway. If genes in a pathway are well-connected in an integrated gene-gene network the results show that new genes that are also well-connected also are more likely to play a role in the pathway. The authors show results in both yeast and worm and thus suggest that the same approach may work for human disease pathways.


Creation and analysis of a database of global gene expression profiles (which we call the ‘stem cell matrix’) that enables the classification of cultured human stem cells in the context of a wide variety of pluripotent, multipotent and differentiated cell types.


Graph diffusion is used to predict new genetic interactions and protein interactions.
Analysis of D. melanogaster eye development is used to introduce concepts of nonlinear modeling in biological systems driven by cell signaling processes.

Network model of co-evolving mutations under evolutionary selective pressure in antibiotic resistance enzyme enables prediction of evolutionary trajectories.

Applications to Human Disease

The first effective therapy targeted at disrupting an aberrant signaling pathway provided a revolutionary breakthrough in cancer treatment and spurred the development of numerous drugs targeted at signaling proteins.

Rare functional variants, weak effects, genetic interactions, and other explanations are proposed for the difficulty in identify-
tifying all the genetic factors underlying human phenotypes.


Describes a probabilistic method for connecting DNA variation to gene expression with application to obesity.


Describes an approach to integrate expression and protein-protein interaction to identify disease-related gene modules.


DEGAS (DysrEgulated Gene set Analysis via Subnetworks) is a method for identifying connected gene subnetworks significantly enriched for genes that are dysregulated in specimens of a disease.


Combine mutations and protein-protein interaction networks to identify driver pathways in cancer.


A probabilistic graphical model (PGM) method for inferring patient-specific genetic activities, incorporating curated path-
way interactions among genes. Each gene is modeled by a factor graph as a set of interconnected variables encoding the expression and known activity of a gene and its products, allowing the incorporation of many types of omic data as evidence. The method predicts the degree to which a pathways activities (e.g. internal gene states, interactions or high-level outputs) are altered in the patient using probabilistic inference.


Describes a method to interpret GWAS data with genetic pathways to identify causal genes in Type 2 Diabetes.


Phenotypes are predicted for combinations of mutations.


The group builds disease classifiers from a large collection of publicly available expression data and uses a Bayesian Network for error-correction.


A model-free approach to building subnetworks of gene relationships putatively critical to cancer cell survival.
Advances in Algorithms


Spectral methods based on eigenvectors provide improved solutions for graph partitioning.


An efficient optimization algorithm solves the NP-hard Prize-Collecting Steiner Tree problem, with broad applications.


Greedy optimizations extend hierarchical stochastic block models to genome-sized networks.


Graph diffusion is used to improve protein similarity searching.


A solution to the NP-hard Prize-Collecting Steiner Tree problem provides optimal active subnetworks.

Many popular methods for graph segmentation fail when a graph has both small and large groups or clusters.

Best Practices


This work summarizes the results of numerous studies that demonstrated proper handling of omics data to insure reproducibility.

Databases

Repositories


The standard European repository for array-based data.


A repository for proteomics data linked to the Tranche data system.


The standard US repository for array-based data.


The Cancer Genome Atlas: repository for data generated in the NCI funded project.
Genomes and Genes


This web-enabled system allows visualization of numerous types of data against the reference genome.


The standard repository maintained by the European Bioinformatics Institute.


The standard repository maintained by the National Library of Medicine.


A tool for searching multiple databases and for visualizing protein structures.

Pathways


A set of curated pathways important to human health.


The standard repository for curated metabolic and signaling pathways in many organisms.


A set of curated pathways based on known protein interactions.


The Pathway Interaction Database hosts a collection of hu-
man signaling pathways based on biomolecular interactions and cellular processes


Pathway commons hosts pathways collected from public pathway databases including BioGrid, Cancer Cell Map, HPRD, HumanCyc, IMID, IntAct, MINT, NCI-PID, and Reactome


WikiPathways is an open, public platform dedicated to the curation of biological pathways by and for the scientific community.

Organism Specific


The D. melanogaster genome and linked data.


The S. cerevisiae genome and linked data.


The C. elegans genome and linked data.


The D. rerio genome and linked data.
Resouces for Network Analysis


VisANT is a web-based software framework for visualizing and analyzing many types of networks of biological interactions and associations.


A software platform for visualization of complex networks


An online repository of genetic and protein interactions in multiple model organisms


A web interface to a broad landscape of data on protein-protein interactions (PPI) consolidated from major public databases including BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPact, MPPI, and OPHID

Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data.


Uses functional association data, including protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity, to find genes that interact