BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 4. 406–419
Advance Access published on 27 January 2012
doi:10.1093/bib/bbs001

Identification of aberrant pathways and network activities from high-throughput data

Jinlian Wang, Yuji Zhang, Catalin Marian and Habtom W. Ressom

Submitted: 6th September 2011; Received (in revised form): 3rd January 2012

Abstract

Many complex diseases such as cancer are associated with changes in biological pathways and molecular networks rather than being caused by single gene alterations. A major challenge in the diagnosis and treatment of such diseases is to identify characteristic aberrancies in the biological pathways and molecular network activities and elucidate their relationship to the disease. This review presents recent progress in using high-throughput biological assays to decipher aberrant pathways and network activities. In particular, this review provides specific examples in which high-throughput data have been applied to identify relationships between diseases and aberrant pathways and network activities. The achievements in this field have been remarkable, but many challenges have yet to be addressed.

Keywords: *pathways; biological networks; biomarker discovery; omics studies; systems biology*

INTRODUCTION

Historically, researchers have been able to demonstrate the relationship between diseases and changes in a single or a few genes using various experimental systems. However, increasing evidence indicates that complex diseases result from multiple genetic aberrancies in biological pathways and networks rather than from single gene changes [1–3]. Biological pathways represent series of actions among molecules that lead to a certain product or a specific change in a cell [4]. These pathways may be involved in metabolism, regulation of genes, signal transmissions and other areas of cellular activities. Extensively studied pathways include signaling, gene regulatory and metabolic pathways. The differences between these three pathways reside in their functioning cellular location, compartments and molecular level. As a result, the functional annotation information and databases attributed to each pathway as well as the analytic methods also differ. Pathways do not function alone. Instead, they occur in extremely complex biological networks [5, 6], which are collections of pathways and interactions between biological entities. Typical biological networks include cell signaling networks, protein–protein interactions and metabolic networks [7]. Increasing evidence indicates that dysregulation of biological pathways and network activities is a hallmark of complex diseases, such as cancer [8, 9]. Therefore, identifying aberrancies involving many genes in specific biological networks may lead to the effective diagnosis and treatment of various diseases.

Corresponding author. Habtom W. Ressom, Lombardi Comprehensive Cancer Center, Georgetown University, 4000 Reservoir Road NW, Washington, DC 20057, USA. E-mail: hwr@georgetown.edu

Jinlian Wang is a postdoctoral fellow at Lombardi Comprehensive Cancer Center, Georgetown University. Her research interest focuses on cancer biomarker discovery and integration of multiple omics data for network and pathway analysis.

Yuji Zhang is an Assistant Professor at the Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo College of Medicine. Her research interest includes the integrative analysis and visualization of high-throughput omics data, systems biology, and analysis of massively parallel sequencing data.

Catalin Marian is an Assistant Professor in the Department of Internal Medicine and the Comprehensive Cancer Center at The Ohio State University. His research interests include cancer biomarker discovery and cancer molecular epidemiology.

Habtom Ressom is an Associate Professor and Director of Genomics and Epigenomics Shared Resource at Lombardi Comprehensive Cancer Center, Georgetown University. He is interested in cancer biomarker discovery and systems biology research by analysis of omics data.

The key question is how to identify aberrancies in biological pathways and networks. Because tens or hundreds of genes are often involved, conventional experimental systems developed for identifying aberrant gene(s) are inadequate for deciphering aberrancies in pathways and network activities. Since high-throughput technologies were first successfully used in DNA sequencing [10], they have quickly gained popularity in the fields of messenger RNA (mRNA) expression profiling, DNA methylation analysis, gene mutation screening, SNP detection, copy number variation (CNV) analysis, microRNA (miRNA) and non-coding RNA (ncRNA) profiling [11]. Herein, we focus on reviewing research contributions from the most recent publications in which high-throughput data (HTD) and analytic methods have been applied to identify relationships between disease and aberrancies in pathways and network activities. In the following sections, we first provide a brief overview of HTD and analytic methods, followed by a review of computational approaches useful for identifying aberrant pathways and networks by integrating multiple types of HTD. Then, we present applications of these approaches to the identification of functional modules and the discovery of biomarkers. Finally, we provide concluding remarks, future perspectives, and guidance to potential users of HTD about identifying aberrant pathways and network activities.

HTD

The emergence of gene expression microarrays in the mid-1990s dramatically changed traditional molecular biological and biochemical approaches, which usually only studied one or a few genes in a tightly controlled experimental system. Various omics studies use HTD generated from technologies that enable the simultaneous detection of a large number of alterations in molecular components to investigate the correlations and dependencies between molecular components. Omics studies not only facilitate the understanding of biological entities at a molecular level but also offer a different perspective on the processes underlying disease initiation and progression. These studies also focus on ways of predicting, preventing or treating disease more accurately and efficiently. For a detailed review about omics studies and HTD, see Schneider *et al.* [11]. Since multiple types of HTD analyses have revealed that our current understanding of molecular biology and chemical biology remains incomplete and fragmentary [12], a number of open-access databases could be a good complement for identifying aberrant pathways and network activities from HTD [13–15]. In addition, numerous tools have been developed for visually exploring and analyzing biological pathways and networks including Cytoscape [16], VisANT [17], GeneGO (http://www.genego.com/), Ingenuity (http://www.ingenuity.com/), and Pathway Studio [18]. Table 1 summarizes high-throughput omics techniques, their applications, related sources and some representative references.

HTD ANALYSIS

Analyses of HTD have allowed us to investigate aberrant pathways and networks at the systems level. Because biological data and related annotation information about genes, proteins, pathways and networks have been accumulated by different research groups, the challenge is to discover ways to integrate these different omics data sources to yield new information about pathways, networks and diseases. In the following section, we review recent advances in the identification of the activities of pathways and networks as well as their aberrancies by integrating heterogeneous sources of HTD. Figure 1 depicts the workflow of a typical study involving high-throughput omics data to identify aberrant pathways and network activities. As shown in the figure, data analysis involves preprocessing steps, such as noise filtering, background subtraction, and normalization, which may affect the performance of subsequent analyses, for example, feature selection (e.g. selection of differentially expressed genes, differentially abundant ions, etc.) and pathway analysis. Two sources of variability tend to impact the analysis of data from high-throughput technologies: (i) Errors which have the same impact on all measurements and thus systematically bias all the data. Since such errors are systematic, they can be addressed by normalizing the data [31]; (ii) Noise, which is stochastic in nature. Such random noise is difficult to remove by normalization methods, but its impact can be reduced through noise-filtering methods and/or by using adequate number of replicates. Data preprocessing methods are useful for minimizing such measurement variabilities and noise. The appropriate method for a specific dataset depends on the specific technology and experimental design used. For example, image or signal processing methods may be necessary to extract features from raw images/data.

Table 1: Omics high-throughput techniques, applications, and useful resources

Omics Studies	Technology	Application	Related sources
Genomics/ Epigenomics	aCGH SNP genotyping arrays Next-gen sequencing DNA methylation array ChIP-chip arrays ChIP-seq small RNA-seq	Determination of variations in the DNA sequence for disease diagnosis, prediction of the risk of future disease in healthy individuals, and identification of unaffected individuals who carry one copy of a gene for a disease that requires two copies of the disease to manifest; methylation profiling for the identification of aberrantly methylated genes in cancer; chromatin Immunoprecipitation arrays or sequencing for the study of DNA-protein interactions, histone modifications; identification and characterization of non-coding RNA molecules. [19, 20, 21–23]	EBI genomes: http://www.ebi.ac.uk/genomes http://www.ensembl.org/index.html Mouse genome: http://www.informatics.jax.org Rat Genome: http://rgd.mcw.edu Saccharomyces genome: http://www.yeastgenome.org AceDB genome: http://www.acedb.org/introduction.shtml HIV Sequence: http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html Gene ontology: http://www.geneontology.org Human mitochondrial genome database: http://www.mitomap.org
Transcriptomics	mRNA arrays miRNA arrays ncRNA arrays RNA-Seq	Gene expression profiling for the molecular classification of tumors with impact on treatment and clinical management, biomarker discovery, and identification of new RNA molecules such as ncRNAs. [24–26]	ArrayExpress: http://www.ebi.ac.uk/arrayexpress/ GEO: http://www.ncbi.nlm.nih.gov/geo miRBASE: http://www.mirbase.org Comparative RNA: http://www.rna.ccbb.utexas.edu Noncoding RNA database: http://www.ncrna.org/frnadb Comprehensive Ribosomal RNA database: http://www.arb-silva.de/ Genomic tRNA database: http://gtrnadb.ucsc.edu miRNA sequences: http://www.ebi.ac.uk/enright-srv/MapMi/
Proteomics	Mass spectrometry Chromatography Protein microarrays Interactomics	Large-scale quantification and characterization of peptides or proteins. Identification of post-translational modifications. Study of protein–protein interactions in a high-throughput manner. [27, 28]	UniProt: http://www.uniprot.org PIR: http://pir.georgetown.edu GO: http://www.ebi.ac.uk/GOA/index.html pfam: http://pfam.sanger.ac.uk Protein Data Bank in Europe: http://www.ebi.ac.uk/pdbe PeptideAtlas: http://www.peptideatlas.org
Metabolomics	Mass spectrometry Chromatography Nuclear magnetic resonance Crystallography	Quantification and characterization of metabolites and other small molecules. Toxicity assessment/toxicology. Prediction of the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. [29, 30]	MMCD: http://mmcd.nmrfam.wisc.edu/ HMDB:http://www.hmdb.ca/ Metlin: http://metlin.scripps.edu/ LipidMaps: http://www.lipidmaps.org/search/search.html GabiPD:http://www.gabipd.org/

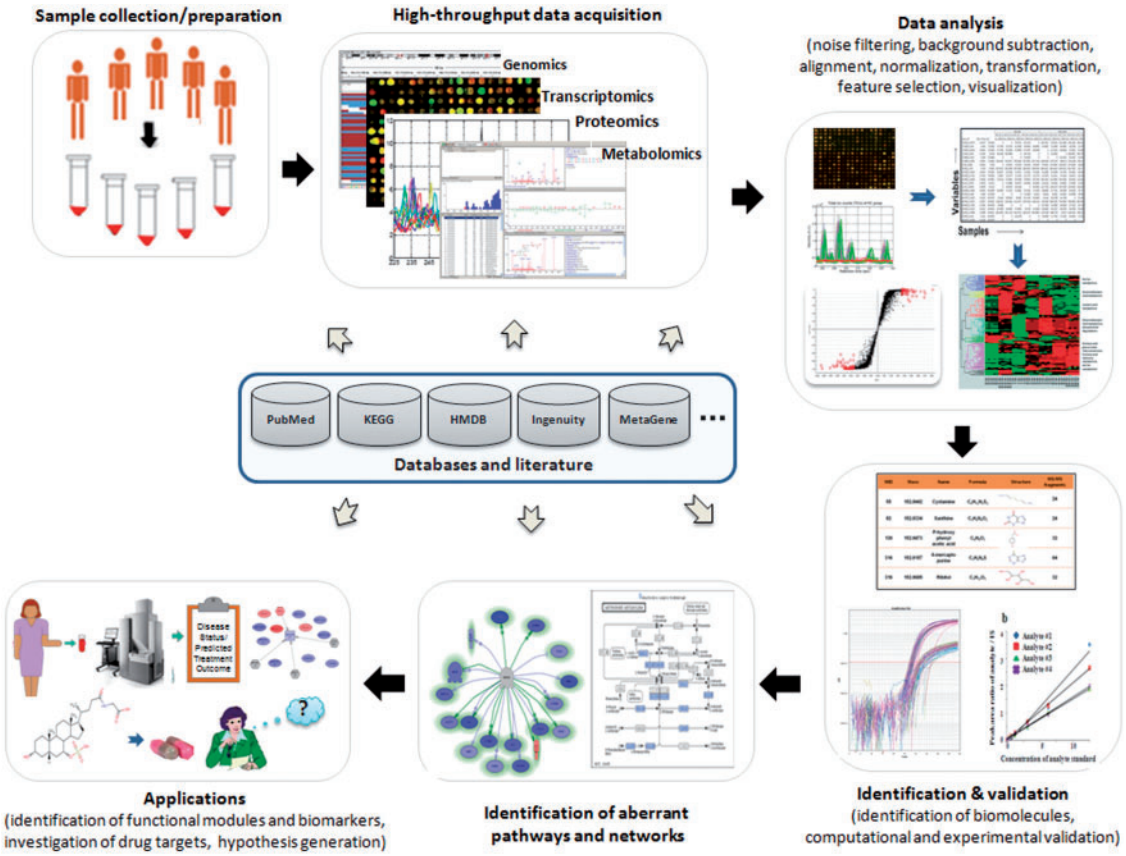

Figure 1: A workflow for identification of aberrant pathways and network activities from high throughput data.

Background subtraction methods allow us to estimate and remove background signals from measurement signals. Normalization methods enable us to reduce systematic bias to enable comparisons between data from various experiments. Transformations such as Z-score and log transforms allow us to modify the data distribution so that it will be better suited for feature selection by statistical or machine learning methods.

An analysis of sensitivity and specificity is important in evaluating the performance of candidate biomarkers or disease outcome predictors identified by HTD analysis. The receiver operating characteristic (ROC) is commonly used to assess the performance of candidate biomarkers [32]. An important weakness of many machine learning methods is that they are not based on a probabilistic model. Therefore, probability levels and confidence intervals cannot be readily associated with their predictions. The confidence that an analyst can have in the accuracy of the prediction results is based purely on the predictor's historical accuracy—how well it has predicted the desired response in other, similar situations. Thus, after learning is completed, the prediction model must be evaluated for its performance through a previously unseen testing data set (also known as a blind validation set). The purpose of this testing is to demonstrate the adequacy or to detect the inadequacy of the selected features (e.g. biomarkers) or of the prediction model. An inadequate performance may be attributable to insufficient or redundant features, inappropriate selection of the model structure, too few or too many model parameters, insufficient training, overtraining, error in the program code or complexity of the underlying system, such as the presence of highly nonlinear relationships, noise and systematic bias. The goal of evaluating a predictor is to ensure that it is able to serve as a general model whose input–output relationships (derived from the training set) apply equally well to a new dataset derived from the same problem, but which was not included in the training set. Thus, ensuring the generalization of the relationships learned on the training set to the new dataset is necessary. Various methods have been used to test the generalization capability of a prediction model. These include the *k*-fold cross-validation, bootstrapping and hold-out methods [33, 34, 27]. In addition to evaluating the

performance of candidate biomarkers or disease outcome predictors by computational methods, examining their performance in samples from a large population and using technologies other than those used during the discovery phase may be necessary. For example, biomarkers discovered by microarray gene expression profiling or by mass spectrometric methods could be validated using real-time polymerase chain reaction (RT-PCR), Western blot, enzyme-linked immunosorbent assay (ELISA), etc. These experimental validation methods can be applied to functional modules as well as to pathways and their cross-talk [35].

Various HTD-based computational methods can be used for modeling aberrant pathways and network activities [27] and for identifying functional modularity [36, 37], as well as for applications that include biomarkers discovery [33, 38, 39]. These computational techniques include statistical methods [40, 41], graph theory/models [42, 43], probabilistic graphical models [44], rule-based inferences approaches [45], logic-based models [45, 46], machine learning methods [37, 47], knowledge-based models including text mining and function annotation [37, 48, 49] and mechanistic differential equation-based models [6, 50] that capture temporal and spatial dynamics at the level of individual reactions. The advantages of these methods are their applicability to situations in which mechanistic information is incomplete or fragmentary. They provide useful insights into the links between disease and pathways and network activities. The choice of an appropriate modeling approach relies on the question being posed, the quality and type of experimental data, and prior knowledge about the pathway/network. Table 2 summarizes computational models and data sources used for analyzing pathways and networks. A detailed review of gene regulatory network modeling approaches can be found in Hecker *et al*. [51]. For a review of metabolic pathways analyses, see Trinh *et al*. [52]. Sardiu *et al*. [53] provide a review about building protein interaction networks.

INTEGRATION OF OMICS DATA FOR IDENTIFICATION OF ABERRANT PATHWAYS AND NETWORK ACTIVITIES

The advantages of integrating omics data to identify aberrant pathways and network activities clearly lie in that we can address systems-specific questions in biology, model systems-wide behavior, estimate cellular structures, generate new insights into basic research, develop new drugs and personalize genomic medicine [71]. Although many researchers have made great strides toward increasing the power of integrative analysis over that of a single measurement platform in order to identify changes in pathway and network activities [34, 27, 54], the need for more effective methods of integrating omics data is still urgent. These methods should be capable of accommodating various sources of binary, categorical and continuous data as well as being suitable for handling missing data, high error rates and systematic biases in the data. In addition, they should be able to deal with reducing the dimensionality of omics data for effective interpretation and visualization of the analysis results. Figure 2 depicts a schematic outlining of the steps involved in integrating omics data sets for the identification of aberrant pathways and network activities. As shown in Figure 2, the identification approach is generally comprised of three steps: (i) identifying the network structure from a high-throughput omics dataset that focuses on interacting transcriptomics, genomics, and protein–DNA interactome information coupled with the known pathway information extracted from public databases such as KEGG [13], Reactome [4] or PID [15]; (ii) decomposing or clustering the network into sub networks, pathways or modules; and (iii) developing cellular models to simulate and predict network activities that give rise to cellular phenotypes. The simplest integration approach is to treat each data set independently and combine the results using union, intersection or majority vote rules. However, no clear method exists for weighting the confidence of different assays. Reviews on the integration of omics data can be found in works by Tsiliki and Kossida [72] and Jacobs and Wang [73]. Methods for omics data integration include network and graph models [54], Bayesian models [58], kernel models and support vector machines (SVM) [74, 75]. In particular, Bayesian methods are commonly used for omics data integration because they allow the combining of highly dissimilar types of heterogeneous omics data and various online databases. They can also express causal relationships and glean from incomplete datasets while avoiding data over fitting. Integration of omics data has led to identifying biomarkers that are associated with cancer states [33, 19], to elucidating aberrant pathways [54, 20],

Table 2: Computational methods and data resources for biological pathway and network analysis

Pathways/Networks	Definition	Methods	Data sources
Signaling pathways	Signaling pathways are chemical reactions in a cell from a stimulus to the response.	Bayesian based model [34][a] Graph theory [54][a] Gene set enrichment analysis [55][a] Fuzzy logic [46] Rule-based inference [45][a] Perturbation measurements [50,[a] 56]	Reactome [4] TRANSPATH NetworKIN PID UCSD [57]
Gene regulatory pathways/ networks	Gene regulatory pathways involve the series of actions by which genes can be activated or repressed. Often a gene is regulated by another gene via a transcription factor. A gene regulatory network (GRN) directs the level of expression for each gene in the cell by controlling whether and how often that gene will be transcribed into RNA.	Gene set enrichment analysis [55][a] Dynamic Bayesian network [58] Subgraphs network [59] Boolean networks [45][a] Probabilistic graphical model [43] PARADIGM [20][a] Ordinary differential equations (ODEs) [58] Bayesian network model [60][a] Unbiased perturbation based model [50][a] Machine learning method. [47][a] Graph based model [54][a]	KEGG [13] TRANSCompel PID [15] IntAct[61] Reactome [4]
Metabolic pathways/ networks	Metabolic pathways are series of enzymatic reactions that convert an initial substrate to the final end-products and are connected by their intermediates. A metabolic network is the complete set of metabolic reactions of a particular cell or organism. These networks are comprised of the sequence of reactions catalyzed by enzymes as well as the regulatory interactions that guide them.	Graph theory [42] Petri-nets [62] Extreme pathway analysis model [63] Stochastic model [64][a] Kinetic based model [6] (pathway cross talk)[a] Kinetically derived flux estimations model. [65] [29][a] Statistics based parameter estimation with kinetic approach [66][a] Petri Net model [67][a] Gaussian graphical modeling [44]	KEGG [13] Reactome [4] BioCarta MetaCyc HMDB EcoCyc [68]
Protein–protein interaction network	Protein–protein interaction networks describe local subsets of protein interactions such as the relationship of proteins within a protein complex or under certain biological conditions. Eventually the network depicts all global protein interactions at the organismal scale.	Graph model [27][a] Generative probabilistic models [69] Text mining model [37][a] Machine learning model [70][a]	PID [15] IntAct [61] Reactome [4] DIP BioCarta NetworKIN HPID

[a]More than two high throughput data types applied.

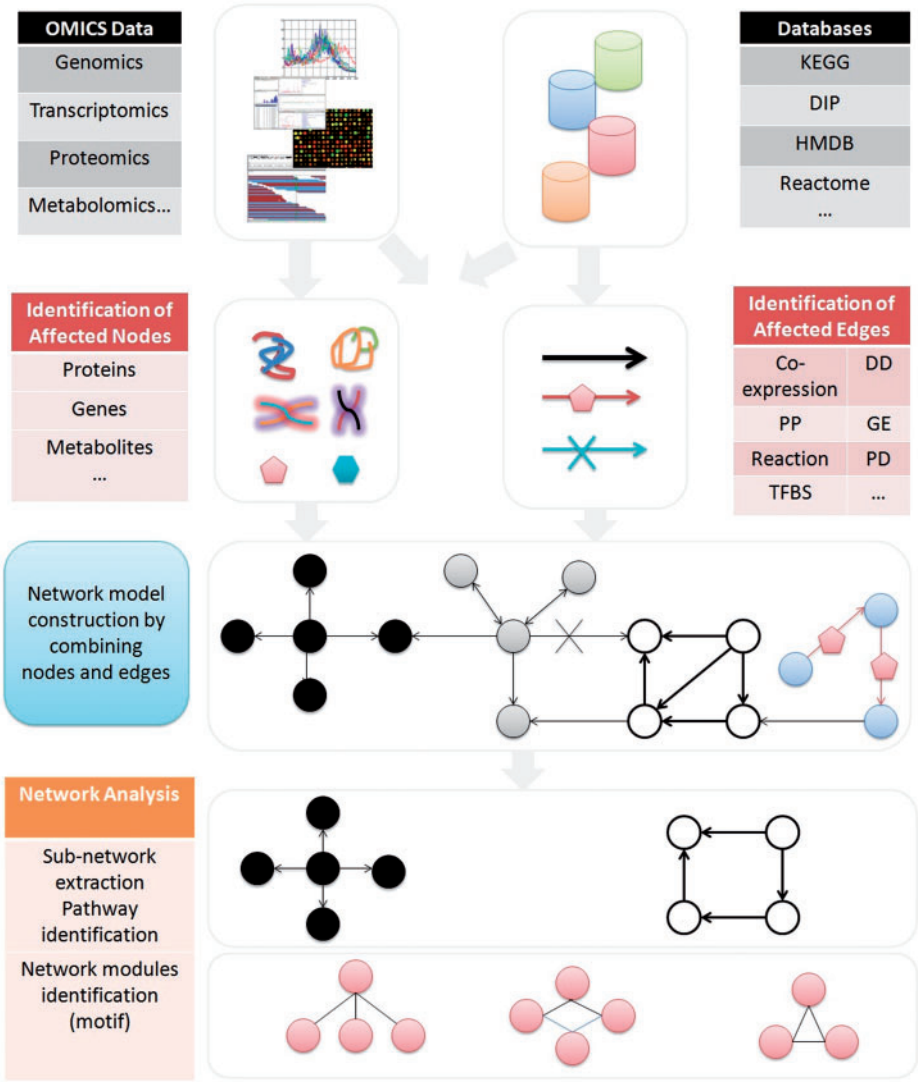

Figure 2: A schematic outlining the steps involved in omics data integration for identification of aberrant pathways and network activities. PP: protein-protein (interaction); PD: Protein DNA (interaction); DD: domain-domain (interaction); GE: gene expression; TFBS: transcription factor binding site.

and to constructing models of dysregulated networks [27, 65]. However, integration of omics data continues to be challenging due to incomplete data, high noise, lack of uniform and standardized data representations, data quality issues, and lab-to-lab variations. The problem is how to interpret and integrate omics datasets in ways that allow researchers and practitioners to understand the principles underlying the regulation of genes, metabolites and proteins. Another problem is comprehending how their combined interactions are associated with variations in phenotype. In the remaining sections, we review recent significant contributions to the identification and application of aberrant pathways and network activities using multiple types of HTD.

IDENTIFICATION OF DYSREGULATED PATHWAYS AND BIOLOGICAL NETWORK ACTIVITIES

One of the most important questions that has been addressed in recent years is how genetic differences between individuals lead to differences in disease pathways and thus to differences in biological phenotypes [33, 20]. Data-driven computational models that have been especially useful in this area include PARADIGM [20] and ResponseNet [54] as well as a few others that have been published recently [2, 8, 47, 50]. In the following, we highlight examples of previously published studies, in which various HTD have been used for the identification of aberrant

pathways and networks. From these examples, we observe that several approaches to integrating omics data from different sources have proved to be promising and biologically meaningful for studying aberrant pathways and network activities. However, the integrated reconstruction and analysis of aberrant biological networks are hampered by difficulties such as insufficient experimental data, annotation differences, multiple interpretations, and integrating heterogeneous and diverse data [76].

Signaling pathways

Recently, Vaske *et al.* [20] developed a probabilistic graphical-based model known as PARADIGM to identify patient-specific pathway activities in The Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM). They integrated four different omics data types including copy number alteration (CNA), mRNA, protein levels and activity of the protein in a single patient to infer the activities of genes, products and abstract process inputs and output for a single PID pathway [15]. A matrix of integrated pathway activities for each patient was used to identify associations with clinical outcomes. Clustering the GBM patients based on PARADIGM revealed patient subtypes correlated with different survival profiles, whereas using single expression data or CNA data did not. In addition, PARADIGM inferred significantly altered gene activities in tumor samples from both GBM and breast cancer with fewer false positives compared with other approaches. PARADIGM has also addressed questions about how the heterogeneity and complexity of pathway changes could affect different individuals [20]. Recently, Masica *et al.* [77] extended this idea by employing an exhaustive model-free pipeline that suggested ways that mutated genes participate in the progression of GBM cancer and, subsequently, allowed for the integration of gene expression and gene sequencing data as well as literature mining to delineate the method. Although the two methods were similar, Masica *et al.*'s was advantageous in that PARADIGM relies on pre-existing knowledge about gene annotations, protein interactions and curated pathways, whereas Masica *et al.*'s did not.

Another example of integrating genomics, transcriptomics and protein–protein interactomes to infer changes in the context of the relationship between signaling protein interactions and transcriptional regulation was reported by Yeger-Lotem *et al.* [54] using ResponseNet. This method treats genetic library screening results and transcriptional changes in the context of the relationship between signaling protein interactions and transcriptional regulation. By integrating gene expression, genetic library and ChIP-chip data into a graph, ResponseNet has successfully identified the pathways involved with α-synuclein toxicity as well as the genes differentially regulated by these pathways. However, this approach relies on downstream transcriptional changes to drive discovery and thus may miss important changes in protein interactions that are not involved with transcriptional change.

Ochs *et al.* [34] proposed a differential expression for a signaling determination (DESIDE) model based on a Bayesian decomposition algorithm. Applying DESIDE to the analysis of microarray data derived from gastrointestinal stromal tumors (GIST) time series cell line and GIST patients, they identified treatment-induced aberrancies in the KIT signaling pathway activities that drive changes in the activity levels of transcriptional regulators [34].

Other interesting contributions include the application of logic models to the identification of biological pathways from HTD [45, 46]. By integrating mutational, transcriptional and proteomic data for 30 breast cancer cell lines, Heiser *et al.* [45] generated a signaling network model that identified the ErbB/MAPK signaling pathway and network modules active in specific subsets of cell lines. They built a unique signaling pathway model for each cell line. The rules of the models represented biochemical reactions, and each model had an initial state that included all proteins present in a particular cell line. Signaling was represented by rule sets based on experimentally derived protein–protein interactions, which determined a sequence of model states. Heiser *et al.* [45] simplified their assumptions by specifically discretizing both the data, i.e. each protein component was either 'present' or 'absent' in each state, and the rules that defined the signaling between these active states. However, the simplicity of this method, especially discritizing the data, may ignore important information about proteins in a system.

Cross-talk pathways

Progress has been made in the quantified analysis of cross-talk pathways. Diseases such as arthritis and diabetes are often seen as resulting from the dysregulation of multi-pathways, dynamic cross-talk and networks [78, 79]. For instance, Wang *et al.* [6]

used Monte Carlo methods and data-driven kinetic modeling to quantify the magnitudes of cross-talk and negative feedback interactions in a signaling network. They found possible cross-talk between the PI3K and Erk pathways: Ras and PI3K activate Erk signaling independently and PI3K enhances Erk activation at points both upstream and downstream of Ras. These results show that integration of kinetic modeling with HTD provides a systematic and quantitative method for understanding pathway cross-talk. Extending this model to the analysis of large scale datasets seems promising.

Regulatory network activities

Recent advances in identifying biological network activities have enabled researchers to obtain a snapshot of the biological process in cells. For instance, Huang *et al*. [27] employed a prize-collecting Steiner tree (PCST) algorithm to construct a regulatory network that explains both known and previously hidden components of yeast pheromone response pathways. This was accomplished by integrating experimental protein–protein interactions and transcriptional data with protein interaction databases. The Steiner tree was successful in balancing the introduction of false positive interactions from experimental data with the loss of key interactions. However, the Steiner tree, which aims to find the tree with the minimum total length, cannot handle large scale networks and requires high quality data sets for seamless integration of multiple data types. ResponsNet has successfully revealed most of the expected pathways in yeast and detected changes that had not been detected by mass spectrometry in pheromone-induced MAKP pathway components such as CPA1, STE11 and BEM1.

Metabolic network activities

Identification of aberrant metabolic networks has been investigated [65, 29]. For example, Yizhak *et al*. [65] proposed an integrative omics-metabolic analysis (IOMA) method, which modeled a metabolic network by integrating quantitative proteomic and metabolomic data and predicted alterations in the metabolic flux under various perturbations. They formulated the problem using quadratic programming to seek a steady-state flux distribution in which the flux through the reactions was measured using proteomic and metabolomic data. Through validation sets, Yizhak *et al*. showed that IOMA had a significant advantage over other commonly used methods of flux balance analysis (FBA) and minimization of metabolomic adjustment (MOMA).

APPLICATIONS OF PATHWAY AND NETWORK-BASED COMPUTATIONAL METHODS

In the following section, we present some of the applications of pathway and network-based approaches including the identification of functional modules and biomarker discovery using HTD.

Identification of functional modules

Great progress has been made in applying information about aberrant pathways and networks to the identification of functional modules, that is, groups of biological entities (e.g. gene, protein) that perform biological tasks (e.g. biological processes) that the constituent parts could not perform if they were dissociated [37, 80]. Sequence mutations, copy number alterations, gene fusion events or epigenetic changes can all lead to changes in the functions of such modules. In the following, we discuss three previously published examples. The first two examples use information from altered network activities to identify functional modules, whereas the third example uses information from aberrant pathways.

Wu *et al*. [37] studied functionally related genes using biological pathway-based inferences to extract, cleanse and filter false positives and noise out of data. The protein functional interactive networks they built included protein–protein interactions, gene co-expressions, protein domain interactions, gene ontology (GO) annotations, and text-mined protein interactions. Applying their method to GBM, breast, colorectal and pancreatic cancers, they found that most samples from these cancers have sequence-altered genes involving commonly known oncogenes and signal transduction modules.

To investigate the combined effect of single markers/genes on mediating complex diseases and traits, Jia *et al*. [36] proposed an integrative approach for identifying candidate subnetworks by integrating the association signal from genome-wide association studies (GWAS) into human protein–protein interaction networks. The core searching algorithm was modified from a previous method [81] that was designed for module searching in gene expression datasets. This approach was able to identify functional modules with higher association signals compared to previously published methods. The top-ranked

module is comprised of 12 genes including BARD1, FGFR2 and GSK3B, which are strongly enriched in breast cancer.

Recently, Koeva *et al.* [82] developed a computational method, stemness meta-analysis pipeline (S-MAP), to test coordinately for modules and individual genes that are up-regulated (stemness-on) or down-regulated (stemness-off) in stem cells. By integrating gene functional modules derived from pathways, protein–protein interactions, homologs and protein complexes, and by extracting multiple gene modules and differential expressed genes from different microarray gene expression databases, they utilized several statistical scores to identify 40 genes and 224 stemness modules upregulated in multiple stem cell types and metastatic populations compared to non-metastatic populations. Their methods demonstrate how omics data can be used to classify normal, cancer and metastatic stem cells.

Biomarker discovery

Biomarkers are primarily molecular markers, such as genes, proteins, metabolites, glycans and other molecules, that can be used for disease diagnosis and prognosis, for predicting therapeutic responses and for developing therapies [38]. Information derived from aberrant pathways and network activities facilitates the detection of diagnosis and treatment biomarkers [83], the identification of novel drug targets [19], the classification of disease types [84] and the prediction of outcomes [33, 84–86]. Omics data-driven methods for identifying diagnosis, prognosis markers and therapeutic targets include machine learning methods [19, 83], graph theory [33] and statistical methods [83]. Many studies have capitalized on aberrant pathway and network activities coupled with omics data integration strategies to identify biomarkers. In the following, we present four published studies. The first example used information from altered network activities to identify biomarkers, whereas the remaining examples utilized information from aberrant pathways.

For example, Taylor *et al.* [33] studied the altered modularity of a protein interaction network to predict breast cancer outcomes by examining the biochemical structure of an interactome. They employed the method proposed by Han *et al.* [87] to use static properties of a network to measure the hubs in protein interaction networks. Hubs associated with cancer were normalized for the frequency of each hub type and for significant differences in the distribution of hubs between cancer and non-cancer genes, as determined by Fisher's exact test. For instance, dynamic network properties connected with breast cancer show that the expressions of BRCA1 and its interactors (e.g. BRCA2 and MRE11) are highly correlated between protein pairs in surviving patients, whereas that organization is lost in patients who die of the disease. This suggests that an altered modularity of these human interactomes may be a biomarker of breast cancer prognosis.

Recently, Andre *et al.* [19] identified greatly amplified changes in the DNA copy number of the FGRR1, VEGFA, E2F3 and NOTCH4 genes. By studying the correlation of DNA copy number with gene expression and the patients' response to therapy, they confirmed that these genes are potential therapeutic targets and could be biomarkers useful for classifying breast cancers. Their integrative analysis of comparative genomic hybridization (CGH) arrays and matched gene expression array data demonstrated that DNA copy number anomalies are strongly associated with chemotherapy efficacy. Therefore, their results included several potential therapeutic target genes that were identified by analyzing anomalies in the DNA copy numbers and gene expression data.

Another interesting example of biomarker discovery is that Sreekumar *et al.* [30] combined prior knowledge of gene expression and transcription factor binding in prostate cancer with liquid and gas chromatography-based mass spectrometry to identify an alteration in a key metabolite, sarcosine, which appeared to be associated with prostate cancer progression. However, subsequent studies have questioned the validity of sarcosine as a biomarker of prostate cancer progression [88–90].

In addition to the discovery of diagnostic biomarkers and disease outcome predictors, one of the promising applications of signaling pathway research is that understanding the dysfunctional responses of signaling pathway associated with cancer could lead to the discovery of more effective and selective therapy targets. Mutational activation of Ras, which is involved in cellular signal transduction, can cause human cancers [59]. K-Ras is an oncoprotein and is an early player in Ras signal transduction pathways. Singh *et al.* [83] used first whole-genome SNP array data to identify molecular features that distinguish K-Ras-dependent and K-Ras-independent cancer cell lines. They found that the vast majority of K-Ras-dependent cell lines exhibited focal *K-Ras*

genomic amplification. Based on gene expression data from K-Ras-dependent and K-Ras-independent cell lines, they identified 'K-Ras dependency signature' genes (involving the kinases SYK and RON), which can potentially serve as candidate therapeutic targets. The identification of these signature genes was performed using the prediction analysis of microarrays (PAM) algorithm. Expression of the Ras dependency signature genes was found predominantly in K-Ras mutant tumors.

The primary goal of most of the approaches reviewed here was to identify disease biomarkers based on integrating HTD and pathway/network information. This has allowed researchers to better understand: i) network-based mechanisms underlying complex common diseases; ii) strategies to improve disease classification; iii) approaches to enhance robustness in biomarker selection and disease classification; and iv) the identification of potential disease drivers or causal agents at various levels of biological organization. However, efforts to integrate different types of HTD and pathway and network information into biomarker discovery studies must continue.

CONCLUSIONS AND OUTLOOK

Recent studies have shown that significant progress has been made in understanding aberrant pathways and identifying network activities using high-throughput techniques. The examples cited in this review demonstrate that a well-built data integration model has the power to correctly explain observations, identify relationships between different biological components, and lead to deepened insight into biological mechanisms.

Many advances in understanding complex diseases have been achieved via HTD. Among them, perhaps the most remarkable are achievements in identifying changes in the genetic entities involved in pathways and networks. Researchers have used diverse data sources from varying perspectives to study the etiology of complex diseases, including cancer.

This review presents studies that used various computational methods coupled with HTD to solve specific problems in the biomedical research field. The successful application of these methods can be instructive by providing guidance into ways to continue uncovering the mechanisms of disease. Although the choice of computational methods depends on the specific problem and the omics data available, some computational methods seem preferential for specific pathway and network inference problems. For example, for signaling pathways, Bayesian networks [34], graph theory [54], Boolean networks [45] and perturbation measurements [50] are preferable. For modeling gene regulatory pathways, dynamic Bayesian models [58], graph theory [91] and gene set enrichment analysis [55] seem to be appropriate. The methods for modeling metabolic pathways have focused on Petri-nets [62], extreme pathway analysis [63], stochastic models [64] and graph theory [42]. The simplest model for gene regulatory networks and pathways is the Boolean network. Amit *et al*. [50] used a perturbation-based approach to build a network of cell signaling pathways that form the mechanistic basis for pathogen-specific responses by dendritic cells with meshed RNA interference screening and global transcriptional profiling [50]. Faust *et al*. [42] applied a subgraph approach to extract metabolic pathways from metabolic networks using metabolomic data. Through the application of these methods, a cohort of biomarkers has been identified for future drug development [19, 83, 92–94]. In spite of the great achievements made to date, identifying aberrant pathways and networks via high-throughput assays is still a new and developing field.

Understanding the aberrancies in biological pathways and networks responsible for complex diseases is far from complete. The use of HTD together with large-scale biological databases (e.g. protein–protein interaction and pathway databases) is crucial for uncovering aberrant biological processes. The current volume of omic data is large and diverse, but even larger and more complete data are required to allow for better understanding of the mechanisms, the heterogeneity and complexity of many common diseases such as cancer. Much work has yet to be done to identify the subnetworks of metabolic reactions associated with each disease. A reliable computational approach for identifying subnetwork-associated diseases is currently limited by the incompleteness of the interactome maps available and by the limitations of the existing tools.

Gaining an integrated understanding of the interactions among the genome, the proteome, the environment and pathophenome, as mediated by the underlying cellular network, may offer a basis for future advances. Currently, some of the most difficult problems in this area include understanding and finding ways to identify dynamic (rather than static) processes in the cell, connecting molecular level network activities to functional behavior at cellular

level, developing a data-driven computational model that reflects the causal relationships between drug targets and biomarkers, measuring the changes that occur in cellular dynamic processes, and intervening in the system to change a given outcome.

An especially challenging research focus of particular importance is to find ways to model changes in biological entities which could affect the dynamics of the biological process using large-scale, diverse omics data. Network-based research is shifting its focus toward integrated multiple networks or toward networks composed of heterogeneous large-scale data. At this point, in order to process large scale omics data, researchers should consider focusing on developing algorithms and tools for studying the relationships between aberrant human genes, proteins and interactome networks, on predicting new human disease-associated genes based on pathways and networks, and on analyzing network perturbations caused by pathogens.

Key points

- Computational methods for analysis of high-throughput omics data to identify characteristic aberrancies in the biological pathways and molecular network activities and elucidate their relationship to the disease.
- Successful applications of pathway and network-based computational methods for functional module identification and biomarker discovery.
- Future directions of research in this field including: (i) the integration of heterogeneous and diverse data at different levels, such as DNA, mRNAs, protein and metabolites; (ii) the need for advanced algorithm and tools to better understand the mechanism of the disease cellular pathways and networks; (iii) ways of modeling aberrant biological entities that could affect the dynamics of biological process using large scale and diverse data.

Acknowledgements

We thank Dr. Rhoda Perozzi for her kindness in performing a thorough review and edit. We thank the anonymous reviewers for their suggestions for improving the manuscript. We appreciate the organizers and participants of the Workshop on Identification of Aberrant Pathway and Network Activity from High-Throughput Data at the 2011 Pacific Biocomputing Symposium for providing new insights into this field.

FUNDING

This work was supported by Grant Number R01GM086746 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

1. Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* 2009;**461**:218–23.
2. Tan CS, Bodenmiller B, Pasculescu A, *et al*. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal* 2009;**2**:ra39.
3. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 2010;**11**:685–96.
4. Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;**39**:D691–7.
5. Chang JT, Carvalho C, Mori S, *et al*. A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol Cell* 2009;**34**:104–14.
6. Wang CC, Cirit M, Haugh JM. PI3K-dependent cross-talk interactions converge with Ras as quantifiable inputs integrated by Erk. *Mol Syst Biol* 2009;**5**:246.
7. Kim TY, Kim HU, Lee SY. Data integration and analysis of biological networks. *Curr Opin Biotechnol* 2010;**21**:78–84.
8. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis* 2010;**31**:2–8.
9. Lu R, Markowetz F, Unwin RD, *et al*. Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature* 2009;**462**:358–62.
10. Sanger F, Air GM, Barrell BG, *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977;**265**:687–95.
11. Schneider MV, Orchard S. Omics technologies, data and bioinformatics principles. *Methods Mol Biol* 2011;**719**:3–30.
12. Ochs MF, Karchin R, Ressom H, *et al*. Identification of aberrant pathway and network activity from high-throughput data - Workshop Introduction. *Pac Symp Biocomput* 2011; 364–8.
13. Kanehisa M, Goto S, Kawashima S, *et al*. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004; **32**:D277–80.
14. Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003; **13**:2363–71.
15. Schaefer CF, Anthony K, Krupa S, *et al*. PID: the Pathway Interaction Database. *Nucleic Acids Res* 2009;**37**: D674–79.
16. Smoot ME, Ono K, Ruscheinski J, *et al*. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2010;**27**:431–2.
17. Hu Z, Hung JH, Wang Y, *et al*. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res* 2009;**37**:W115–21.
18. Nikitin A, Egorov S, Daraselia N, *et al*. Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* 2003;**19**:2155–7.
19. Andre F, Job B, Dessen P, *et al*. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clin Cancer Res* 2009; **15**:441–51.

20. Vaske CJ, Benz SC, Sanborn JZ, *et al*. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;**26**:i237–45.
21. Tan CS, Pasculescu A, Lim WA, *et al*. Positive selection of tyrosine loss in metazoan evolution. *Science* 2009;**325**: 1686–8.
22. Chari R, Thu KL, Wilson IM, *et al*. Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer Metastasis Rev* 2011;**29**:73–93.
23. Martin-Subero JI, Esteller M. Profiling epigenetic alterations in disease. *Adv Exp Med Biol* 2011;**711**:162–77.
24. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. *Mol Oncol* 2011;**5**:5–23.
25. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* 2011; **11**:426–37.
26. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**: 87–98.
27. Huang SS, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal* 2009;**2**:ra40.
28. Samaga R, Saez-Rodriguez J, Alexopoulos LG, *et al*. The logic of EGFR/ErbB signaling: theoretical properties and analysis of high-throughput data. *PLoS Comput Biol* 2009;**5**: e1000438.
29. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* 2010;**6**:401.
30. Sreekumar A, Poisson LM, Rajendiran TM, *et al*. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 2009;**457**:910–14.
31. Quackenbush J. Microarray data normalization and transformation. *Nat Genet* 2002;**32(Suppl)**:496–501.
32. Soreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *J Clin Pathol* 2009;**62**:1–5.
33. Taylor IW, Linding R, Warde-Farley D, *et al*. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 2009;**27**:199–204.
34. Ochs MF, Rink L, Tarn C, *et al*. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res* 2009; **69**:9125–32.
35. Raichur S, Fitzsimmons RL, Myers SA, *et al*. Identification and validation of the pathways and functions regulated by the orphan nuclear receptor, ROR alpha1, in skeletal muscle. *Nucleic Acids Res* 2010;**38**:4296–312.
36. Jia P, Zheng S, Long J, *et al*. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics* 2010;**27**: 95–102.
37. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* 2010;**11**:R53.
38. Ransohoff DF, Gourlay ML. Sources of bias in specimens for research about molecular markers for cancer. *J Clin Oncol* 2010;**28**:698–704.
39. Siena S, Sartore-Bianchi A, Di Nicolantonio F, *et al*. Biomarkers predicting clinical outcome of epidermal growth factor receptor-targeted therapy in metastatic colorectal cancer. *J Natl Cancer Inst* 2009;**101**:1308–24.
40. Loss LA, Sadanandam A, Durinck S, *et al*. Prediction of epigenetically regulated genes in breast cancer cell lines. *BMC Bioinformatics* 2010;**11**:305.
41. Ma S, Kosorok MR. Identification of differential gene pathways with principal component analysis. *Bioinformatics* 2009; **25**:882–9.
42. Faust K, Dupont P, Callut J, *et al*. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics* 2010;**26**:1211–18.
43. Novershtern N, Regev A, Friedman N. Physical Module Networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics* 2011;**27**:i177–85.
44. Krumsiek J, Suhre K, Illig T, *et al*. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol* 2011;**5**:21.
45. Heiser LM, Wang NJ, Talcott CL, *et al*. Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biol* 2009;**10**:R31.
46. Aldridge BB, Saez-Rodriguez J, Muhlich JL, *et al*. Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/ insulin-induced signaling. *PLoS Comput Biol* 2009;**5**: e1000340.
47. Chien CH, Sun YM, Chang WC, *et al*. Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res* 2011; **39**(21):9345–56.
48. Kemper B, Matsuzaki T, Matsuoka Y, *et al*. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics* 2010;**26**:i374–81.
49. Kupershmidt I, Su QJ, Grewal A, *et al*. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One* 2010. **5(9)**. pii: e13066.
50. Amit I, Garber M, Chevrier N, *et al*. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 2009;**326**:257–63.
51. Hecker M, Lambeck S, Toepfer S, *et al*. Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* 2009;**96**:86–103.
52. Trinh CT, Wlaschin A, Srienc F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl Microbiol Biotechnol* 2009;**81**:813–26.
53. Sardiu ME, Washburn MP. Building protein–protein interaction networks with proteomics and informatics tools. *J Biol Chem* 2011;**286**:23645–51.
54. Yeger-Lotem E, Riva L, Su LJ, *et al*. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* 2009;**41**: 316–23.
55. Keller A, Backes C, Gerasch A, *et al*. A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics* 2009;**25**: 2787–94.
56. Xu TR, Vyshemirsky V, Gormand A, *et al*. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci Signal* 2010;**3**: ra20.
57. Fujita PA, Rhead B, Zweig AS, *et al*. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2011;**39**: D876–82.

58. Li Z, Li P, Krishnan A, *et al*. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* 2011;**27**:2686–91.
59. Liao JC, Boscolo R, Yang YL, *et al*. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA* 2003;**100**:15522–7.
60. Alakwaa FM, Solouma NH, Kadah YM. Construction of Gene Regulatory Networks using biclustering and Bayesian networks. *Theor Biol Med Model* 2011;**8**:39.
61. Kerrien S, Alam-Faruque Y, Aranda B, *et al*. IntAct—open source resource for molecular interaction data. *Nucleic Acids Res* 2007;**35**:D561–5.
62. Zevedei-Oancea I, Schuster S. Topological analysis of metabolic networks based on petri net theory. *Stud Health Technol Inform* 2003;**162**:17–37.
63. Rezola A, de Figueiredo LF, Brock M, *et al*. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics* 2010;**27**:534–40.
64. Mithani A, Preston GM, Hein J. A stochastic model for the evolution of metabolic networks with neighbor dependence. *Bioinformatics* 2009;**25**:1528–35.
65. Yizhak K, Benyamini T, Liebermeister W, *et al*. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* 2010;**26**:i255–60.
66. Berthoumieux S, Brilli M, de Jong H, *et al*. Identification of metabolic network models from incomplete high-throughput datasets. *Bioinformatics* 2011;**27**:i186–95.
67. Goffard N, Frickey T, Weiller G. PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways. *Nucleic Acids Res* 2009;**37**:W335–9.
68. Karp PD, Riley M, Saier M, *et al*. The EcoCyc Database. *Nucleic Acids Res* 2002;**30**:56–8.
69. Schweiger R, Linial M, Linial N. Generative probabilistic models for protein–protein interaction networks–the biclique perspective. *Bioinformatics* 2011;**27**:i142–8.
70. Bui QC, Katrenko S, Sloot PM. A hybrid approach to extract protein–protein interactions. *Bioinformatics* 2011;**27**:259–65.
71. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 2006;**7**:198–210.
72. Tsiliki G, Kossida S. Fusion methodologies for biomedical data. *J Proteomics* 2011;**74**:2774–85.
73. Jacobs F, Wang L. Adeno-associated viral vectors for correction of inborn errors of metabolism: progressing towards clinical application. *Curr Pharm Des* 2011;**17**:2500–15.
74. Johannes M, Frohlich H, Sultmann H, *et al*. pathClass: an R-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics* 2011;**27**:1442–3.
75. Wang YC, Zhang CH, Deng NY, *et al*. Kernel-based data fusion improves the drug-protein interaction prediction. *Comput Biol Chem* 2011;**35**:353–62.
76. Bauer-Mehren A, Furlong LI, Sanz F. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol* 2009;**5**:290.
77. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res* 2011;**71**:4550–61.
78. Hart GW, Slawson C, Ramirez-Correa G, *et al*. Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu Rev Biochem* 2011;**80**:825–58.
79. Wang Z, Udeshi ND, Slawson C, *et al*. Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Sci Signal* 2010;**3**:ra2.
80. Qiu YQ, Zhang S, Zhang XS, *et al*. Identifying differentially expressed pathways via a mixed integer linear programming model. *IET Syst Biol* 2009;**3**:475–86.
81. Chuang HY, Lee E, Liu YT, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
82. Koeva M, Forsberg EC, Stuart JM. Computational integration of homolog and pathway gene module expression reveals general stemness signatures. *PLoS One* 2011;**6**:e18968.
83. Singh A, Greninger P, Rhodes D, *et al*. A gene expression signature associated with "K-Ras addiction" reveals regulators of EMT and tumor cell survival. *Cancer Cell* 2009;**15**:489–500.
84. Gatza ML, Lucas JE, Barry WT, *et al*. A pathway-based classification of human breast cancer. *Proc Natl Acad Sci USA* 2010;**107**:6994–9.
85. Cerami E, Demir E, Schultz N, *et al*. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 2010;**5**:e8918.
86. Chen J, Sam L, Huang Y, *et al*. Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. *J Biomed Inform* 2010;**43**:385–96.
87. Han JD, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 2004;**430**:88–93.
88. Struys EA, Heijboer AC, van Moorselaar J, *et al*. Serum sarcosine is not a marker for prostate cancer. *Ann Clin Biochem* 2010;**47**:282.
89. Jentzmik F, Stephan C, Lein M, *et al*. Sarcosine in prostate cancer tissue is not a differential metabolite for prostate cancer aggressiveness and biochemical progression. *J Urol* 2011;**185**:706–11.
90. Colleselli D, Stenzl A, Schwentner C. Re: Florian Jentzmik, Carsten Stephan, Kurt Miller, *et al*. Sarcosine in urine after digital rectal examination fails as a marker in prostate cancer detection and identification of aggressive tumours. *Eur Urol* 2010;**58**:12–18. Eur Urol 58:e51.
91. Alon N, Dao P, Hajirasouliha I, *et al*. Biomolecular network motif counting and discovery by color coding. *Bioinformatics* 2008;**24**:i241–9.
92. Barbie DA, Tamayo P, Boehm JS, *et al*. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 2009;**462**:108–12.
93. Luo J, Emanuele MJ, Li D, *et al*. A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* 2009;**137**:835–48.
94. Scholl C, Frohling S, Dunn IF, *et al*. Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* 2009;**137**:821–34.