Published in final edited form as:

Adv Drug Deliv Rev. 2013 June 30; 65(7): 966–972. doi:10.1016/j.addr.2012.12.004.

Drug target inference through pathway analysis of genomics data

Haisu Ma¹ and Hongyu Zhao^{1,2,*}

¹Program in Computational Biology and Bioinformatics, Yale University, New Haven CT 06511; haisu.ma@yale.edu

²Department of Biostatistics, Yale School of Public Health, New Haven CT 06520 USA; hongyu.zhao@yale.edu

Abstract

Statistical modeling coupled with bioinformatics is commonly used for drug discovery. Although there exist many approaches for single target based drug design and target inference, recent years have seen a paradigm shift to system-level pharmacological research. Pathway analysis of genomics data represents one promising direction for computational inference of drug targets. This article aims at providing a comprehensive review on the evolving issues in this field, covering methodological developments, their pros and cons, as well as future research directions.

Keywords

Drug target inference; pathway analysis; genomics; statistical modeling; factor model; data mining; optimization

1. Introduction

The dominant paradigm in drug discovery is the “one drug-one target” approach, which aims to design the most selective drug molecules to act on individual targets [1]. However, this paradigm ignores the cellular and physiological context of the drugs' mechanism of action, and makes it difficult to address the safety and toxicity issues in drug development [2]. In addition, many complex diseases (such as cancer, cardiovascular diseases, neurological disorders) result from the dysfunction of multiple pathways rather than a small number of individual genes, it is important to consider all relevant pathways involved for the design of effective therapies. Recent technological advances have allowed researchers to collect large-scale datasets on various properties of compounds [3], features of target genes/proteins [4], as well as responses in the human physiological system [5]. These high dimensional data sets present unique opportunities to delineate complex interactions and responses among biological pathways to drug treatments, but they also present great challenges to researchers because they have very high dimensionality, complex structure, and are of distinct types. Although various computational and statistical methods have been

© 2012 Elsevier B.V. All rights reserved

*Corresponding author: Yale School of Public Health, PO Box 208034, 60 College Street, New Haven, CT 06520-8034; Tel: (203)785-3613; Fax: (203) 785-6912.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

proposed to analyze these data for drug target inference, not all aspects of the data complexity are well addressed to appropriately analyze and interpret these data. In this review, we provide an overview of the existing methods for drug target inference through pathway analysis of genomic data and discuss their advantages and limitations of these with the goal of stimulating further methodology developments in this critical research area.

2. Omics Technology platforms utilized in drug target discovery

The choice of mathematical models for drug target inference highly depends on the type of data used. In order to measure the biological effects of drug perturbation in a high throughput manner, a number of experimental approaches have been developed. Generally speaking, these approaches fall in three categories: genomics, proteomics and metabolomics [6].

2.1. Genomic technologies

Genomic technologies aim to characterize the physiological state of cell lines, tissues, as well as organisms, from the perspective of the genome. Among different -omics approaches, genomic technologies are more developed and robust. The types of data collected include DNA variations, transcription levels, epigenetic changes, and histone modifications among others.

Microarrays were developed in the mid-1990s and were quickly adopted for genotyping and expression profiling [7]. At very affordable cost, gene expression arrays can measure the transcription levels of all the genes in the genome at the gene or exon levels. Genotyping arrays now can query up to 5 millions of DNA variations in the genome. Arrays can also measure copy number variations. For example, comparative genomic hybridization (CGH), a molecular-cytogenetic method for the analysis of copy number variation (CNV) in the DNA content (amplification or deletion of chromosomal regions) of given samples [8], was developed around two decades ago. Since many complex diseases including cancer are characterized by large-scale disorganization of the genome, CGH represents a powerful platform for analyzing the contribution of genomic instability to pathogenesis [8]. CNV can also be inferred from genotyping arrays. Microarrays have been used to gather information from hundreds of thousands of biological samples, and many of them can be found in GEO [9]. For example, the NCI 60 cell lines have been analyzed by National Cancer Institute and National Institute of Health [10]. The expression profiles may capture the bioactivities of various genes and proteins, and therefore provide valuable information on the biological states, for example, healthy/diseased conditions and before/after drug treatment. Comparative analysis of transcriptional profiles has led to the identification of biomarkers for diseases biomarkers.

Recently, notable progress in next generation sequencing (NGS) technologies has attracted tremendous interests in the bioscience community. They are more sensitive and accurate than microarrays and also have broader application areas. They have been employed to identify genetic alterations (deletion, insertion, repetition), measure transcript levels, discover novel isoforms, protein-DNA interactions, and infer epigenetic status (DNA methylation and histone modification) [6]. For example, RNA-seq [11], also known as “Whole Transcriptome Shotgun Sequencing” (WTSS) [12], can provide deep coverage and base-level resolution. Compared to microarrays, more detailed information on gene expression, including gene alleles, differently spliced transcripts, non-coding RNAs, post-transcriptional mutations, and gene fusions can be gathered [13].

2.2. Proteomics technology

Compared with messengers that are targeted by genomic technologies (such as microarrays and RNA-seq), proteins are the functional units in biological systems and more direct targets for drugs. Applications of proteomic technologies in drug discovery and development include target identification, efficacy/toxicity biomarkers discovery, protein/drug interaction analysis, and drug action mechanism investigation [14]. Two-dimensional gel electrophoresis (2-DE) and mass spectrometry can be utilized to profile protein expression levels [15]. Techniques have also been developed to achieve higher throughput and greater coverage of the proteome, among which PST (protein sequence tags), MudPIT (multidimensional protein identification technology), ICAT (isotope-coded affinity Tagging) are representative examples [15]. iTRAQ (isobaric tags for relative and absolute quantitation) [16] and MRM (multiple reaction monitoring) [17] are two recently developed popular assays.

2.3. Metabolomic technologies

Metabolomics is relatively new compared with genomics and proteomics. It measures the concentrations of small molecule metabolites at the system level using nuclear magnetic resonance, liquid chromatography, mass spectrometry and other technologies [18]. Different from transcripts and proteins, metabolites refer to a wide range of small molecules in a cell or organism, which may include both endogenous and exogenous entities such as carbohydrates, organic acids, food additives, and others [18]. Compared with genomics and proteomics, metabolomics has the advantage of measuring the quantitative changes of a much wider spectrum of different types of biomolecular entities within shorter time ranges (metabolic responses are often measured in seconds or minutes whereas genetic response takes days or weeks) [18]. Therefore, metabolomic technologies may capture important information on physiological phenotypes which is beyond that can be measured by genomic and proteomic technologies.

3. Computational methods for genome-wide drug-target interaction prediction

As discussed above, comprehensive understanding of the targeting spectrum of various drugs is critical for drug mechanism studies and the developments of multi-drug treatment, the so-called polypharmacology [19]. Even in the case that the molecular mechanisms for a drug/compound are well studied, unexpected side effects or toxicity of are often observed [19–21]. Therefore, it is of great value and importance to construct comprehensive drug-target network using diverse sources of information.

Current approaches for computational drug target identification can be roughly categorized into three groups: ligand-based, target-based, and phenotype-based [19]. Ligand-based approaches assume that similarity in chemical structures between drugs indicates similar targeting activities. A well known example in this category is the QSAR method (Quantitative Structure Activity Relationship), which uses 2D topological feature vectors (to encode atom types and their bonding structure) of drugs to train machine learning models in order to predict their binding activity towards specific target proteins [19, 22]. However, decent performance of ligand-based approach requires a large enough number of known ligands for target proteins of interest, which may be difficult to meet in practice [20].

A second category of methods looks at the similarity between target proteins to predict drug-target interactions, based on proteins' structure, sequence, evolutionary, as well as functional information [23]. Several recent studies integrate sequence features of targets with ligand fingerprints to train machine learning models for drug-target interaction prediction [24–27].

One well known method in this category is docking analysis, which predicts the preferred orientation of drug candidates to potential target proteins when they are bound to each other to form a stable complex [28]. However, docking analysis cannot be applied to proteins whose 3D structures are unknown. It is of limited utility on a genome-wide scale [29].

The third class of methods associates different drugs by comparing the biological phenotype responses, such as cell lines' gene expression profiles or proteomic data [23]. Seminal work in this direction includes the national NCI-60 project [30], which screened 60 human tumor cell lines against more than 100,000 compounds and constructed a public repository for the basal gene expression and drug sensitivity information. The Connectivity Map project initiated by the Broad Institute went a step further [31, 32]. Although it only focused on five human cancer cell lines, the project generated genome-wide expression profiles both before and after drug treatment for 1,309 compounds. In this way, compounds can be connected into a network by comparing their ranked lists of up- and down-regulated genes [33, 34]. Other phenotype information such as cell imaging and side effects have also been utilized to associate different drugs and to make inference about their potential targets [21, 35]. Pros and cons of the above three approaches are summarized in Table 1.

4. Pathway analysis of genomic data for drug target inference

In the following, we focus on statistical methods belonging to the third group of drug target prediction methods mentioned above, which analyze genomic phenotype data and additional drug-related information to enable drug target discoveries.

4.1 Types of data utilized in drug-related pathway analysis

The genomic, proteomic, and metabolic technologies discussed above all have been employed to generate data sets that convey useful information for pathway-based drug target prediction. Because of the longer history and relatively robust platforms, data generated from the genomic technologies, such as gene expression profiles, and data collected from drug sensitivity analysis are the two most commonly used data sources for drug target prediction. As for gene expression data, they can be collected before (base-line) and after (response) drug treatments. In many instances, expression profiles at multiple time points after treatment are collected. Baseline expression provides an overview of the static landscape of the transcriptome activities. Combined with after-treatment measurements, gene expression profile alternations in response to drug treatment can be analyzed to infer drug targets. The NCI-60 project represents a successful initiative to comprehensively utilize baseline gene expression data to investigate drug mechanisms [22]. It has been shown that basal pathway activity levels may be used to infer the response to drug treatments [36]. In contrast, the Connectivity Map project aimed to generate data about gene expression changes induced by compound treatments [23, 24]. The data generated from the Connectivity Map project allow a more direct association between genome-wide expression responses and various chemical compounds, and as a result, more accurate understanding on potential drug targets [20]. Apart from expression data, additional genomic datasets available for pathway analysis include copy number variation, DNA methylation status and fingerprints, sequence mutations of exons and introns, and others [37].

Biological response or sensitivity to drug treatments is usually measured by GI50, which is the minimum concentration of the drug needed to inhibit the growth of cells by 50%. Therefore, higher GI50 indicates drug-resistant response whereas lower GI50 indicates drug-sensitive response. Additional available information on drugs may include chemical properties, absorption, distribution, metabolism, elimination and toxicity [38], but these types of information have not been well utilized in computational analysis, partly due to the qualitative, instead of quantitative, nature of these data.

4.2 General background on pathway analysis

4.2.1 Definition of biological pathways—The definition of a biological pathway, based on the website of National Human Genome Research Institute, is “a series of actions among molecules in a cell that leads to a certain product or a change in a cell.” (<http://www.genome.gov>). In terms of function, biological pathways can be broadly classified into three groups: metabolic pathways, gene regulation pathways and signal transduction pathways. Metabolic pathways are responsible for the chemical reactions involved in the biosynthesis or decomposition of various metabolites (such as proteins, nucleic acids, lipids). Gene regulations pathways control the on/off genetic information flow which determines protein expression both qualitatively and quantitatively. Signal transduction pathways work as mailmen, carrying signals from the exterior environment to the interior cellular compartment, via the interactions among signaling molecules, receptors on the cell surface, and additional information transporter proteins which often go through certain chemical reactions (i.e. phosphorylation/dephosphorylation) in order to accomplish signal transmission. Figure 1 illustrates these three different types of biological pathways discussed above (from the National Human Genome Research Institute <http://www.genome.gov>). A collection of interacting biological pathways form a biological network. Usually discovered via molecular and biochemical experiments in model organisms, biological pathways have varying degrees of conservation across the species, with highly-conserved pathways performing fundamental physiological functions. An extensive understanding and annotations of pathways is of great importance to facilitate disease-related biomedical researches.

4.2.2 Bioinformatics methods for pathway analysis—Depending on whether the members in a pathway being analyzed are known a priori or not, pathway analysis roughly falls into two approaches: candidate pathway significance test and pathway identification.

Thanks to the availability of a large number of pathway databases constructed by the research community, computational and experimental biologists now have access to rich information regarding the genes involved in each specific biological pathway. The knowledge in these databases (which will be briefly described in the following section) may be acquired via either hands-on experimental studies or online literature mining of published articles. Once the pathway affiliation of all the genes is determined, the significance of the differential expression of the pathways across samples belonging to distinct phenotype groups can be assessed. For example, after assembling gene expression profiles from both normal and cancer cell lines, the varying activities of disease-related pathways between the two groups can be computed. Gene Set Enrichment Analysis [39] is a well-known computational method to identify pathways enriched for genes showing different expression patterns across different conditions (e.g. cancer versus normal) or treatment groups (e.g. different compounds). See [40] for a good review and method comparison. This class of approaches has the advantage of clear biological interpretation; however, it relies on the prior knowledge of gene sets (genes in the same pathway or related to the same disease), which is still far from complete. Moreover, under many circumstances, only a subset of genes in a pathway exhibit expression alteration, and testing on the whole set may lead to false negative results [41].

An alternative approach in pathway analysis aims to identify unknown pathways representing biological functional units. In such analyses, the term “pathway” is often used interchangeably with “gene set”, “module” and “network”, which refers to a group of genes without specific topological structures designated. Here the goal is to identify a subset of genes that are most representative of the condition-specific changes. This class of methods

usually has two main components: a scoring function quantifying the alternation of a given sub-network between different conditions; and a search algorithm to extract the highest-scoring sub-networks [42]. A seminal work in this field [43] introduced a scoring function which measures the differential expression of individual genes (adjusted for the total number of genes selected), and chose simulated annealing as the search algorithm. A later method [44] used a scoring function which measures gene-gene correlation rather than single gene alternations. Apart from global optimization methods such as simulated annealing, local greedy search algorithms [45–49], mathematical programming methods, and other exact approaches based on graph-theory were also proposed. Recent examples include: mixed integer linear programming model [50], prize-collecting Steiner tree problem (PCST) [51], iterative local optimization algorithm [52], and regression model with diffusion kernel [53]. We also developed a method called “COSINE” (COndition-Specific sub-NEtwork) to consider both differential expression and differential co-expression coupled with a global optimization method for sub-network inference [42]. A comparison of these methods is shown in Table 2. In this table, “node-based” methods are those that focus on differential expressions, “edge-based” methods are those focusing on differential correlations between genes, and methods in the “combined” category consider both types of differences. As for optimization methods, “global optimization” methods try to find the globally optimal sub-networks but usually require longer computational time, whereas “local optimization” methods are faster but the results may be sub-optimal.

4.2.3 Useful pathway databases and software—There has been a growing number of pathway databases and analysis software developed among the bioinformatics research community. The journal of Nucleic Acids Research has an annual issue on databases, which provides a good catalog for pathway resources [54]. Many other papers, e.g. [55, 56], also provide excellent reviews on pathway analysis tools.

4.3 Statistical methods to infer pathway-drug association relationships

It can be easily seen that there is distinction between gene-drug and pathway-drug association relationships. Apart from the one-to-one versus one-to-many correspondence, pathway-drug association provides more physiological or functional context of the mechanism of action for the chemical compounds used to treat diseases. One gene can be involved in multiple pathways; therefore, merely knowing the direct molecular target of the drugs may be insufficient to understand how a given drug perturbs the biological system at the phenotype level. Although both pathway analysis and drug target prediction are research fields with a long history and mature methodologies, the inference for the pathway-drug associations remains a relatively young area. In the following we will review some recent explorative attempts in this direction.

4.3.1 Pathway enrichment analysis—An intuitive approach is to apply the well-known Gene Set Enrichment Analysis (GSEA, <http://www.broadinstitute.org/gsea>) [57] to datasets which measure the genomic profiles before and after drug treatment. The analysis procedure of GSEA can be briefly described as follows: Given genome-wide expression profiles from two or more classes of samples (e.g. two cancer subtypes, or drug-treated v.s. control samples), as well as a pre-defined gene set S (e.g. genes associated with the same metabolic pathway, or from the same Gene Ontology category [58]), the goal of GSEA is to assess whether members of set S are enriched among the genes whose expression levels show significant correlation with the group label.

The first step of the algorithm is to calculate the “enrichment score” for each gene set. All the genes are ranked as list L by their correlation with the group label. Then, a running sum statistic is computed by walking down the list L . This statistic is increased every time a gene

within set S is encountered and decreased when a gene not in set S appears. The enrichment score (ES) is defined as the maximum deviation from 0 (either positive or negative) across the whole walk, which can be regarded as a weighted Kolmogorov–Smirnov-like statistic.

The second step is to calculate the significance level of ES by permuting the group label of each sample. After each permutation, the ES can be recalculated and then a null distribution of the score is derived. In this way, an empirical P-value can be obtained for the original ES .

The final step of GSEA is adjustment of the P-value for multiple hypotheses testing when a large number of gene sets are analyzed.

The output of GSEA is an enrichment score ranging from -1 to 1, reflecting the magnitude and direction of pathway activity changes between two or more sample groups. For a specific drug, the pathways can be ranked by this enrichment score to understand the treatment effects. Despite its common use, GSEA has several limitations: first, computation must be done one at a time for each pathway-drug pair; second, all the genes belonging to the same pathway have equal weights, which may not be an optimal strategy when a subgroup of genes act as the crucial interaction partners for a certain drug.

4.3.2 Sparse factor modeling approach—Several studies have employed statistical modeling to predict pathway-drug associations using genomics data [59, 60]. Recently, we have developed two methods for the inference of target pathways for a large number of drugs simultaneously. The first method “iFad” was designed for the analysis of paired baseline gene expression and drug sensitivity (GI50) data [36], and the second method, “FacPad”, for the analysis of gene expression data both before and after drug treatment [20].

In the “iFad” model (an integrative factor analysis model for drug-pathway association inference), gene expression dataset is denoted as matrix Y_1 , with dimension G_1 by J where G_1 is the number of genes and J is the sample size. The paired GI50 dataset is denoted by matrix Y_2 , with dimension G_2 by J where G_2 is the number of drugs. The model aims to jointly decompose the two matrices, explaining the sample variation in both gene expression and GI50 in terms of the latent activities of a much fewer number of biological pathway, denoted by matrix X with dimension K by J , where K is the pre-defined number of pathways. The decomposition is in the form of a Bayesian factor analysis mode as follows:

$$Y_1 = W_1 X + E_1, E_1 \sim N(0, \Sigma_1), \Sigma_1 = \text{diag}\{\tau_{g_1}^{-1}\}$$

$$Y_2 = W_2 X + E_2, E_2 \sim N(0, \Sigma_2), \Sigma_2 = \text{diag}\{\tau_{g_2}^{-1}\}$$

$$X_{k,j} \sim \text{Normal}(0, 1)$$

$$\tau_{g_1} \sim \text{Gamma}(\alpha_1, \beta_1), g_1 = 1, 2, \dots, G_1$$

$$\tau_{g_2} \sim \text{Gamma}(\alpha_2, \beta_2), g_2 = 1, 2, \dots, G_2$$

In most cases, certain prior knowledge on the gene-pathway and drug-pathway association structure (matrices L_1 and L_2) may be available, which may be utilized to guide the inference of loading matrices W_1 and W_2 . The Spike-and-Slab mixture prior [61] provides a natural solution for this. Suppose the actual sparsity patterns of W_1 and W_2 are indicated by binary matrices Z_1 and Z_2 , L_1 and L_2 will influence the estimation of Z_1 and Z_2 in the following flexible way:

$$P(Z_{g,k}=1)=\pi_{g,k}=\begin{cases} \eta_0, & \text{if } L_{g,k}=0 \\ 1-\eta_1, & \text{if } L_{g,k}=1 \end{cases}$$

$$P(W_{g,k})=(1-\pi_{g,k})\delta_0(W_{g,k})+\pi_{g,k}Normal(W_{g,k}|0,\tau_w^{-1})$$

δ_0 is the unit point mass at zero (a.k.a. the Dirac delta function), whereas $\pi_{g,k}$ denotes the Bernoulli probability that $W_{g,k}$ is non-zero.

The basic setup of “FacPad” is also a sparse Bayesian factor model, yet it aims at decomposing a single matrix: the genome-wide expression changes upon treatment of different drugs. Expression changes are calculated as the ratio of post-treatment values to pre-treatment measurements. Statistical inference algorithm for both methods can be developed based on collapsed Gibbs sampling [20, 36].

4.3.3 Other pathway identification methods—Many other methods have been developed in this fast evolving field. A recent study aims to integrate biomedical ontologies from various pharmacogenomic databases to identify significant associations between pathways and drugs [62]. By integrating a number of drug databases including PharmGKB, DrugBank and CTD, powerful queries across multiple ontologies can be performed. These links to disease/chemical ontologies enable queries for disease associations and functional similarity between drugs. A statistical enrichment analysis was also performed using these links, which reveals disease-pathway and drug-pathway associations.

Another study combines existing drug target information, drug response expression data, and human physiological interaction network to identify drug responsive pathways that are not pre-defined, but extracted from seed expansion [63]. The algorithm pipeline includes five steps: (a) Constructing drug-specific subnetworks by connecting drug targets to differently expressed gene sets. (b) Identifying over-represented short paths of four proteins. (c) Constructing a non-redundant set of seed paths. (d) Seed expansion by overlapping paths to yield the full pathways. (e) Associating drugs with pathways if the pathway's proteins significantly overlap the drug-specific subnetwork. Using the newly inferred pathways, they developed a pathway-based drug-drug similarity measure, which was shown to improve the prediction of drug side effects.

Co-expression network analysis has also been applied in drug response analysis [64]. It has been shown that coexpression network modules are generally enriched for genes involved in known biological pathways, linked to common genetic loci, or associated with certain diseases. By associating coexpression modules with drug response (and also other clinical phenotypes), network models may help to underline the actions of drugs on a larger scale.

4.3.4 Pros and cons of different pathway analysis methods—Bayesian sparse factor analysis models, such as iFad and FacPad, try to identify the target biological pathways for drugs with unclear mechanism of action. Their advantages lie in that they

allow natural incorporation of prior knowledge about the connectivity structure of biological pathways (e.g. KEGG pathway), and simultaneously relate the underlying pathway activity to both gene expression levels and/or drug response. Due to this sparse formulation, the sample size needed to achieve satisfactory inference result can be much smaller than the number of features in either dataset. Different from GSEA, they enable genome-wide analysis all at once, and provide estimation for not only drug-pathway associations but also the different weights of genes belonging to the same pathway. However, due to its Bayesian setting, the computation cost is high. In addition, the total number of latent factors has to be chosen by the user, which may prove difficult when prior information regarding the sparsity structure is limited. Another drawback of Bayesian factor modeling methods compared with alternatives is their strong dependence on the accuracy of prior information. Therefore, it is highly desirable to try different analysis procedures on the same dataset and compare the results.

5. Conclusion and future perspectives

Despite the great promises offered by the recently collected large and diverse types of data related to drug treatment responses, these data present great challenges to computational biologists to develop efficient and comprehensive pathway analysis methods to facilitate drug target inference. Although we are in the age of “big data” coming from various “Omic” platforms, searching and extracting the real signals are likely much more demanding than generating these huge datasets. Several issues worthy of note include: first, how to appropriately incorporate prior information to avoid completely blinded, unsupervised data analysis whose result may be difficult for biological interpretation; second, which types of statistical models should be utilized to analyze the data; Different models have different assumptions, sometimes explicit and often times implicit. It is extremely important to have a good understanding of the original patterns of the data before applying certain statistical modeling; third, how to devise efficient computational and visualization tools to allow comprehensive data analysis; fourth, how to choose the gold standard datasets in the literature and design further experiments to validate the inference result. Current catalogs for drug targets are often inaccurate and experimental approaches may be more convincing for final confirmation. Cooperation between experimental and computational scientists is crucial for enabling efficient and meaningful analysis of high-throughput data during drug discovery and development.

Another potential future direction is to analyze genomic data from *in vivo* rather *in vitro*. In this dissertation, all real datasets are from human cell lines. Although cell lines are easier to handle and to perform large-scale drug screen on, they might not represent the true gene expression patterns within the human bodies. In addition, since complex diseases require combinatorial therapy, it will be of great interest to study the drug combination effects and how different drugs interact on the biological pathways.

Acknowledgments

This work was supported in part by the National Institutes of Health grant R01 GM59507 to H.Z., R21-GM084008 to Ning Sun, and National Science Foundation grant DMS 1106738 to H.Z.

References

- [1]. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008; 4:682–690. [PubMed: 18936753]
- [2]. Iskar M, Zeller G, Zhao XM, van Noort V, Bork P. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Current opinion in biotechnology. 2011

- [3]. Giuliano KA, Haskins JR, Taylor DL. Advances in high content screening for drug discovery. Assay and drug development technologies. 2003; 1:565–577. [PubMed: 15090253]
- [4]. Hughes JE. Genomic technologies in drug discovery and development. Drug discovery today. 1999; 4:6–6. [PubMed: 10234143]
- [5]. Petriz BA, Gomes CP, Rocha LAO, Rezende TMB, Franco OL. Proteomics applied to exercise physiology: A cutting-edge technology. Journal of cellular physiology. 2012; 227:885–898. [PubMed: 21503895]
- [6]. Zhao S, Iyengar R. Systems Pharmacology: Network Analysis to Identify Multiscale Mechanisms of Drug Action. Annu Rev Pharmacol. 2012; 52:505–521.
- [7]. Fernandes TG, Diogo MM, Clark DS, Dordick JS, Cabral JMS. High-throughput cellular microarray platforms: applications in drug discovery, toxicology and stem cell research. Trends Biotechnol. 2009; 27:342–349. [PubMed: 19398140]
- [8]. Bayani J, Squire JA. Comparative genomic hybridization. Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.]. 2005; Chapter 22(Unit 22):26.
- [9]. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles - database and tools update. Nucleic Acids Res. 2007; 35:D760–D765. [PubMed: 17099226]
- [10]. Shoemaker RH. The NCI 60 human tumor cell line screen: An information-rich screen informing on mechanisms of toxicity. In Vitro Cell Dev-An. 2006; 42:5A–5A.
- [11]. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
- [12]. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques. 2008; 45:81–94. [PubMed: 18611170]
- [13]. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009; 458:97–101. [PubMed: 19136943]
- [14]. Walgren JL, Thompson DC. Application of proteomic technologies in the drug development process. Toxicol Lett. 2004; 149:377–385. [PubMed: 15093284]
- [15]. Macri J, Rapundalo ST. Application of proteomics to the study of cardiovascular biology. Trends Cardiovas Med. 2001; 11:66–75.
- [16]. Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, Pandhal J, Smith D, Pham TK, Karunakaran E, Zou X, Biggs CA, Wright PC. An insight into iTRAQ: where do we stand now? Anal Bioanal Chem. 2012; 404:1011–1027. [PubMed: 22451173]
- [17]. Waybright T, Xiao Z, Xu X, Faupel-Badger J. Quantitation of Prolactin using Multiple Reaction Monitoring (MRM) Mass Spectrometry. Protein Science. 2012; 21:111–112.
- [18]. Wishart DS. Applications of metabolomics in drug discovery and development. Drugs R&D. 2008; 9:307–322.
- [19]. Xie L, Xie L, Kinnings SL, Bourne PE. Novel Computational Approaches to Polypharmacology as a Means to Define Responses to Individual Drugs. Annu Rev Pharmacol. 2012; 52:361–+.
- [20]. Ma H, Zhao H. FacPad: Bayesian Sparse Factor Modeling for the Inference of Pathways Responsive to Drug Treatment. Bioinformatics. 2012
- [21]. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008; 321:263–266. [PubMed: 18621671]
- [22]. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010; 26:i246–i254. [PubMed: 20529913]
- [23]. Xie L, Kinnings SL, Bourne PE. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. Annual review of pharmacology and toxicology. 2012; 52:361–379.
- [24]. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008; 24:i232–240. [PubMed: 18586719]

- [25]. Nagamine N, Shirakawa T, Minato Y, Torii K, Kobayashi H, Imoto M, Sakakibara Y. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *Plos Comput Biol.* 2009; 5:e1000397. [PubMed: 19503826]
- [26]. Vina D, Uriarte E, Orallo F, Gonzalez-Diaz H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. *Molecular pharmaceutics.* 2009; 6:825–835. [PubMed: 19281186]
- [27]. He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, Chou KC. Predicting drug-target interaction networks based on functional groups and biological features. *PloS one.* 2010; 5:e9603. [PubMed: 20300175]
- [28]. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews. Drug discovery.* 2004; 3:935–949.
- [29]. Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics.* 2010; 26:i246–254. [PubMed: 20529913]
- [30]. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nature reviews. Cancer.* 2006; 6:813–823.
- [31]. Lamb J. The Connectivity Map: a new tool for biomedical research. *Nature reviews. Cancer.* 2007; 7:54–60.
- [32]. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science.* 2006; 313:1929–1935. [PubMed: 17008526]
- [33]. Iorio F, Tagliaferri R, di Bernardo D. Identifying network of drug mode of action by gene expression profiling. *Journal of computational biology : a journal of computational molecular cell biology.* 2009; 16:241–251. [PubMed: 19183001]
- [34]. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences of the United States of America.* 2010; 107:14621–14626. [PubMed: 20679242]
- [35]. Young DW, Bender A, Hoyt J, McWhinnie E, Chirn GW, Tao CY, Tallarico JA, Labow M, Jenkins JL, Mitchison TJ, Feng Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol.* 2008; 4:59–68. [PubMed: 18066055]
- [36]. Ma H, Zhao H. iFad: an integrative factor analysis model for drug-pathway association inference. *Bioinformatics.* 2012; 28:1911–1918. [PubMed: 22581178]
- [37]. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, Doroshow J, Pommier Y. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research.* 2012; 72:3499–3511. [PubMed: 22802077]
- [38]. Pakhomov S, McInnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, Bhise N, Lamba V, Birnbaum AK. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. *Journal of biomedical informatics.* 2012; 45:862–869. [PubMed: 22564551]
- [39]. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P Natl Acad Sci USA.* 2005; 102:15545–15550.
- [40]. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *Bmc Bioinformatics.* 2009; 10
- [41]. Yan XT, Sun FZ. Testing gene set enrichment for subset of genes: Sub-GSE. *Bmc Bioinformatics.* 2008; 9
- [42]. Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics.* 2011; 27:1290–1298. [PubMed: 21414987]

- [43]. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002; 18(Suppl 1):S233–240. [PubMed: 12169552]
- [44]. Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Zhu J, Zhang M, Yang D, Rao S, Wang J. Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. *Bioinformatics*. 2007; 23:2121–2128. [PubMed: 17545181]
- [45]. Nacu S, Critchley-Thorne R, Lee P, Holmes S. Gene expression network analysis and applications to immunology. *Bioinformatics*. 2007; 23:850–858. [PubMed: 17267429]
- [46]. Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC bioinformatics*. 2004; 5:100. [PubMed: 15272936]
- [47]. Rajagopalan D, Agarwal P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*. 2005; 21:788–793. [PubMed: 15509611]
- [48]. Ulitsky I, Karp RM, Shamir R. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles, *Research in Computational Molecular Biology. Proceedings*. 2008; 4955:347–359.
- [49]. Ulitsky I, Shamir R. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*. 2009; 25:1158–1164. [PubMed: 19297352]
- [50]. Qiu YQ, Zhang S, Zhang XS, Chen L. Identifying differentially expressed pathways via a mixed integer linear programming model. *Iet Syst Biol*. 2009; 3:475–486. [PubMed: 19947773]
- [51]. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*. 2008; 24:1223–1231. [PubMed: 18586718]
- [52]. Wang Y, Xia Y. Condition specific subnetwork identification using an optimization model. *Lect Notes Oper Res*. 2008; 9:333–340.
- [53]. Qiu YQ, Zhang SH, Zhang XS, Chen LN. Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC bioinformatics*. 2010; 11
- [54]. Galperin MY, Fernandez-Suarez XM. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res*. 2012; 40:D1–D8. [PubMed: 22144685]
- [55]. Tsui IF, Chari R, Buys TP, Lam WL. Public databases and software for the pathway analysis of cancer genomes. *Cancer informatics*. 2007; 3:379–397. [PubMed: 19455256]
- [56]. Glez-Pena D, Reboiro-Jato M, Dominguez R, Gomez-Lopez G, Pisano DG, Fdez-Riverola F. PathJam: a new service for integrating biological pathway information. *Journal of integrative bioinformatics*. 2010; 7
- [57]. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 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*. 2005; 102:15545–15550. [PubMed: 16199517]
- [58]. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
- [59]. Desai K, Brott D, Hu X, Christianson A. A systems biology approach for detecting toxicity-related hotspots inside protein interaction networks. *Journal of bioinformatics and computational biology*. 2011; 9:647–662. [PubMed: 21976381]
- [60]. Fechete R, Heinzl A, Perco P, Monks K, Sollner J, Stelzer G, Eder S, Lancet D, Oberbauer R, Mayer G, Mayer B. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics. Clinical applications*. 2011; 5:354–366. [PubMed: 21491608]
- [61]. West M. Bayesian factor regression models in the “Large p, Small n” paradigm. *Bayesian Statistics*. 2003; 7:733–742.
- [62]. Hoehndorf R, Dumontier M, Gkoutos GV. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. *Bioinformatics*. 2012; 28:2169–2175. [PubMed: 22711793]

- [63]. Silberberg Y, Gottlieb A, Kupiec M, Ruppin E, Sharan R. Large-scale elucidation of drug response pathways in humans. *Journal of computational biology : a journal of computational molecular cell biology*. 2012; 19:163–174. [PubMed: 22300318]
- [64]. Kasarskis A, Yang X, Schadt E. Integrative genomics strategies to elucidate the complexity of drug response. *Pharmacogenomics*. 2011; 12:1695–1715. [PubMed: 22118053]

Figure 1.
An illustration of metabolic pathways, gene regulation pathways and signal transduction pathways (from <http://www.genome.gov>)

Table 1

Comparison of the three approaches for drug target interaction predictions

	Pros	Cons
Ligand-based	Easily applied to the target prediction for new drugs sharing similar structural or chemical properties with known drugs	It requires a large enough number of known ligands for target proteins of interest, which may be difficult to meet in practice
Target-based	Usually there is rich information on various target proteins	Models and algorithms are not designed for efficient genome-scale computation
Phenotype-based	Genome-scale computation is feasible	May overlook valuable information from other types of data sources

Table 2

Classification of methods for module (aka. sub-network) identification

	Global optimization	Local optimization
Node-based	[43, 48–51]	[45–47]
Edge-based	[44]	/
Combined	[42, 53]	[52]