Published in final edited form as:
Pharmacogenomics. 2011 September ; 12(9): 1349–1360. doi:10.2217/pgs.11.76.

Systems biology data analysis methodology in pharmacogenomics

Andrei S Rodin[†,1], **Grigoriy Gogoshin**[1], and **Eric Boerwinkle**[1,2]

[1]Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, TX 77030, USA

[2]Institute of Molecular Medicine, University of Texas Health Science Center, Houston, TX 77030, USA

Abstract

Pharmacogenetics aims to elucidate the genetic factors underlying the individual's response to pharmacotherapy. Coupled with the recent (and ongoing) progress in high-throughput genotyping, sequencing and other genomic technologies, pharmacogenetics is rapidly transforming into pharmacogenomics, while pursuing the primary goals of identifying and studying the genetic contribution to drug therapy response and adverse effects, and existing drug characterization and new drug discovery. Accomplishment of both of these goals hinges on gaining a better understanding of the underlying biological systems; however, reverse-engineering biological system models from the massive datasets generated by the large-scale genetic epidemiology studies presents a formidable data analysis challenge. In this article, we review the recent progress made in developing such data analysis methodology within the paradigm of systems biology research that broadly aims to gain a 'holistic', or 'mechanistic' understanding of biological systems by attempting to capture the entirety of interactions between the components (genetic and otherwise) of the system.

Keywords

biological networks; data analysis methodology; genome-wide association studies; metabolomics; pharmacogenomics; systems biology

Existence of wide variation in individual responses to pharmacotherapy led to the recent emergence of the research fields of pharmacoge-netics and pharmacogenomics. While the two terms are historically somewhat interchangeable, the latter is commonly associated with the current proliferation of genomic technologies that led to rapid (and continuing) accumulation of the large-scale genome-wide genetic epide-miological datasets. Thus, pharmacogenomics aims to elucidate the genetic factors (genetic variability) behind the individual's response to drug therapy. The five practical questions are: What are the genetic factors that influence the drug efficacy? What are the genetic factors that influence the adverse effects of the therapy? By identifying these genetic factors, can we gain a better understanding of the drug mechanism of action? Can it lead to an existing drug

[†]Author for correspondence: Tel.: +1 713 500 9845, Fax: +1 713 500 0900 andrei.s.rodin@uth.tmc.edu.

Financial & competing interests disclosure
The authors have no other relevant affiliations or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.
No writing assistance was utilized in the production of this manuscript.

improvement, or a new drug design? Armed with the answers, at least partial, to the above questions, can we achieve the ultimate goal of individually tailored or personalized pharmacotheraphy?

It is clear that all these questions/research activities have at least one thing in common: attaining some degree of understanding of how the drug interacts with the genetic and other components of biological systems is highly desirable before we can follow-up with developing new drugs and personalized therapy guidelines.

An example of a typical pharmacogenomic approach is genome-wide association study (GWAS), in which a number of individuals are genotyped for a large number of genome-spanning genetic markers (typically SNPs) in case–control or population-based epidemiological study frameworks. Subsequently, the SNPs significantly associated with a phenotype of interest are used to identify relevant genes using linkage disequilibrium and bioinformatics tools. Some recent examples of drug therapy phenotypes in GWAS include antihypertensive response to thiazide diuretic [1], adverse dermatological reaction to drug therapy [2], modulation of warfarin treatment dosage [3] and anti-depressant treatment outcomes [4], to name just a few. A detailed discussion of the pros and cons of GWAS in the pharmacogenomic context is outside the scope of this article (see [5,6] for a comprehensive discussion); however, two important points emerge: first, even if successful, GWAS can only point to a certain (hopefully sufficiently narrow) part of the genome in which a genetic susceptibility factor resides, without any guarantee that the found SNPs are in fact genetically causative; second, it is very difficult to account for possible gene–gene and gene–environment interactions (let alone gene–other variables), mostly because of the computational and statistical reasons. The first difficulty can be somewhat alleviated through the bioinformatics analysis follow-up, including utilization of available ontology and pathway databases, and we will return to this later; it is the second challenge that is central to our narrative.

Indeed, without data analysis methodology capable of capturing the interactions between the genetic factors, phenotypes and other variables corresponding to different components of biological systems, it would be difficult to address the five goals of pharmacogenomics outlined above. Traditional statistical genetics analysis methodology, being predominantly univariate in its nature, is ill-equipped to deal with the massive genomic data, which was first recognized in the field when the problem of multiple testing (or multiple comparisons) emerged in the early stages of GWAS (see [7] for a recent perspective). The interaction modeling problem is, if anything, even more challenging, because it presents substantial (and theoretically insurmountable) computational difficulties – intuitively, any methodology capable of modeling high-order interactions between the variables simply will not seamlessly scale up to a large number of variables, and a large number of variables is precisely what the modern genomic technologies capture and measure. The issue is exacerbated by the fact that the genome-wide genotyping is increasingly being augmented and supplanted by the newer genomic technologies promising to deliver even larger and more diverse datasets (whole-genome sequencing, exome sequencing, deep resequencing of candidate genes, metabolomics, epigenomics and other omics). The additional difficulty here lies in incorporating different data types (e.g., SNPs from GWAS and metabolite measurements from metabolomics studies) into the same analysis framework, which is something that the traditional parametric statistical analysis methods are not particularly efficient at either. Finally, pharmacogenomics studies usually rely on smaller samples (compared with the typical genetic epidemiology studies), often being limited by the logistics of clinical study designs. Therefore, the ratio of the number of variables to the number of individuals/observations grows even higher, placing an additional strain on the analysis methods (colloquially known as ‘dimensionality curse’ problem).

Systems biology

Systems biology is a relatively novel approach to the study and modeling of biological systems that can be briefly described as looking at a biological system in its entirety, including the (usually complex) network of interactions between the factors and components comprising the system. In this sense, it is antithetical to the reductionist approach (e.g., “let us perturb the value of one variable and see if it causatively changes the value of another variable – if yes, then the first variable influences the second; now, let us take another variable…”).

The important features of the systems biology approach are as follows: first, the biological systems under study are ‘emergent’, in that scrutinizing separate variables/components one-by-one and adding corresponding effects in simplistic manner is unlikely to predict the system behavior as a whole. However, they can be compartmentalized into the smaller subsystems (organisms into organs, tissues, cells, intracel-lular pathways and genes), reflecting the ‘modularity’ of biological systems. This modularity property is intrinsic to the systems that were not designed from scratch to optimize a global performance criterion, but rather evolved naturally following evolution by gene and domain duplication and subsequent sub- and neo-functionalization. In general, biological systems exist (and should be analyzed and modeled) on many hierarchical levels of biological abstraction, exemplified by the genomic, transcriptomic, epigenetic, proteomic and so on, levels in the genetics context. Finally, a concept of ‘biological network’ is central to the field of systems biology – just as we outlined the five principal questions asked in pharmacogenomics research, it could be argued that the main goal and activity of systems biology research is to reconstruct a biological network (reflecting the complex pattern of interactions between the biological components on the relevant hierarchical levels) from the ‘flat’ large-scale ‘omic (at this time, principally genomic) data. Analytically, the emphasis is on the balance of complexity and scalability (inferring high-dimensional interactions from the massive datasets), pursuing causation (as opposed to correlation), mix-and-matching different data types (reflecting corresponding biological components belonging to the different hierarchical levels), and hypothesis generation (rather than hypothesis testing). An overview of a systems biology approach to pharmacoge-netics and pharmacogenomics can be found in Koster *et al.* [8]; Nadeau and Dudley [9] present a useful introduction to the systems biology from the viewpoint of genetics and genomics.

Systems biology paradigm does not change the fact that the ultimate goal of the genetic research remains to establish the causative genotype-phenotype link, and so the systems biology models, such as genetic networks, remain largely phenotype-centric. This, of course, includes intermediate phenotypes on different hierarchical levels, making the systems biology approach especially fitting for pharma-cogenomics, as in a typical pharmacogenomics study there are at least two phenotypes – the trait (disease) of interest, and the drug response.

To summarize, the systems biology paradigm is intuitively appealing, biologically sound and is rapidly becoming a practical necessity with the research community facing the everincreasing onslaught of various omics datasets. Ignoring the systems biology aspect when analyzing such datasets can lead to costly errors of omission. The limiting factor, however, is the relative paucity and immaturity of available data analysis methods and tools that take advantage of the systems biology approach. In the following sections, we review the existing approaches and identify some promising directions in systems biology data analysis methods development. We conclude by outlining the emerging trends in the application of the systems biology data analysis tools to pharmacogenetics and pharmacogenomics.

Systems biology data analysis methodology

Existing systems biology methods historically tend to belong to one or more of the following categories: epistasis (gene–gene interaction) modeling methods, machine learning and data mining methods that typically employ some combination of classification and variable selection, biological network reconstruction methods, and various approaches that utilize prior knowledge (often in form of known biological networks and pathways). Regardless of their genesis, all of them share certain common features (such as primary emphasis on the hypothesis generation rather than hypothesis testing) and issues, the principal one being that of scalability and overfitting, or dealing with the 'dimensionality curse'.

Consider a prototypical GWAS with 1 million SNPs. In order to test their association, one by one, with a phenotype of interest, 1 million uni-variate statistical tests must be performed. This is a computationally trivial task (linear in number of variables, i.e., SNPs), although the multiple testing problem must somehow be addressed. If, however, one wishes to test all pairwise SNP combinations for association, approximately 5×10^{11} tests will have to be performed (approximately 1.7×10^{17} for three SNP combinations and so on). In an idealized systems biology model, all possible combinations of components (variables) should be considered. This presents obvious problems: first, computational intractability ('learning complex models is NP-hard', meaning that it is simply not practically feasible for the large number of variables: see [10] for an example of exact theoretical treatment) and, second, an overwhelming false-positives (specificity) problem, because with the exponential expansion of the model space (that follows from allowing high-order interactions), the number of models that will fit the observed data quite well by chance alone will also increase dramatically.

Other issues that uniquely arise with the systems biology data analysis are visualization, interpretation and validation of the resulting models and, last but not least, efficient handling of the different data types within the same analysis framework. The latter include data storage and conversion, discretization and missing data handling/imputation. One should keep in mind all of the above when evaluating the available methods, tools and software.

Box 1 summarizes a prototypical systems biology data analysis flow – a multistage strategy predominantly focused on reconstructing biological model(s) from the 'flat' omics datasets. Specific data analysis techniques detailed in the following sections often cover more than one step of the process (or skip some steps implicitly); however, this generalized scheme is a useful template for initially approaching newly generated large omics datasets. Importantly, separation into data preprocessing, hypotheses generation and hypotheses testing (validation) is essential for successful analysis of large datasets. While systems biology data analysis methods (and, by extension, this article) are largely involved with the second, hypotheses generation, stage, the first and the third stages should not be ignored.

Modeling epistasis

Before one proceeds with capturing interactions between all biological variables, it would make sense, especially in the context of the genomic studies, to concentrate on the gene–gene (or epistatic) interactions. Then, corresponding methodologies can be expanded to include more diverse data and variable types, with the ultimate goal of being able to handle all kinds of biological interactions. Not surprisingly, analytical methods that can account for the epistatic phenomena were historically some of the earliest examples of systems biology data analysis in genomic research. The importance of epistasis was recognized early on in the GWAS era when it became clear that, at least for the complex or common traits, separate SNP signals often could not explain away much of the trait's heritability. The current paradigm is that the interplay of many genes contributes to the genetic makeup of complex traits, and that the high-order gene-gene interactions might be a rule rather than an exception [11]. This dovetails perfectly with our increasing appreciation of the importance of higher levels of genetic control (epigenetic, transcriptomic and so on). A comprehensive review of epistasis-capturing methods (with primary emphasis on GWAS, and using prior knowledge to alleviate associated computational burden) can be found in [12]. In general, their algorithmic foundation is a variable (SNP) selection 'wrapper' [13], wherein numerous combinations (models) of SNPs are scored based on the strength of association with a phenotype using an explicitly defined genetic model, and the highest-scoring subsets of interacting SNPs are selected via exhaustive (for lower-order interactions) or heuristic (for higher-order iterations) search through the model space. By design, these methods are simultaneously variable selection and classification/estimation techniques (Box 1). Depending on the specific implementation and hardware, these methods are computationally feasible in up to the 500,000–1 million SNPs range. Notable examples include combinatorial partitioning method (CPM) [14], multifac-tor dimensionality reduction and its extensions (model-based, family-based and so on) [15,16], and the set-association method [17].

An interesting feature of the latter is that it explicitly penalizes for the model complexity (the 'model', in general, being the subset of SNPs or other variables and their interactions, and its complexity being directly proportional to the number of variables and interactions it contains). The set-association method does that by increasing degrees of freedom in the statistical association test, but in its generalized form the complexity penalty concept is central to both machine learning and statistical learning fields. It prescribes that not only the model should fit the observed data well, but it should also be as compact as possible. For example, if the two models – one with three SNPs and one with four – explain the same amount of heritability, then the former is to be preferred. The appeal of this computer-science take on the Occam's razor principle to the field of genomics is obvious, as it aids in combating the multiple testing/false-positives problem (known as 'overfitting' in machine learning vernacular).

The principal shortcomings of epistasis-capturing methods (in addition to the obvious difficulties with incorporating non-SNP factors in the analysis framework) are limited scalability and replicability of the resulting models in light of dimensionality curse – too many variables (SNPs) for too few observations (tens of thousands of individuals in a typical genetic epidemiology study). Some useful suggestions for the meta-analysis of epistasis modeling results in GWAS are summarized in [18].

Machine learning methods

Data analysis methods derived from machine learning and artificial intelligence research have been steadily gaining wider acceptance in bioinformatics and genomics over the last two decades. Historically, their primary advantage (and raison d'etre) was ability to perform automated knowledge discovery (data mining) in large datasets, and this translated well to bioinformatics applications when the first high-throughput technologies emerged in the late 1990s [19]. While the terms 'data mining' and 'machine learning' are often used indiscriminately and interchangeably, the former is more of an approach and 'activity', while the latter encompasses a specific group of methods that share common origins in computer science and artificial intelligence research. A detailed treatment can be found in [20,21], respectively.

The two types of machine learning methods relevant for the purposes of this article are variable selection methods [13,22,23] and classification algorithms (or classifiers), the latter being a natural fit for the case–control genetic epidemiology studies. Estimation algorithms play largely the same role in the context of continuous phenotypes, the differences between estimation and classification algorithms being mostly of mathematical rather than conceptual nature. Of course, many epistasis modeling methods, as well as traditional statistical methods such as multivariate logistic regression, are in essence classifiers, so the subdivision of data analysis methods into these groups is necessarily somewhat artificial (similarly, variable selection and classification are also closely related). Instead, it might be more natural to categorize machine learning (and other) methods following the activities listed in Box1.

In variable selection, the ‘irrelevant’ variables (e.g., SNPs in GWAS that do not influence the phenotype of interest) are removed from the analysis using ‘filters’ or ‘wrappers’ (such as some of the epistasis modeling algorithms mentioned earlier). Filters are the simplest variable selection techniques that rely on univariate tests based on the measure of genetic association, prior biological knowledge or some other measure of importance. Traditional statistical methods for multiple testing correction belong to this category. By their nature, such methods are incapable of modeling interactions between the variables, and therefore their appeal in the context of systems biology is very limited. Classification algorithms often have some built-in (‘embedded’) capacity for variable selection and ranking based on the concept of ‘usefulness’ to a classifier – would adding another SNP to a model improve generalization classification accuracy (correct assignment to a ‘case’ or ‘control’ class, for example), and if yes, by how much? Examples of the variable selection approach to genomic data include various combinations of filters, wrappers and embedded (typically, into classifiers) algorithms [24–28].

In our opinion, while undeniably useful, pure variable selection methods are limited in their utility for the systems biology data analysis, as the danger of ‘throwing away the wheat with the chaff’ is inherent to the filter-based methods. As far as the wrapper-based methods are concerned, they are in principle no different from the classifiers and other descriptive and predictive modeling machine learning methods with embedded variable selection capability, the only distinction being that in the wrapper methods the variable selection loop resides outside of the classification algorithm proper. For example, least absolute shrinkage and selection operator (LASSO) can be defined as both a wrapper variable selection method and a classifier, and has been successfully used in genetic epidemiology context as such [25,27,28].

Classification algorithms are both predictive and descriptive modeling methods – not only do they aim to predict the class instance (e.g., case or control) based on the values of potentially predictive variables (e.g., SNPs), but they also build a model capturing the relationships between the variables. With some classification algorithms (logistic regression and artificial neural networks) the descriptive modeling ability is limited, and such algorithms are commonly known as ‘black boxes’. However, the classification algorithms based on decision trees (see [21] for in-depth analysis) express the underlying model in a format (‘if-then’ statements) that is both appealing to the human experts and reflective of the underlying biological relationships. Decision tree-based classifiers have two other key advantages. First, methods that use randomized decision tree ensembles (instead of single decision trees), such as Random Forests [29], repeatedly encounter and evaluate many combinations of variables, and thus are capable of accounting for the variable interaction. However, this property had never been rigorously studied analytically, in contrast to the epistasis modeling methods (see above) that rely on the explicit genetic inheritance models. The other key advantage has to do with the ‘usefulness’ principle, wherein the variables can be ranked by judging how ‘useful’ they are to a classifier in predicting correct class instance. This allows us to ‘mix-and-match’ different variable types (continuous, discrete, genetic, environmental and so on) provided the decision tree algorithm itself can handle them, something of perhaps limited value to the GWAS, but crucially important in the context of systems biology data analysis.

Random Forests, in particular, is a randomized decision tree ensemble that has attractive scalability properties (proportional to the square root of the number of variables) in the approximately 500,000–1 million variables range, which makes it very appealing to GWAS and similar analyses ([26,30,31], see also [32] for a recent overview). Numerous software implementations specifically aimed at the genomic data exist. Other approaches include, to name just a few, classification and regression trees [33,34], logic regression and logic forests [35], and multivariate adaptive regression splines [36,37].

Recent comparative studies of different classification algorithms [38,39] provide important insights into their behavior under the ‘dimensionality curse’ conditions; more empirical/simulation studies of this kind are needed. Similarly, Genetic Analysis Workshop 16 (GAW16) ([40] and references therein) proved very fruitful in comparing various variable selection and classification approaches as applied to the analysis of real standardized GWAS datasets.

Principal shortcomings of classification algorithms are similar to those of epistasis modeling methods – limited scalability and, if one is not careful when estimating generalization classification accuracy, proneness to overfitting (learning random noise from the data). However, classifiers can in principle handle more diverse data types and variables (and, therefore, can reconstruct more general models) than the epistasis modeling methods.

This said, both epistasis modeling and machine learning methods discussed so far have one important aspect in common: there is a single ‘target’ variable (class variable, dependent variable and so on), usually a phenotype of interest, and a (typically very large) number of other potentially predictive variables (SNPs, environmental variables and so on) that, in a linear or a more complex combination, are supposed to predict the value of a target variable. While the methods differ in how well they capture and visualize the interplay of the potentially predictive variables, all of the above approaches can be expressed as a combination of variable selection process and classification (or, in case of continuous phenotypes, estimation). However, in idealized systems biology modeling there is no such thing as a single target or dependent variable – rather, a network of diverse variables (corresponding to biological entities) and their interactions is reconstructed from the data. Of course, while all the variables are ‘equal’, some, like pharmacotherapy response, are ‘more equal than the others’, and it would make sense to concentrate on the subnetworks centered around such variables.

Reverse-engineering biological networks

Reconstruction and visualization of the biological networks from the omics datasets is a central part of the systems biology toolkit. Its main appeal lies with the network format being a natural representation for the biological processes at many levels of biological hierarchy. Network-based descriptive modeling is naturally more flexible than the predominantly classification-based approaches we have reviewed so far; however, that model flexibility comes at a price of increased computational burden and propensity for overfitting. Not surprisingly, the main shared drawback of all biological network reconstruction methods is limited scalability. A variety of methods for reverse-engineering biological networks from the data exist, including the information-theoretic based approaches (integration of pairwise mutual information or entropy measurements) ([41] and references therein), ordinary differential equations (ODE) [42], structural equation [43], Granger causality [44], Bayesian networks (BN) and dynamic Bayesian networks (DBN). In this section, we will focus predominantly on the BN and DBN approaches; however, many of the considerations discussed (scalability, visualization and interpretability) apply equally to all of these techniques. An empirical comparison of some of these methods can be found in [41,45]. While these methods have different mathematical motivations, conceptually they are quite similar: a network of nodes (representing biological objects) is automatically reconstructed from the flat datasets. Intuitively, nodes showing the strongest mutual correlation should be connected in the reconstructed network. Strengths of pairwise connections can be quantified using information-theoretic metrics [41], explicit biological process modeling [42,43], or statistical metrics (BNs and DBNs), and then the optimal pattern of connections (network topology) can be found that fits observed data the best. This framework can be extended by incorporating time (dynamic) or causal [44] dimension, wherein a node is not just ‘correlated’ with another node, but ‘causes’ it. ODE and structural equation modeling methods require explicit formulation of underlying biological processes, while information-theoretic, BN and Granger causality approaches are more agnostic. At this time, it is unclear whether any of these approaches holds distinct advantage in terms of performance (accuracy of network structure reconstruction) – it is likely to be strongly dependent on the actual dataset(s) under consideration – but as one of the more mature methodologies, BN modeling has a distinct edge in both scalability and quantity and quality of available software implementations.

The principal feature of the BN approach is treating the network nodes (corresponding to the biological entities, such as SNPs, genes, proteins, metabolites or phenotypes) as random variables. Accordingly, the edges (connections) in the network represent the dependencies between the variables, and their absences – conditional independencies. The network itself is, therefore, a graphical representation of a joint probability distribution of the random variables. Just like the actual biological networks being modeled, BNs are ‘sparse’ (not everything is connected with everything else), and the task of reverse-engineering a BN from the data can be deconstructed into inferring a BN topological structure (stable network edges) and then propagating the probabilistic inference through the network to estimate its local parameters. After searching through the model space of many possible BN topological structures, the ones fitting the data the best are selected; the model fit being equal, more ‘compact’ (fewer nodes and lower edge density) models are preferred to avoid overfitting. The amount of over-fitting can be controlled by adjusting sensitivity/ specificity balance. The validity and robustness of the reconstructed BN can be ascertained using likelihood ratio-like tests and statistical resampling techniques, such as bootstrapping and cross-validation. BN modeling goes back to the seminal works such as [46,47]; however, only in the last two decades nontrivial BN modeling became computationally feasible (see [48,49] for early BN modeling applications to the genetic data). The theoretical foundation of the BN modeling can be found in [50,51,101]; see [52] for a less formal introduction.

An important aspect of BN modeling is that the nodes in the network can represent different types of biological objects (unlike, for example, uniformly discrete SNP variables in epistasis modeling methods). Therefore, an additional data analysis challenge arises – handling different data types simultaneously. In practice, this is achieved by discretizing the continuous variable values, or by implementing more complex (‘hybrid’) local probability models [48]. Typically, the actual implementations are software-specific and, to our knowledge, there is no ‘universal’ approach available at this time. Of particularly high importance to the field of pharmacogenomics is potential application of BN modeling to the metabolomic data, which invokes dealing with a completely new data type (metabolomic measurements) – continuous variables with largely unknown distribution(s) – that is also dependent on quality control and imputation algorithms built into the data capture platform. Data-type heterogeneity also complicates missing data imputation algorithms. In BNs, advanced imputation algorithms typically take advantage of assessing the immediate network neighborhood of a variable with missing values, in addition to the simpler standard imputation by assignment of an extra variable state, majority or application of proximity rules. Finally, optimizing data storage formats becomes very important when dealing with large heterogeneous datasets – the most computationally efficient algorithms convert all the mixed-type data into numerical single-type data, separate these from annotation, and store the values of variables ‘row-wise’ on the disk in order to achieve the quickest possible access to the ‘column’ (variable) data within a given programming environment (see [53] for a GWAS example).

In general, the scalability of BN reconstruction is limited by two factors. One is the innate complexity of BN models – using local search heuristics (such as hill-climbing) or more sophisticated optimization algorithms (hill-climbing with restarts and simulated annealing) somewhat alleviates the problem, but at the expense of likely achieving suboptimal BN model(s) for any realistic dataset. Recent attempts to scale the BN methodology up to the hundreds of thousands of SNPs [54,55], while promising, essentially reduce the flexibility of the general BN model to the levels of a naive Bayes classifier [56], thus partially sidestepping the problem. The second factor limiting the scalability of BN modeling is the ability of a human expert to ‘grasp’ the reconstructed BN in its entirety. Possible solutions to these scalability, search exhaustiveness and visualization problems include using optimized ‘row-wise’ data storage formats [53] for speeding up the input/output process, using ‘sparse candidate’-type algorithms [102] for preestablishing the edges (or absence thereof) between the SNPs or other common variables before reconstructing the network structure, using highly sparse BN structure priors, heavy parallelization, when possible, and ‘looking-glass’ visualization of the reconstructed BN that allows the human expert to zoom in on network neighborhood(s) of particular variable(s). The latter can be conveniently done by exporting the reconstructed BN in a graph markup language script that is subsequently processed by the graph visualization rendering engine (e.g., GraphViz [103]).

The wealth of recent literature in BN modeling in the context of systems biology illustrates other important aspects of BN reconstruction, software implementation and applications. One intriguing research direction is moving from correlations, or associations, to inferring causality by means of experimental intervention [57,58]. DBNs can be a better fit for modeling time-varying biological systems, and are especially appropriate in dealing with metabolic pathways, gene regulatory networks and gene-expression data [59,60]. A useful DBN modeling tool is introduced in [61]. In addition, DBN-derived graphic duration models can be used in application to survival analysis, which is highly relevant to the pharmacogenomics phenotypes and endpoints. Majority of recent BN modeling applications are centered around large-scale GWAS and gene-expression data-sets, indicating attainment of a certain level of methodological maturity [62–66]; however, scalability will remain a principal challenge in foreseeable future. Fortunately, just as with the systems biology data analysis in general, BN modeling can substantially benefit from merging the purely data-driven approach (BNs derived strictly from the omics datasets, with no prespecified network topological structure priors) with the existing expert, or prior knowledge. Using prior knowledge (including known molecular pathways) to augment the GWAS analysis is extensively discussed in [12,18]; one of the major benefits being increased scalability and specificity. This approach naturally translates to the BN reconstruction: by selectively choosing biologically realistic network topology priors, and by ‘forcing’ or ‘forbidding’ certain edges, one could, in principle, drastically shrink the network model search space. In particular, the choice of biologically-motivated network structure priors before reconstructing the BNs from the datasets can substantially influence the quality of the resulting networks [67–70]. Another intriguing opportunity lies in formally contrasting BNs reconstructed from the data with the literature (ontology)-derived pathways and BNs [104,105]; see also [71] for a recent comprehensive comparison. However, this has not been extensively studied so far (but see [72] for a promising approach). It remains to note that three comprehensive general-purpose BN reconstruction software lists can be found at [106,107]. In addition, specialized R packages that can be used to reconstruct biological regulatory networks using BNs and DBNs are also available [108,109]. Another useful resource is the Dialogue for Reverse Engineering Assessments and Methods (DREAM) project [110].

Conclusion

First attempts to reverse-engineer gene networks from heterogeneous omics (specifically, genome-wide microarray expression) data using BN modeling in the context of pharmacogenomics and drug-related phenotypes (human endothelial cell fenofibrate treatment, *S. cerevisiae* systemic drug response) go back to 2005 [73,74]. This approach successfully led to the automated generation of ‘druggable’ biological networks, genes and pathways, thus directly contributing to achieving the primary goals of pharmacogenomics research (as outlined at the beginning of the article). Subsequent work resulted in much more comprehensive understanding of the pharmacogenomics of antihyperlipidemia drug response [75,76]. At this time, use of BN modeling in pharmacogenomics data analyses is becoming increasingly commonplace [77–79].

Arguably, three of the more important domains within current pharmacogenomics research are personalized cancer pharmacotherapy, toxicogenomics, and explicit incorporation of biochemical (pharmacodynamic–pharmaco-kinetic) principles into genetic models of drug response. It is, therefore, highly indicative that some of the latest research in these areas largely relies on the systems biology paradigm and methodology ([80–82], respectively).

In summary, it is abundantly clear that the systems biology approach can be very fruitful within the domain of pharmacogenomics and pharmacogenetics research. For any given pharmacotherapy phenotype of interest, moving from the captured relevant large-scale ‘omic datasets to the descriptive and predictive models incorporating the interplay of numerous genetic (and other) factors contributing to this phenotype is the first, and critical, step on the road to individualized prescribing and new drug discovery. The main question is: at this time, do we possess the mature data analysis methods and tools to accomplish this task? In our opinion, the answer is (a somewhat guarded) yes.

Future perspective

The three major challenges that we face in development and successful application of the systems biology data analysis methods remain increasing scalability and specificity, incorporating many different data types (and biological hierarchy levels) into the data analysis framework, and seam-lessly incorporating prior biological and expert knowledge. While the latter approach allows us to significantly increase the scalability and specificity of existing model search algorithms, it would be more appropriate to evaluate where we are right now from the more objective perspective (of purely data-driven analysis). As an example of what can be currently accomplished, it took us several days to reconstruct and visualize a BN from a 500,000 SNP GWAS dataset using a high-end eight-core workstation. This is representative of the current crop of systems biology data analysis tools that take advantage of many algorithmic refinements and shortcuts, ranging from more advanced local search and optimization algorithms to parallelization to efficient data storage and memory utilization. It should be noted that modularity of biological systems, reflecting their evolutionary origins, allows us to adopt a bottom-up modeling approach, where we can first concentrate on a number of smaller semi-independent subsystems (networks) and later integrate them into a larger framework (e.g., moving from the SNPs within a gene to a gene interaction network).

Over the last few years, substantial progress has been achieved in both understanding of how, and why, systems biology data analysis should be applied to the data generated by the pharmacogenomic studies, and refining the relevant data analysis algorithms and software. Coupled with the increasing availability of high-performance computing resources, and our ability to harness them more efficiently, it bodes well for the future of systems biology approach to the pharmacogenomics research. Most importantly, we predict increasing availability and variety of software tools that can automatically reverse-engineer biological networks from the large-scale omics data, and increasing number of applications of the aforementioned tools to the real-world pharmacogenomics studies and datasets.

Acknowledgments

The authors are especially grateful to the anonymous reviewers, whose thoughtful suggestions have substantially improved the manuscript.

Andrei S Rodin is supported by grants from PhRMA foundation and NIH 5R03LM009738, 3U01HG004402, 5R01HL072810, 5R01Hl086694, 2U01GM074492, 1R01AI085014 and 5R01HL084099. Grigoriy Gogoshin is supported by grants from PhRMA foundation and NIH 5R03LM009738, 3U01HG004402 and 5R01HL072810. Eric Boerwinkle is supported by grants from NIH 3U01HG004402, 5R01HL072810, 5RC2HL102419, 5R01HL083498 and 5P50GM065509.

Bibliography

Papers of special note have been highlighted as:

- ■ of interest
- ■■ of considerable interest

1. Turner ST, Bailey KR, Fridley BL, et al. Genomic association analysis suggests chromosome 12 locus influencing antihypertensive response to thiazide diuretic. Hypertension. 2008; 52(2):359–365. [PubMed: 18591461]

2. Shen Y, Nicoletti P, Floratos A, et al. Genome-wide association study of serious blistering skin rash caused by drugs. Pharmacogenomics J. 2011 Epub ahead of print. 10.1038/ tpj.2010.84

3. Takeuchi F, McGinnis R, Bourgeois S, et al. A genome-wide association study confirms *VKORC1*, *CYP2C9*, and *CYP4F2* as principal genetic determinants of warfarin dose. PLoS Genet. 2009; 5(3):e1000433. [PubMed: 19300499]

4. Laje G, McMahon FJ. Genome-wide association studies of antidepressant outcome: a brief review. Prog Neuropsychopharmacol Biol Psychiatry. 2010 Epub ahead of print. 10.1016/j.pnpbp 2010.11.031

5■. Motsinger-Reif AA, Jorgenson E, Relling MV, et al. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet Genomics. 2010 Epub ahead of print. Comprehensive review of genome-wide association study pros and cons in the pharmacogenomics context. 10.1097/FPC.0b013e32833d7b45

6■. Bailey KR, Cheng C. Conference Scene: the great debate: genome-wide association studies in pharmacogenetics research, good or bad? Pharmacogenomics. 2010; 11(3):305–308. More informal overview of the pros and cons of genome-wide association study in the pharmacogenomics context. [PubMed: 20235786]

7■. Johnson RC, Nelson GW, Troyer JL, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010; 11:724. Comprehensive treatment of the multiple testing problem. [PubMed: 21176216]

8■. Koster ES, Rodin AS, Raaijmakers JA, Maitland-van der Zee AH. Systems biology in pharmacogenomic research: the way to personalized prescribing? Pharmacogenomics. 2009; 10(6):971–981. Useful introduction to systems biology for pharmacologists. [PubMed: 19530964]

9■. Nadeau JH, Dudley AM. Genetics Systems genetics. Science. 2011; 331(6020):1015–1016. Concise introduction to systems biology for geneticists. [PubMed: 21350153]

10. Nikoloski Z, Grimbs S, May P, Selbig J. Metabolic networks are NP-hard to reconstruct. J Theor Biol. 2008; 254(4):807–816. [PubMed: 18682254]

11. Templeton, A. Epistasis and complex traits. In: Wade, MW.; Brodie, W., III; Wolf, J., editors. Epistasis and the Evolutionary Process. Oxford University Press; Oxford, UK: 2000. p. 41-57.

12■. Ritchie MD. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. Ann Hum Genet. 2011; 75(1):172–182. Comprehensive overview of epistasis capturing methods. [PubMed: 21158748]

13■. Guyon E, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–1182. Comprehensive overview of variable selection techniques.

14. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 2001; 11(3): 458–470. [PubMed: 11230170]

15. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003; 24(2):150–157. [PubMed: 12548676]

16. Cattaert T, Calle ML, Dudek SM, et al. Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. Ann Hum Genet. 2011; 75(1):78–89. [PubMed: 21158747]

17. Ott J, Hoh JJ. Set association analysis of SNP case–control and microarray data. Comput Biol. 2003; 10(3–4):569–574.

18■. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86(1):6–22. Overview of current state-of-the-art statistical data analysis approaches to genome-wide association study. [PubMed: 20074509]

19. Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. ACM SIGKDD. 2003; 5(2):1–5.

20. Hastie, R.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer; NY, USA: 2003.

21. Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach. 2. Prentice Hall; NJ, USA: 2003.

22. Reunanen J. Overfitting in making comparisons between variable selection methods. J Mach Learn Res. 2003; 3:1371–1382.

23. Kohavi R, John G. Wrappers for feature selection. Artific Intelli. 1997; 97:273–324.

24. He Q, Lin DY. A variable selection method for genome-wide association studies. Bioinformatics. 2011; 27(1):1–8. [PubMed: 21036813]

25. Li M, Romero R, Fu WJ, Cui Y. Mapping haplotype-haplotype interactions with adaptive LASSO. BMC Genet. 2010; 11:79. [PubMed: 20799953]

26■. Rodin AS, Litvinenko A, Klos K, et al. Use of wrapper algorithms coupled with a random forests classifier for variable selection in large-scale genomic association studies. J Comput Biol. 2009; 16(12):1705–1718. Software package for SNP variable selection using Random Forests classifier is available directly from the authors upon request. [PubMed: 20047492]

27. D'Angelo GM, Rao DC, Gu CC. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. BMC Proc 15. 2009; 3(Suppl. 7):S62.

28. Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC. Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. Genet Epidemiol. 2011; 35(2):111–118. [PubMed: 21254218]

29. Breiman L. Random Forests. Mach Learn. 2001; 45(1):5–32.

30. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of random forests to a genome-wide association dataset: methodological considerations and new findings. BMC Genet. 2010; 11:49. [PubMed: 20546594]

31. Nonyane BA, Foulkes AS. Application of two machine learning algorithms to genetic association studies in the presence of covariates. BMC Genet. 2008; 9:71. [PubMed: 19014573]

32. Sun YV. Multigenic modeling of complex disease by random forests. Adv Genet. 2010; 72:73–99. [PubMed: 21029849]

33. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and Regression Trees. Wadsworth International Group; Belmont, CA, USA: 1984.

34. García-Magariños M, López-de-Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP–SNP interaction. Ann Hum Genet. 2009; 73(Pt 3):360–369. [PubMed: 19291098]

35. Wolf BJ, Hill EG, Slate EH. Logic forest: an ensemble classifier for discovering logical combinations of binary markers. Bioinformatics. 2010; 26(17):2183–2189. [PubMed: 20628070]

36. Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. Stat Methods Med Res. 1995; 4(3):197–217. [PubMed: 8548103]

37. York TP, Eaves LJ, van den Oord EJ. Multivariate adaptive regression splines: a powerful method for detecting disease-risk relationship differences among subgroups. Stat Med. 2006; 25(8):1355–1367. [PubMed: 16100739]

38■. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinfor. 2010; 11:447. Comparative overview of different classification algorithms in the biological data analysis context.

39. Nicodemus KK, Malley JD. Predictor correlation impacts machine learning algorithms: implications for genomic studies. Bioinformatics. 2009; 25(15):1884–1890. [PubMed: 19460890]

40. Cupples LA, Beyene J, Bickeböller H, et al. Genetic Analysis Workshop 16: Strategies for genome-wide association study analyses. BMC Proc. 2009; 3(Suppl. 7):S1. [PubMed: 20017962]

41. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007; 3:78. [PubMed: 17299415]

42. Liu B, Zhang J, Tan PY, et al. A computational and experimental study of the regulatory mechanisms of the complement system. PLoS Comput Biol. 2011; 7(1):e1001059. [PubMed: 21283780]

43. Das M, Mukhopadhyay S, De RK. Gradient descent optimization in gene regulatory pathways. PLoS One. 2010; 5(9):e12475. [PubMed: 20838430]

44. Zou C, Ladroue C, Guo S, Feng J. Identifying interactions in the time and frequency domains in local and global networks – a granger causality approach. BMC Bioinfor. 2010; 11:337.

45. Cantone I, Marucci L, Iorio F, et al. A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. Cell. 2009; 137(1):172–181. [PubMed: 19327819]

46. Wright S. The method of path coefficients. Ann Math Stat. 1934; 5:161–215.

47. Rao DC, Morton NE. Path analysis of family resemblance in the presence of gene-environment interaction. Am J Hum Genet. 1974; 26(6):767–772. [PubMed: 4440682]

48. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000; 7(3–4):601–620. [PubMed: 11108481]

49■. Rodin A, Mosley TH Jr, Clark AG, Sing CF, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks: application to *APOE* gene variation and plasma lipid levels. J Comput Biol. 2005; 12(1):1–11. Software package for building Bayesian networks (BNs) from genomic data is available directly from the authors upon request. [PubMed: 15725730]

50. Pearl, J. Probabilistic reasoning in intelligent systems. Morgan Kaufmann; San Mateo, CA, USA: 1984.

51. Pearl, J. Causality Models, Reasoning, and Inference. Cambridge University Press; Cambridge, UK: 2000.

52■. Krause PJ. Learning probabilistic networks. Know Eng Rev. 1998; 13(14):321–351. Broad introduction to BN modeling.

53. Nielsen J, Mailund T. SNPFile – a software library and file format for large scale association mapping and population genetics studies. BMC Bioinformatics. 2008; 9:526. [PubMed: 19063732]

54. Jiang X, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. Genet Epidemiol. 2010; 34(6):575–581. [PubMed: 20568290]

55. Jiang X, Neapolitan RE, Barmada MM, Visweswaran S, Cooper GF. A fast algorithm for learning epistatic genomic relationships. AMIA Ann Symp Proc. 2010; 2010:341–345.

56. Friedman ND, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning. 1997; 29:131–163.

57. Pe'er D. Bayesian network analysis of signaling networks: a primer. Sci STKE. 2005; 281:14.

58. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005; 308(5721):523–529. [PubMed: 15845847]

59. Grzegorczyk M, Husmeier D. Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. Bioinformatics. 2011; 27(5):693–699. [PubMed: 21177328]

60. Lèbre S. Inferring dynamic genetic networks with low order independencies. Stat Appl Genet Mol Biol. 2009; 8(1) Article 9.

61. Paluszewski M, Hamelryck T. Mocapy++ – a toolkit for inference and learning in dynamic Bayesian networks. BMC Bioinfor. 2010; 11(126)

62. Bauer S, Gagneur J, Robinson PN. GOing Bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Res. 2010; 38(11):3523–3532. [PubMed: 20172960]

63. Villanueva E, Maciel CD. Modeling associations between genetic markers using Bayesian networks. Bioinformatics. 2010; 26(18):632–637. [PubMed: 20080511]

64. Chu JH, Weiss ST, Carey VJ, Raby BA. Agraphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. BMC Syst Biol. 2009; 3:55. [PubMed: 19473523]

65. Tamada Y, Imoto S, Araki H, et al. Estimating genome-wide gene networks using nonparametric Bayesian network models on massively parallel computers. IEEE/ACM Trans Comput Biol Bioinform. 2011; 8(3):683–697. [PubMed: 20714027]

66. Wang Y, Zhang XS, Xia Y. Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. Nucleic Acids Res. 2009; 37(18):5943–5958. [PubMed: 19661283]

67. Djebbari A, Quackenbush J. Seeded Bayesian networks: constructing genetic networks from microarray data. BMC Syst Biol. 2008; 2:57. [PubMed: 18601736]

68. Keilwagen J, Grau J, Posch S, Grosse I. Apples and oranges: avoiding different priors in Bayesian DNA sequence analysis. BMC Bioinfor. 2010; 11:149.

69. Steele E, Tucker A, Hoen PA, Schuemie MJ. Literature-based priors for gene regulatory networks. Bioinformatics. 2009; 25(14):1768–1774. [PubMed: 19389730]

70. Shah A, Woolf P. Python environment for Bayesian learning: inferring the structure of Bayesian networks from knowledge and data. J Mach Learn Res. 2009; 10:159–162. [PubMed: 20161541]

71. Shmelkov E, Tang Z, Aifantis I, Statnikov A. Assessing quality and completeness of human transcriptional regulatory pathways on a genome-wide scale. Biol Direct. 2011; 6(1):15. [PubMed: 21356087]

72. Chang, R.; Shoemaker, R.; Wang, W. IEEE/ACM Trans Comput Biol Bioinform. 2011. A novel knowledge-driven systems biology approach for phenotype prediction upon genetic intervention. Epub ahead of print

73. Tamada Y, Imoto S, Tashiro K, Kuhara S, Miyano S. Identifying drug active pathways from gene networks estimated by gene expression data. Genome Inform. 2005; 16(1):182–191. [PubMed: 16362921]

74. Imoto S, Tamada Y, Araki H, et al. Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. Pac Symp Biocomput 2006. 2006:559–571.

75. Tamada Y, Araki H, Imoto S, et al. Unraveling dynamic activities of autocrine pathways that control drug-response transcriptome networks. Pac Symp Biocomput. 2009; 14:251–263. [PubMed: 19209706]

76. Araki H, Tamada Y, Imoto S, et al. Analysis of PPARα-dependent and PPARα-independent transcript regulation following fenofibrate treatment of human endothelial cells. Angiogenesis . 2009; 12(3):221–229. [PubMed: 19357976]

77. Himes BE, Wu AC, Duan QL, et al. Predicting response to short-acting bronchodilator medication using Bayesian networks. Pharmacogenomics. 2009; 10(9):1393–1412. [PubMed: 19761364]

78. Lara J, Xia G, Purdy M, Khudyakov Y. Coevolution of the hepatitis C virus polyprotein sites in patients on combined pegylated interferon and ribavirin therapy. J Virol. 2011; 85(7):3649–3663. [PubMed: 21248044]

79. Deforche K, Camacho RJ, Grossman Z, et al. Bayesian network analyses of resistance pathways against efavirenz and nevirapine. AIDS. 2008; 22(16):2107–2115. [PubMed: 18832874]

80■. Gonzalez-Angulo AM, Hennessy BT, Mills GB. Future of personalized medicine in oncology: a systems biology approach. J Clin Oncol. 2010; 28(16):2777–2783. Useful introduction to systems biology for oncologists with a special emphasis on personalized therapy. [PubMed: 20406928]

81. Audouze K, Juncker AS, Roque FJ, et al. Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. PLoS Comput Biol 20. 2010; 6(5):e1000788.

82. Ahn K, Luo J, Berg A, Keefe D, Wu R. Functional mapping of drug response with pharmacodynamic-pharmacokinetic principles. Trends Pharmacol Sci. 2010; 31(7):306–311. [PubMed: 20488563]

Websites

101■■. Heckerman, DA. Tutorial on learning with Bayesian networks Technical report MSR-TR-95–06, Microsoft research. 1995. http://research.microsoft.com/en-us/um/people/heckerman/tutorial.pdf Detailed introduction to BN modeling

102. Friedman, N.; Nachman, I.; Pe'er, D. Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm. Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence. UAI '99; 1999. p. 196-205.www.cs.huji.ac.il/~nirf/Papers/FPN1.pdf

103. GraphViz (Graph Visualization) software. www.graphviz.org

104. Ariadne Pathway Studio pathway analysis software. www.ariadnegenomics.com/products/pathway-studio

105. KEGG Pathway Database. www.genome.jp/kegg/pathway.html

106■. Bayesian networks and Bayessian classifier software. www.kdnuggets.com/software/bayesian.html Comprehensive BN reconstruction software list

107■. Software packages for graphical models/Bayesian networks. www.cs.ubc.ca/~murphyk/Software/bnsoft.html Comprehensive BN reconstruction software list

108■. Empirical Bayes Dynamic Bayesian Network Inference. www.stat.purdue.edu/~arau/EBDBN/ebdbNet.html Specialized R package for modeling and visualizing regulatory networks using BNs and dynamic BNs

109. The q-order partial correlation graph learning software. http://functionalgenomics.upf.edu/software/ qpgraph/index.html

110■. Dialogue for Reverse Engineering Assessments and Methods. http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project The Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative is a multipronged project incorporating annual conferences, challenges and discussions aimed at improving robustness and accuracy of biological network reverse-engineering methodology

Executive summary

Systems biology

- A useful paradigm for 'making biological sense' of large-scale omics datasets, particularly in the pharmacogenomics studies.
- Availability and variety of systems biology data analysis tools is limited and so is their scalability.

Systems biology data analysis methodology

- Diverse group of methods that share one common goal – reconstructing biological model(s) from the 'flat' datasets – and one common concern: how to overcome the 'dimensionality curse' (too many variables by too few observations/individuals).

Modeling epistasis

- A group of well-established statistical genetics methods that can account for gene–gene interactions and are directly applicable to pharmacogenomics studies. However, they are somewhat limited by the explicit genetic models employed.

Machine learning methods

- Can be roughly subdivided into variable selection, classification (estimation) and biological network reconstruction approaches.
- The latter are less limited by the model constraints (i.e., can approximate almost any real biological network), but must pay the price of added computational complexity.

Reverse-engineering biological networks

- This activity is the cornerstone of systems biology data analysis. Many methods are available; however, few of them scale up to the typical modern large-scale omics datasets.

Conclusion

- The systems biology paradigm is becoming increasingly useful (and widely accepted) in pharmacogenetics and pharmacogenomics.
- However, systems biology data analysis method development has not kept pace with the rapid progress in genomic technologies.
- We believe that in the next 5–10 years systems biology data analysis methods, in particular biological network reverse-engineering methods, will become standard tools for pharmacogenomics research.

Box 1

Typical systems biology data analysis flow

Data preprocessing stage

- Data parsing, reformatting, conversion and storage optimization
- Imputation and (if necessary) discretization
- Variable selection
- Filter methods (including traditional statistical multiple testing/comparisons filters)
- Wrapper methods
- Embedded methods

Biological modeling stage (emphasis on hypotheses generation)

- Classification (or estimation, for continuous phenotypes)
- Traditional statistical methods (e.g., multivariate logistic regression)
- Statistical genetics methods (e.g., capturing epistatic interactions)
- Machine learning methods (e.g., decision tree ensembles)
- Reverse-engineering of biological networks (e.g., structural equations, Granger causality, Bayesian networks)
- Visualization of resulting models

Model validation stage (emphasis on hypotheses testing)

- Traditional statistical hypothesis testing (e.g., for SNP–phenotype associations)
- Resampling (bootstrapping and cross-validation)
- Replication (using different, external, datasets)
- Integration with expert and literature-derived knowledge