A survey of current integrative network algorithms for systems biology

Andrew K. Rider[1], Nitesh V. Chawla[1], Scott J. Emrich[1]
1 Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
∗ **E-mail: nvchawla@nd.edu**

Abstract

The goal of systems biology is to gain a more complete understanding of biological systems by viewing all of their components and the interactions between them simultaneously. Until recently, the most complete global view of a biological system was through the use of gene expression or protein-protein interaction data. With the increasing number of high-throughput technologies for measuring genomic, proteomic, and metabolomic data, scientists now have the opportunity to create complex network-based models for drug discovery, protein function annotation, and many other problems. Each technology used to measure a biological system inherently presents a limited view of the system. However, the combination of multiple technologies can provide a more complete picture. Much recent work has studied integrating these heterogeneous data types into single networks. Here we provide a survey of integrative network-based approaches to problems in systems biology. We focus on describing the variety of algorithms used in integrative network inference. Ultimately, the survey of current approaches leads us to the conclusion that there is an urgent need for a standard set of evaluation metrics and data sets in this field.

Keywords: Network inference, Integrative networks, Systems biology

1 Introduction

The history of genetics has been a process of uncovering increasing amounts of complexity and depth in biological systems. In the past, we knew that DNA was transcribed into RNA and then translated to proteins. Our growing knowledge of alternative splicing and other post-transcriptional regulation complicated this view. We knew that transcription factors were the primary regulators of gene expression. This view became complicated by our increasing knowledge of the regulating effect of phosphorylation on transcription factors. Given the complexity of biological systems and the certain knowledge that we do not fully understand fundamental aspects of biology, it is

important to carefully consider how prior knowledge and diverse data types are incorporated into computational models.

As we learn more about genetics, it is becoming increasingly clear that the traits and behaviors of organisms are emergent: they are the product of complex interactions between numerous biological components. In systems biology, networks are used to capture this complexity by modeling an entire biological system. This approach gives scientists a global view of a biological system that can enable further understanding of the nature of human disease as well as new tools to understand the processes driving life [1].

Networks are a versatile tool that have been used to model interactions between numerous different biological concepts. Nodes can be used to represent genes, proteins, metabolites, or any other discrete biological component or concept. Edges in the network may represent the relationship between a gene and a protein, similarity of function between genes, or any other pair of biological concepts. Edges may represent multiple types of relationships simultaneously. Each type of relationship reveals unique information about an organism. For example, protein-protein interaction (PPI) data reveals which proteins can physically interact, but alone it does not impart knowledge about how an organism will react to stimuli. Similarly, gene expression data can reveal how an organism responds to stimuli in terms of the amount of RNA produced but it does not impart any knowledge about the physical mechanisms that cause change in the organisms behavior. Therefore, the key to furthering our understanding of biological systems the integration of diverse data types.

Differences in the underlying architecture of networks can affect their utility. Directed networks such as Bayesian networks or networks that use asymmetric edge weighting metrics implicitly contain some indication of causality [2, 3]. These methods are well suited for making specific inferences about how the effects of a perturbation to one or more genes will propagate through the network. Undirected networks make fewer assumptions about how nodes are connected and are often less computationally demanding to construct but may yield less specific information.

1.1 Contributions

There are a number of review articles that cover network inference. De Smet et al. reviews network inference and integrative methods in the context of how they approach the problem of underdetermination [4]. Sharan et al. reviews several integrative network methods in the context of clustering [5]. Hecker et al. covers network models for time course behavior of gene expression data and integration of heterogeneous data sources. They discuss a wide range of network inference algorithms both within and outside of the context of integrative approaches. They cover the inclusion of previous biological knowledge

such as expected network topology. In terms of the integration heterogeneous data types, they primarily cover Bayesian networks [6]. Califano et al. review a number of integrative networks approaches in terms of the combinations of data used [7]. They describe how different approaches use different combinations of data types to uncover specific relationships in the data. They also address the need for more focus on awareness of context specific regulation in network models. Bebek et al. focus on integrative approaches specifically used for the identification of biomarkers and the betterment of clinical science [8]. In this work, we focus on presenting a wide range of integrative model types and exclusively on the integration of heterogeneous data. Our purpose is to provide a familiarity with the variety of algorithms used for integration in network models.

In Section 2 we discuss some of the most commonly used data types in integrative network models. Section 3 covers the problem of network inference in the abstract. In Sections 3.1 through 3.4 we discuss various approaches to network inference and cover examples of each in some detail. Section 3.1 and 3.2 cover Bayesian and other probabilistic networks. Section 3.3 discusses integration methods based on machine learning techniques. In Section 3.4 we cover techniques that rely on identifying modules in networks and context specific regulatory patterns. Finally, in Section 4 we discuss some of the patterns that emerge from examining the variety of methods discussed in the previous sections. We conclude that there is an urgent need for consensus about how to evaluate and compare models.

2 Data

Precursors to integrative networks used microarray expression data alone to infer regulatory and other types of relationships between genes. Microarrays enable high-throughput measurement of the expression level of genes. Expression levels measure the relative amount of RNA produced from the transcription of genes. RNA levels give some indication about the amount of protein that is expected to be produced. Since proteins are the primary causes of change in a cell, expression data can give an indirect evidence towards answering many different questions in systems biology. Many studies have relied on clustering and network models to identify functionally similar genes or infer regulatory networks based on expression data [9, 10, 11, 12, 13].

Protein-protein interactions provide more direct information in the form of which proteins physically interact. There are many distinct methods to measure PPI, each with different strengths and weaknesses [14]. For example, Affinity Capture-MS protein interactions are determined by using a “bait” protein that is “captured” by a polyclonal antibody or an epitope tag. The associated partner is then identified by mass spectrometry. The co-immunoprecipitation approach isolates a protein with antibodies. Interacting

partner proteins are then detected with western blotting. Like expression data, PPI data is commonly used to cluster or infer networks to identify novel interactions or determine function [15, 5].

Some data types are themselves integrative. The ChIP-chip technique combines microarrays with chromatin immunoprecipitation to allow the identification of protein binding sites on DNA [16]. This is particularly useful for the study of transcription factors (TFs) which are proteins that transcribe DNA into RNA and are thought to play a major role in the regulation of gene expression. Motifs or identifiable strings of DNA can also be located computationally from sequence data to identify potential transcription factor binding sites (TFBS). Expression quantitative trait loci (eQTLs) use genetic variation between individuals in combination with gene expression data to measure the association between expression levels and genotypes. An expression trait refers to the amount of RNA produced by a gene. Each eQTL represents a strong association between a position or locus in the genome and the expression level of a gene. eQTLs describe the relationship between genotype and phenotype and enable inferences about the regulatory interactions between genes [17].

Annotation data can come from many sources and can describe experimentally or computationally derived knowledge such as functions associated with biological components or pathways that components are a part of. The Gene Ontology (GO) keeps curated functional annotations for genes [18]. Annotations exist in a hierarchy such that a gene my have a number of general and specific functions. Pathway information describes the chain of biological components involved in causing some event or fulfilling some function in a cell. Many databases exist to curate pathway and other data types, often for specific organisms [14, 19, 20].

One consideration that affects many data types is the experimental conditions under which measurements are taken. For example, the expression level of genes can change drastically based on environmental and genetic conditions [17, 21]. Common genetic conditions include gene knockout experiments, in which a gene is made inoperative, and chemical or environmental treatments. Measurements may also involve an element of time under a condition or after a treatment.

3 Network Inference

The basic problem of network inference is to create a network that has a meaningful topology. Ultimately this means creating a sparse network in which only important edges are present. This is accomplished in various ways by different algorithms. In the abstract, there are two general types of networks: distance-based networks and probabilistic networks.

Network inference algorithms universally depend on some measure of dependence or distance between biological components. The approach used to calculate edge weight can have a significant effect on what is contained in the resulting network [22, 23]. Mason et al. compared co-expression networks based on Pearson's correlation to co-expression networks based on the absolute value of Pearson's correlation and showed that modules in the signed network are more biologically coherent [24]. Probabilistic network inference faces a similar problem in that conditional probabilities can be calculated in a number of different ways.

The fundamental assumption in relevance networks and other distance-based networks is that relationships between biological components can be accurately ranked in some meaningful way. Once the relationships between all components have been quantified, edges are removed from the network. This results in a sparse network with some meaningful topology that is determined in part by the edge weighting method and in part by the pruning criterion.

Each approach makes different underlying assumptions that can impact the information contained in the network. Relevance networks make inherent assumptions in the choice of weighting method and the pruning approach. The underlying assumption is that the weighting method correctly ranks edges in terms of importance. Zhou et al. use Pearson's correlation to infer a co-expression network for yeast [25]. They use the shortest paths between all nodes in the network to identify functionally related genes. This approach assumes that transitive relationships that are represented in the network may are as important to understand relationships between genes as direct relationships. ARACNE makes the opposite assumption and explicitly disallows triangles in the network, assuming that all triangles contain an indirect relationship that should not be explicitly represented in the network [12].

Other approaches can be described by the category of distance-based networks focus on machine learning techniques such as feature selection and decision trees. MRNET uses a maximum-relevance minimum-redundancy feature selection method to identify important neighbors for every node. After the pairwise mutual information between expression levels of all genes is calculated, edges are effectively pruned by the feature selection algorithm. For each node, the algorithm selects the neighbor with the highest mutual information that has the lowest redundancy with the neighbors already selected. Neighbor selection stops when the score of the next best neighbor is below a threshold [13].

Probabilistic or graphical models represent the dependence between random variables as nodes in a network. Edge weights represent conditional probabilities. This approach naturally captures the noise and stochastic nature of biological data.

3.1 Bayesian networks

Bayesian networks are one of the most commonly used methods of integrating diverse biological data types. They describe biological data as random variables. Using this approach, measurements of a gene's expression levels may be interpreted as samples from a random variable. Relationships in Bayesian networks are directed, reflecting the conditional dependence between variables. As such, they are often interpreted as causal. This interpretation allows Bayesian networks to represent pathways and to be used to predict the effect of perturbations to the system. Bayesian networks can be discrete, continuous, or a mixture of both.

Discrete Bayesian networks model the probability of discrete states. For example, an edge between nodes A and B can indicate the probability that gene B is highly expressed given the state of gene A. Discrete Bayesian networks may require that each node have a prior distribution to represent the possible prior states of the variable. A model relying only on the frequency of observed values may be unable to assign a probability to new observations if they do not fall within the observed range. Discrete Bayesian networks can model relationships in the data relatively concisely with a conditional probability table for each node that lists the probability of each state given the inputs. One drawback is that discretization of the data may lead to information loss. Bayesian networks that use continuous variables rely on conditional probability densities instead of conditional probability tables. Continuous variables may also be modeled using linear conditional densities, in which the conditional density of a node X is dependent on its parents as shown in Equation 1. The equation shows that the conditional density of X given its parents p is linearly dependent on the values of the parents. It is common to use a normal distribution in this approach. Continuous Bayesian networks do not lose information due to discretization but it is more computationally complex to infer the continuous model than the discrete model.

$$P(X|p_1, ..., p_n) = N(\beta_0 + \sum_{i}^{n} \beta_i * p_i, \sigma^2) \tag{1}$$

There are three major steps in Bayesian network inference. First, a structure must be proposed. Second, the parameters or probabilities associated with edges and nodes must be set. Third, networks must be evaluated to determine how well they model the data. These steps are commonly used iteratively to propose a structure and parameters, then evaluate it against further structural changes. This process allows a search through potential Bayesian network models.

Identifying edges in the network is a critical step in Bayesian network inference, as the direction of edges can greatly affect the interpretation of the model. The presence or lack of edges between nodes can also have a large effect as it determines the conditional relationships between variables.

The most straightforward method to infer network structure would be to exhaustively compare every possible network. This approach is prohibitively expensive, as the number of possible networks grows super exponentially with the number of nodes [26]. Practical methods rely on sampling or heuristics to reduce the search space dramatically.

The sparse candidate algorithm relies on simple local statistics such as correlation to identify potential parents for each gene [2]. It greatly reduces the search space by evaluating edges only between a node and its candidate set. The algorithm can then use hill-climbing or a divide and conquer approach to determine edges. Choices made early in the assignment of edges can result in a restricted search space. Therefore, the algorithm iteratively creates a network then updates the candidate parent sets for each node by replacing nodes in node X's candidate set with a transitive relationship with nodes that had a weaker dependency with X.

Sampling methods such as the Metropolis-Hastings algorithm can be used to reduce computational cost of structure learning at the expense of an accurate description of the data. Sampling and other inexact techniques are often used repeatedly and then averaged to form a single network. Alternatively, one model or a few 'good' models can be selected as representative of all possible models. This process is called model selection when one network is chosen or selective model averaging if multiple representative networks are averaged [27].

Model parameters in Bayesian networks are conditional probability distributions or tables. A continuous node may assume that the observed data come from a normal distribution. However, the parameters of the distribution, the mean and standard deviation, may be incorrect. If the assumed distribution or prior is incorrect then the calculated probability of an observed instance and the fit of the network to the data will be incorrect. Parameter fitting is the process of calculating the priors and conditional probabilities in the network.

In Equation 2, D is the data, E is background knowledge, and θ is the model. $p(\theta|E)$ and $p(\theta|D,E)$ are the prior and posterior probability distributions for the model θ, respectively. The prior describes the agreement between the prior knowledge and the network. The posterior describes how well the model fits the observed data. We direct the reader to Heckerman and Needham et al. for a more thorough treatment of parameter fitting and the selection of priors [28, 26].

$$p(\theta|D,E) = \frac{p(\theta|E)p(D|\theta,E)}{p(D|E)} \tag{2}$$

There are two primary ways to include prior knowledge in Bayesian networks. The first is to constrain the edges in the structure learning step. This is a commonly used approach to integrate heterogeneous biological knowledge [29, 30]. The second is to update the priors in an iterative process. Often, a Bayesian network will be inferred and the parameters fitted to one type of

biological knowledge, then priors are updated to take into account additional sources of data iteratively [31, 32].

Zhu et al. use a mixture of constrained and prior-updated techniques to integrate data types into a Bayesian network. They use the sparse candidate algorithm to infer structure in Bayesian networks based on only expression data, based on eQTL data, and based on expression data, eQTL data, TFBS, and PPI data [29]. For each network type, they learned 1000 networks and determined a consensus network that consisted of edges that were present in at least 30% of the networks. Loops were resolved by removing the weakest edge. Prior knowledge gained from eQTL data was incorporated by constraining edge direction such that genes with cis-acting eQTLs (as defined in Doss et. al. 2005) are considered as potential parent nodes for genes with trans-acting eQTLs in the same region of the genome [33]. Representative genes were used to incorporate TFBS and PPI data. They used a set of genes that were determined to be the most strongly associated with a transcription factor to represent each transcription factor in the network. The prior probability that the gene associated with a transcription factor is the parent of other genes that carry the TFBS was proportional to the number of expression traits correlated with the transcription factor's expression levels. The inferred networks were evaluated in terms of predicting functional categories from the Gene Ontology, predicting genes regulated by various transcription factors, and predicting the response of gene expression to gene knockout experiments.

3.2 Other probabilistic networks

While Bayesian networks are a popular approach to integrating diverse data types, there are many other network models that rely on a probabilistic interpretation of the data.

Tu et al. use a stochastic network to integrate PPI, TFBS, phosphorylation, eQTL, and expression data in order to identify causal genes and regulatory pathways [34]. Their model works under the assumption that causal or regulating genes in the network regulate their targets through either direct or indirect affects on the activity of transcription factors. They take into account the possibility that transcription factors can be regulated at the protein level. They also make the common assumption that gene activity correlates with gene expression. Protein-protein interactions are represented in the network as undirected edges, protein phosphorylation and TFBS are represented as directed edges. Each node has a set of transcription factors that bind to it and a set of genes with eQTLs that are candidate regulators. For each node in the network they estimate the likelihood that every neighboring gene is the cause for its expression by calculating Pearson's correlation between the expression level of the two genes. The algorithm determines the causal regulator of gene G by taking random walks without cycles along the edges in

the network until it reaches a candidate eQTL gene. They used this algorithm on subsets of expression data from specific treatments as well as with bootstrapped samples to observe variation in transcription factor activity and account for variation in expression levels. The method was evaluated by comparing predicted relationships against a compendium of gene knock-out expression data.

Lee et al. propose a method to represent functional associations between biological components. They use a Bayesian statistics approach to determine the likelihood that genes are functionally linked based on evidence from heterogeneous data sources [35]. They use microarray data, philogenetic profiles, PPI, functional linkages from text mining, as well as four other data types. Their log-likelihood score compares the frequency of linkages in each data type between genes that share a pathway to the frequency of linkages between genes that do not share a pathway. In Equation 3, $P(L|E)$ is the frequency of linkages (L) in a data type (E) between genes in the same pathway, $\sim P(L\|E)$ is the frequency of linkages between genes in different pathways for the data type. P(L) and $\sim$ P(L) are the total frequency across data types of all linkages between genes sharing a pathway and not sharing a pathway, respectively.

$$LLS = \frac{P(L|E)/ \sim P(L|E)}{P(L)/ \sim P(L)} \tag{3}$$

This method relies on the use of the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway and sub-cellular location data as ground truth data for the calculation of LLS [19]. The use of a common ground truth allows scores for different types of data to be meaningfully compared. The resulting integrative network showed improved accuracy in terms of linking genes that share pathways in the KEGG database over other methods.

Other methods integrate diverse data types and model the stochastic nature of biological systems use hidden Markov models, Markov random fields, and naïve Bayes models for the data [36, 37, 38].

3.3 Statistical and machine learning approaches

Machine learning and statistical approaches are distance based as many provide some confidence or probability that a prediction is correct. They tend to be different from other distance based methods in that the distances are often determined in a supervised manner.

SEREND is a semi-supervised network construction method that integrates TFBS, DNA sequence binding motifs, and gene expression data to predict transcription factor-gene interactions [39]. It uses a logistic regression classifier for expression data and sequence motif data, then combines the two in a hierarchical classification scheme by training a third logistic

regression classifier on the output of the other two classifiers. Features for the classification of expression data were from 455 expression experiments from a compendium of treatment experiments. Each instance corresponded to a gene. Class labels were activated by a transcription factor, repressed by a transcription factor, or not regulated by a transcription factor. The motif classifier used only a single feature to classify genes as regulated by the transcription factor or not regulated by the transcription factor. If the meta-classifier found that there was enough evidence that a non-regulated gene was regulated by a transcription factor, then the algorithm would switch the label from not regulated to regulated and update the weights for all classifiers. This process allows SEREND to iteratively expand its predictions about transcription factor-gene relationships until they converge. SEREND was evaluated in terms of how well it recovered gene targets that were verified in a ChIP-chip data set.

Hwang et al. use a few statistical methods to combine p-values from different data sets [40]. They use an ensemble of Fisher's weighted F, Mudholkar-George's weighted T, and Liptak-Stouffer's weighted Z where the weight is a measure of the relative statistical power for each data set. They determine a combined weight by comparing a hypothetical weight distribution to an observed distribution. The resulting integrative network has a p-value for each node and edge that indicates the confidence that the node or edge belongs in the network. Multiple approaches were tested on simulated data sets, which allowed a comparison on the basis of ground truth data.

The modENCODE Consortium is a group that collects a great deal of diverse data about the model organism *Drosophila* [41]. They use correlated activity patterns from over 700 data sets to define a functional regulatory network. They use logistic regression to classify promoters as active or inactive based on chromatin modification, TFBS, and nucleosome physical properties. The resulting probabilities are used to weight the confidence of each regulatory edge in the network. They evaluated inferred networks based on the enrichment in the network compared to randomized networks of GO terms, correlation of gene expression across time, frequency of protein-protein interactions in the network, and other metrics.

The STRING (Search Tool for the Retrieval of Interacting Genes) database is a collection of data for the understanding of functional interactions among proteins [42]. Interactions in the database come from many curated data sets from multiple organisms as well as from text mining the literature, predicted interactions from gene co-expression and cross-genome homology. Each interaction in the database has a confidence score assigned to it based on benchmarks against a trusted PPI data source, the KEGG database. Each data source is individually benchmarked and them combined in a naïve Bayesian approach by simply multiplying the normalized scores together. Interactions with more support from multiple sources of data will naturally have a higher combined score. STRING is properly a search tool rather than an integrative network inference method. As such, it does not attempt to evaluate the re-

sulting network but provides the ability to alter the data types included, as well as access the raw data.

An alternative approach to modeling heterogeneous data in a single network is to use multiple edge types in what is called a multi-relational network. Davis and Chawla use a multi-relational network approach to make predictions about disease occurrence in patients and study the relationship between diseases and genes [43]. They combine a network of disease co-morbidity data with a network of genes related to each other by their relationship to the same disease. They then use a link prediction method that uses a triad census (counting the occurrences of sets of three nodes with each possible combination of edges) as the basis to predict unknown genetic links. Predicted links were benchmarked against a number of canonical link prediction methods and performance was measured in terms of area under the ROC curve, and the precision-recall curve.

3.4 Modular networks and condition specific regulators

One of the fundamental problems in creating a network model for regulatory interactions in the genome is that the regulatory program of a cell appears to change under different conditions [44]. Network modules can be viewed as discrete groups composed of many types of molecules whose function is separable from other modules. The aggregate expression of these modules may have condition specific regulators. Integrative network approaches to modeling condition specific regulatory networks rely on compendiums of expression data from different experimental conditions and commonly use TFBS, ChIP-chip, or other protein-DNA interaction data [45, 46, 47].

SAMBA integrates heterogeneous data from gene expression, PPI, phenotypic sensitivity, and TFBS sources into a probabilistic bipartite network in order to identify genes with common behavior across experiments [48]. The nodes on one side of the network are genes and the other side are properties of genes or proteins. Weighted edges in the network between node N and property P are interpreted as the probability that node N has property P. Property nodes can indicate anything from interaction with a specific protein to different levels of discretized gene expression. Subgraphs are scored based on the log ratio of the observed topology under two statistical models, a model for the dependency expected in modules and a model for the background dependency. Biclustering is used to identify gene sets that share sets of properties. Modules are evaluated in terms of functional enrichment based on the Gene Ontology. It finds complete bipartite subgraphs with high density by using a hashing technique to find 'seed' nodes and then using a local search to identify other nodes in the module.

DISTILLER is an integrative framework to identify condition-dependent modularity and regulatory relationships [49]. It uses an efficient item set min-

ing algorithm to identify modules. It starts with "seed" modules, consisting of a small number of genes that are co-expressed in a sufficiently large number of conditions and share motifs for the same regulators. Seed modules are expanded to nodes that do not violate the module properties. A drawback of the item set mining approach is that it can be difficult to identify the most interesting modules from the large amount of potentially redundant output. DISTILLER ranks modules by a measure that takes into account how much they help to cover the entire condition space and their redundancy with already ranked modules. DISTILLER was evaluated in terms of precision and recall on a ChIP-chip gold standard data set.

4 Discussion

While there are many benefits to integrating diverse data types, integration of prior knowledge may reinforce bias in network models to the detriment of new discoveries. For example, a number of networks papers have observed that many biological networks appear to have scale-free topology [50, 51]. In response, methods to infer or evaluate networks based on their topology have been developed. Networks inferred using this criterion will systematically overlook possible networks with alternative architectures [52]. There is evidence that this may be happening as many of the observed scale-free topologies in biological networks may not truly be scale free. Clauset et al. showed that the methods used to measure scale-free topology in many preceding studies of biological networks were unable to distinguish between power-law distributions (such as scale-free) and a number of other distinct distributions [53]. Bias may also enter into models through other prior knowledge. For example, Zhu et al. and Tu et al. both constrain their models to use trans-acting eQTLs to constrain edges but the definition of trans acting is different [29, 34].

Network inference methods that are constrained to include edges from PPI, TFBS, eQTL, or other data may reinforce bias in the models as they do not allow room for error in the data. Less constrained approaches avoid this problem but may add a more subtle bias to the model. Many integrative network approaches construct a single network by integrating data based on a single algorithm [29, 30, 39]. As is the case with different types of data, different algorithms contain different biases. Bayesian approaches that create an ensemble or consensus model with Monte Carlo techniques may suffer less from this type of bias but may reduce bias further by use of fundamentally different algorithms.

The problem of evaluation is made extraordinarily difficult in systems biology by the scarcity of ground truth data. Even curated data sets such as PPI data from KEGG that are used to benchmark novel methods are based on uncertain data. The problem of network evaluation has been noted

before in the single data type network inference problem [54]. Marbach et al. propose a unifying approach to the evaluation of network models that includes common evaluation metrics and simulated data. While these are excellent suggestions, the problem is made much more complicated by the diversity of data involved in integrative methods.

Any single type of data presents a one-dimensional view of a biological system. Therefore, evaluation based on a single data type may not be a baseline for the performance of an integrative method. Furthermore, different approaches tend to use different amounts and types of data, making the actual methods themselves very difficult to compare. There are, of course, high-confidence experimentally derived interactions, but it can be difficult to locate and identify them. Databases such as STRING, KEGG, and modENCODE will be critical for the future progress of integrative network models because they provide this service. The creation of a common body of data for evaluation and a standard for evaluation methods for integrative network approaches would allow integrative network algorithms to be truly compared. This in turn could help us to better understand the complex interplay of diverse data types.

References

1. Schadt EE, Friend SH, Shaywitz DA (2009) A network view of disease and compound screening. Nature Reviews Drug Discovery 8: 286–295.
2. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology 7: 601–620.
3. Rao A, Hero AO, States DJ, Engel JD (2007) Using directed information to build biologically relevant influence networks. Computational systems bioinformatics / Life Sciences Society Computational Systems Bioinformatics Conference 6: 145–156.
4. De Smet R, Marchal K (2010) Advantages and limitations of current network inference methods. Nat Rev Micro 8: 717–729.
5. Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Molecular Systems Biology 3.
6. Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R (2009) Gene regulatory network inference: Data integration in dynamic models–A review. Biosystems 96: 86–103.
7. Califano A, Butte A, Friend S, Ideker T, Schadt EE (2011) Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era .
8. Bebek G, Koyutürk M, Price ND, Chance MR (2012) Network biology methods integrating biological data for translational science. Briefings in Bioinformatics .
9. Hanisch D, Zien A, Zimmer R, Lengauer T (2002) Co-clustering of biological networks and gene expression data. Bioinformatics 18: S145–S154.
10. Datta S, Datta S (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19: 459–466.
11. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863–14868.
12. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7: S7+.

13. Meyer P, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics 9: 461+.
14. Christie KR, Hong EL, Cherry JM (2009) Functional annotations for the Saccharomyces cerevisiae genome: the knowns and the known unknowns. Trends in Microbiology 17: 286–294.
15. Sen T, Kloczkowski A, Jernigan R (2006) Functional clustering of yeast proteins from the protein-protein interaction network. BMC Bioinformatics 7: 355+.
16. Aparicio O, Geisberg JV, Sekinger E, Yang A, Moqtaderi Z, et al. (2005) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Current protocols in molecular biology / edited by Frederick M Ausubel [et al] Chapter 21.
17. Jansen R (2001) Genetical genomics: the added value from segregation. Trends in Genetics 17: 388–391.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.
19. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 40: D109–D114.
20. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, et al. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. Nucleic Acids Res 33.
21. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109–126.
22. Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics 18: S231–S240.
23. de Matos Simoes R, Emmert-Streib F (2011) Influence of Statistical Estimators of Mutual Information and Data Heterogeneity on the Inference of Gene Regulatory Networks. PLoS ONE 6: e29279+.
24. Mason M, Fan G, Plath K, Zhou Q, Horvath S (2009) Signed weighted gene coexpression network analysis of transcriptional regulation in murine embryonic stem cells. BMC Genomics 10: 327+.
25. Zhou X, Kao MCC, Hung W (2002) Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A 99: 12783–12788.
26. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2007) A primer on learning in Bayesian networks for computational biology. PLoS computational biology 3: e129+.
27. Maxwell Chickering D, Heckerman D (1997) Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. Machine Learning 29: 181–212.
28. Heckerman D (1995) A tutorial on learning with bayesian networks. Technical report, Microsoft Research, Redmond, Washington.
29. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics 40: 854–861.
30. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2002) Combining location and expression data for principled discovery of genetic regulatory network models. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing : 437–449.
31. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, et al. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. Bioinformatics 19 Suppl 2.
32. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, et al. (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. Proceedings / IEEE Computer Society Bioinformatics Conference IEEE Computer Society Bioinformatics Conference 2: 104–113.

33. Doss S, Schadt EE, Drake TA, Lusis AJ (2005) Cis-acting expression quantitative trait loci in mice. Genome Research 15: 681–691.
34. Tu Z, Wang L, Arbeitman MN, Chen T, Sun F (2006) An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics 22: e489–e496.
35. Lee I, Date SV, Adai AT, Marcotte EM (2004) A Probabilistic Functional Network of Yeast Genes. Science 306: 1555–1558.
36. Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z (2007) Reconstructing dynamic regulatory maps. Molecular Systems Biology 3.
37. Deng M, Chen T, Sun F (2004) An integrated probabilistic model for functional prediction of proteins. J Comput Biol 11: 463–475.
38. Ucar D, Beyer A, Parthasarathy S, Workman CT (2009) Predicting functionality of protein-DNA interactions by integrating diverse evidence. Bioinformatics 25: i137–144.
39. Ernst J, Beg QK, Kay KA, Balázsi G, Oltvai ZN, et al. (2008) A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. PLoS computational biology 4: e1000044+.
40. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, et al. (2005) A data integration methodology for systems biology. Proceedings of the National Academy of Sciences of the United States of America 102: 17296.
41. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science (New York, NY) 330: 1787–1797.
42. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2010) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research 39: D561–D568.
43. Davis DA, Chawla NV (2011) Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks. PLoS ONE 6: e22670+.
44. Segal MR, Dahlquist KD, Conklin BR (2003) Regression approaches for microarray data analysis. Journal of Computational Biology 10: 961–980.
45. Kim H, Hu W, Kluger Y (2006) Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. BMC Bioinformatics 7: 165+.
46. Gao F, Foat B, Bussemaker H (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics 5: 31+.
47. Luscombe NM, Madan Babu M, Yu H, Snyder M, Teichmann SA, et al. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431: 308–312.
48. Tanay A, Sharan R, Kupiec M, Shamir R (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences of the United States of America 101: 2981–2986.
49. Lemmens K, De Bie T, Dhollander T, De Keersmaecker S, Thijs I, et al. (2009) DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. Genome Biology 10: R27+.
50. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. Nature 407: 651–654.
51. van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO reports 5: 280–284.
52. Yip A, Horvath S (2007) Gene network interconnectedness and the generalized topological overlap measure. BMC bioinformatics 8: 22+.
53. Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. SIAM Review 51: 661+.

54. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences 107: 6286–6291.