Briefings in Bioinformatics Advance Access published August 29, 2007

doi:10.1093/bib/bbm038

Current progress in network research: toward reference networks for key model organisms

Balaji S. Srinivasan, Nigam H. Shah, Jason A. Flannick, Eduardo Abeliuk, Antal F. Novak and Serafim Batzoglou

Submitted: 21st May 2007; Received (in revised form): 22nd July 2007

Abstract

The collection of multiple genome-scale datasets is now routine, and the frontier of research in systems biology has shifted accordingly. Rather than clustering a single dataset to produce a static map of functional modules, the focus today is on data integration, network alignment, interactive visualization and ontological markup. Because of the intrinsic noisiness of high-throughput measurements, statistical methods have been central to this effort. In this review, we briefly survey available datasets in functional genomics, review methods for data integration and network alignment, and describe recent work on using network models to guide experimental validation. We explain how the integration and validation steps spring from a Bayesian description of network uncertainty, and conclude by describing an important near-term milestone for systems biology: the construction of a set of rich reference networks for key model organisms.

Keywords: *pathways; reference networks; systems biology; data integration; network alignment; machine learning*

INTRODUCTION

The term 'post-genomic era' became a cliche even before the human genome was sequenced, but it has a definite meaning. It refers to the refocusing of effort on tasks that were insurmountable without the genome as a platform, such as the construction of hybridization probes for every human gene [1] or the phenotyping of knockout strains for every yeast ORF [2]. Many different kinds of these genome-scale datasets are now available [3–9], and each analysis tells the same story: the components of biological systems are not free-floating parts, but are organized into functional modules [10].

Systems biology is the science of quantitatively defining and analyzing these modules [11]. While a continuum of strategies exists [12], it is useful to divide the field into three levels of increasingly detailed modeling: global characterizations of an organism's interactome [13] or metabolome [14], deterministic models of kinetics and diffusion [15, 16], and detailed stochastic models of variation in isogenic cell lines [17]. While global interactome models can be derived from assays on populations of cells, deterministic models require temporally and sometimes spatially [18] resolved data, and stochastic models require even more data in the form of

Corresponding Author. Balaji S. Srinivasan, 318 Campus Drive, Clark Center S251, Stanford, CA 94305, USA. Tel: (650) 380-0695; Fax: (650) 725-1449; E-mail: balajis@stanford.edu

Balaji Srinivasan is a VIGRE postdoctoral fellow in the Stanford University Department of Statistics, where he teaches classes in computational genomics and machine learning. His research interests include data integration, scalable ontologies and population genetics.

Nigam Shah is a Research Scientist in the Medical Informatics Department at Stanford University. His research interests revolve around developing ontology-based approaches to integrate diverse kinds of biological information.

Jason Flannick is a PhD candidate in Computer Science at Stanford University. His research interests include systems biology and the analysis of functional modules.

Eduardo Abeliuk is a PhD candidate in Electrical Engineering at Stanford University. He is currently interested in regulatory motif identification and transcriptional networks.

Antal Novak is a PhD candidate in Computer Science at Stanford University. His current research interests focus on network visualization and alignment.

Serafim Batzoglou received his PhD from MIT in 2000. He is an Assistant Professor of Computer Science at Stanford University and his research focus is computational biology.

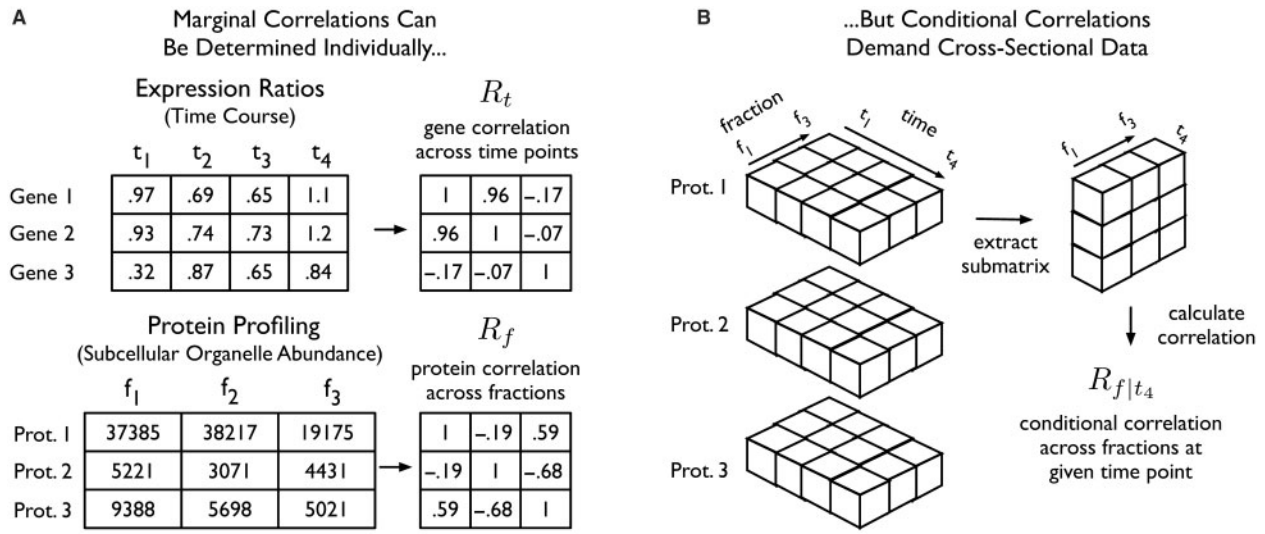


Figure 1: Data availability constrains network detail. (**A**) Given a cell-cycle time course of gene expression measurements, we can determine which genes are temporally coexpressed [35]. Similarly, from protein correlation profiling [4], we can determine which proteins are abundant in the same subcellular organelles, and thereby derive a rough measure of colocalization. (**B**) Suppose, however, that we wish to determine whether a given protein pair is colocalized at a particular time in the cell cycle. To calculate this conditional correlation we must (i) sharply increase the number of data points in our experiment and (ii) collect both kinds of data on the same object at the same time. This may be difficult or impossible to do experimentally; for example, the methods for determining protein abundance across organelles are very different from those for determing an mRNA abundance time series. As more kinds of variables are incorporated (chemical stimuli, genetic background, etc.) the requisite number of data points increases exponentially. These constraints fundamentally limit the extent to which conditional interactions can be probed.

population ensembles. Though there have been some signal successes for both deterministic [19] and stochastic [20] modeling of systems for which many parameters are available, many believe that this area will remain 'data starved' [21] until high-throughput methods for the determination of rate constants [22] and spatial structure [4, 23] become commonplace.

For this reason, we focus here primarily upon the relatively data-rich level of systems biology: the inference and analysis of a global network of interactions for a single organism in which subgraphs of tightly interconnected objects represent functional modules [24]. Some of these networks come from direct measurements of pair-wise interactions [25], including physical [5, 7], signaling [26, 27], transcriptional [28, 29], metabolic [30] and epistatic [3, 31, 32] networks. Other networks are inferred through indirect correlations, including coexpression under the same conditions [8], in the same tissues [33], or at the same time points [34, 35]; coinheritance in the same species [36, 37]; collocation on chromosomes [38]; coevolution of residues [39] or shared mutant phenotype [40]. These indirect networks are constructed by using variation along one dimension (time, space, environmental perturbation, etc.) to inform the construction of the global network. For example, proteins that are abundant in the same subcellular organelles [4] are likely to functionally interact, as are genes that are expressed at the same time [35]; such interacting sets represent subgraphs in the global interaction network.

While it might seem suboptimal to collapse variation in this way, consider the problem of determining a conditional network of interactions or correlations in each subcellular organelle. As Figure 1 shows, this seemingly simple request dramatically increases the amount of data that must be collected. Moreover, in many cases the extra resolution is simply unavailable with current experimental techniques. Microfluidic automation of basic laboratory procedures [41, 42] may make such cross-sectional measurements feasible in the future, but with few exceptions, such as the high-throughput construction and characterization of deletion strains [43], fine-grained conditional data is usually unavailable. Even in large scale

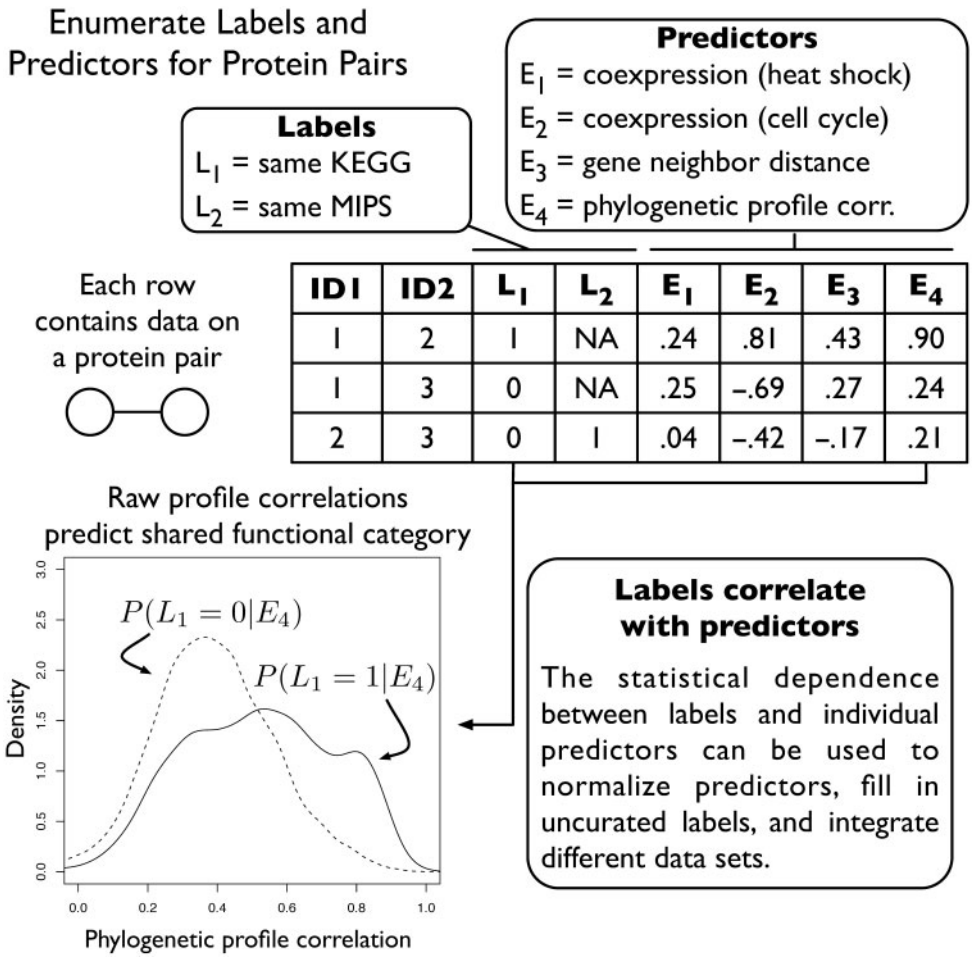


Figure 2: Enumerating labels and predictors for data integration. For each protein pair, we can compute labels and predictors. At the top of the figure, two kinds of labels and four predictors have been tabulated for each pair of proteins; given N proteins, this table will have $N(N-1)/2$ rows. Labels are directly useful to humans while predictors represent raw experimental data. Importantly, many labels correlate with predictors. For example, calculating conditional density estimates (lower left) for the phylogenetic profile correlation over all pairs in *Mycoplasma genitalium* shows that highly coinherited pairs are likely to functionally interact in the same KEGG category [65]. This statistical dependence can be used to put predictors on the same scale, by normalizing them in terms of their ability to recapitulate functional interactions. It can also be used to fill in uncurated labels and integrate different data types (Figure 3).

studies, data is usually collected on only one variable at a time.

Thus, the limitations of the available data tend to force us towards a static 'lowest common denominator' map of interactions for most organisms, averaged over time, space, perturbation and other variables. All is not lost, however, as this static network is still a significant conceptual leap beyond the raw genome sequence of an organism. Moreover, variation of different kinds (e.g. up-regulation of genes or spatial localization of proteins) can be visualized by superimposing tracks and layouts (also see Figure 4) upon such static networks [44, 45], in the same way we view gene and motif tracks upon a genome assembly [46].

Here, we review methods for the inference of these static networks from multiple data sources, along with allied methods for network alignment, network visualization and network-guided experimental prioritization. We then describe a common Bayesian formulation which unifies the steps of network integration and experimental validation (Figures 2 and 3). By analogy to the concept of a reference genome assembly [47, 48], we conclude with a discussion of how recent large scale efforts at network determination, such as the recent Connectivity Map [8] and the proposed

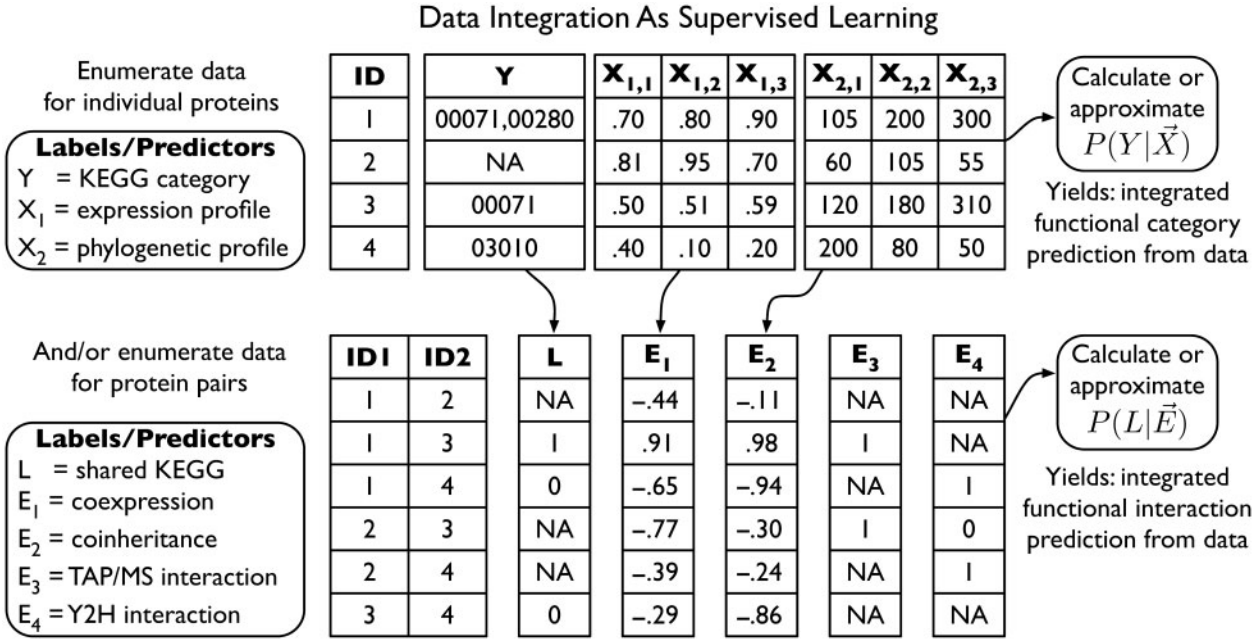


Figure 3: Data integration as supervised learning. For each biological object, we tabulate labels and predictors as in Figure 2. Rather than comparing predictors in terms of their correlation with the label, we use all the predictors at the same time to estimate the label. If we do this for individual proteins, we can obtain an integrative prediction of protein *function*. If instead, we do this for pairs of proteins, we can obtain an integrative prediction of protein *interaction*. Note that some of the columns in the pair table are only defined for pairs (in this case, the TAP/MS and Y2H data), while other quantities can be computed from the protein table. Note also that for statistical reasons, the interaction prediction problem can be easier than the function prediction problem. In the former case, we have a multiclass classification problem with only a few thousand data points, while in the latter case we have a binary classification problem with millions of data points [148]. Importantly, the supervised learning framework can be applied to many other kinds of biological objects besides proteins and protein pairs.

Human Interactome Project [49], can be focused to produce ontologically labeled 'reference networks' (also see Figures 5 and 6).

DATA SOURCES AND NETWORK INTEGRATION

Data sources and data types

Data sources

As hundreds of large-scale datasets are now available, it has become essential to consult meta-databases. Among the most useful are Pathguide [50], BiowareDB [51], BioGRID [52], the yearly Nucleic Acids Research Database [53] and Web Server [54] issues, and a recent compilation of more than 150 publicly available functional genomic resources [55].

Labels versus predictors

For the purposes of data integration, a useful dataset is one that provides measurements on at least one type of biological object, such as genes, proteins or protein pairs (Figure 2). Such datasets can be divided into two broad categories: labels and predictors. Predictors, such as expression ratio measurements on a gene [1] or phylogenetic profiles of a protein [36], are often 'dense' in that they are available for most instances of a biological object and are acquired in a high-throughput way. For example, because most genes are present on standard microarrays, expression profiles are available for most genes (modulo missing values). In contrast, labels such as GO consortium gene annotations [56] or phosphorylation interactions culled from the literature [57] tend to be sparse and of high quality. One of the most important recent discoveries [13] in functional genomics is that these curated labels, which represent directly useful information, can be statistically predicted from combinations of uncurated predictors.

Early methods for clustering and integration

The road to this discovery began with early attempts at unsupervised integration and clustering. When the

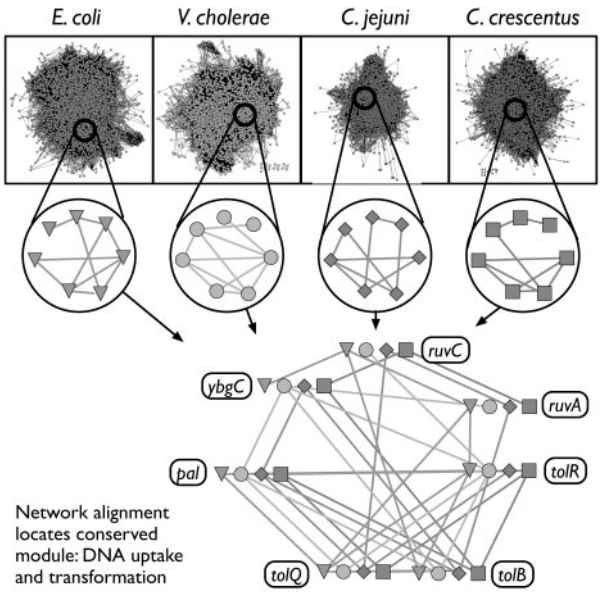


Figure 4: Network alignment. A sample network alignment calculated with the Graemlin algorithm [95]. In the top row, integrated association networks for four microbes are depicted. In these large graphs, nodes represent proteins and edge weights are probabilities of association between proteins. Calculating a global network alignment finds several conserved modules, including one consisting of seven conserved protein families: *ruvC*, *ruvA, tolR, tolB, tolQ, pal and ybgC*. Each family contains four homologous proteins, one in each species; node shape denotes the species of origin and proteins from a given family are grouped near each other. Moreover, the pattern of functional associations between protein families (as revealed by the edges) displays significant conservation. The alignment suggests a possible function for the module: exogenous DNA is allowed into the cell by the *tol*/*exb* membrane channel proteins and then incorporated into the chromosome by the *ruv* recombination proteins. The literature supports this hypothesis, as insertional disruption of *tol*/*exb* family proteins in *Pseudomonas stutzeri* reduces transformational efficiency to 20% of its previous level [149]. This strongly suggests that exogenous DNA travels through these channels before chromosomal incorporation.

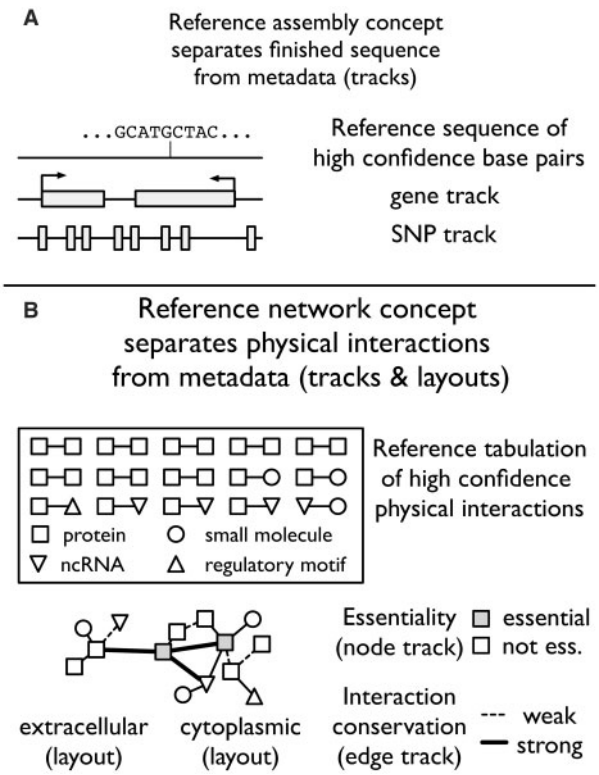
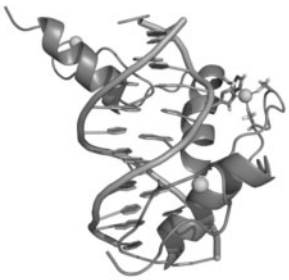


Figure 5: Reference assemblies and reference networks. (**A**) The concept of a reference assembly allows us to enforce a divide between data and metadata. Everything other than finished sequence data is visualized and represented as a metadata track associated with the raw sequence [46]. (**B**) Enforcing a similar kind of separation for a reference network will have key advantages. By enumerating a static list of highly probable physical interactions which occur for an 'average cell' of a given species (averaged over condition, space, time, etc.), we will obtain a lowest common denominator of interaction information to compare between species. Given this physical backbone, metadata can then be visualized via tracks and layouts. For example, we can apply a node track to flag essential nodes, an edge track to highlight strongly and weakly conserved edges and a layout to mirror the known physical separation of modules.

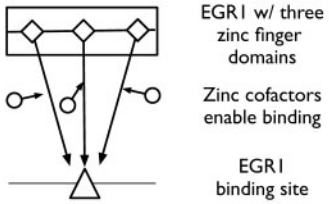
first microarray datasets became available, dozens of different algorithms for unsupervised clustering of these datasets were published [58, 59]. These techniques were also applied to other datasets, such as phylogenetic profiling [36]. While individual clusters of genes were sometimes experimentally validated [37, 60], it was difficult to assess the extent to which any given clustering reflected the 'true' modules of the organism. Given the fuzziness of the module concept, the fact that genes (and other biological objects) can belong to more than one module, and the often conditional nature of intra-module interactions, it was not clear whether the concept of a 'true' set of modules was even a useful one.

This problem became more pronounced when investigators began to combine interaction networks inferred from different assays, which in turn had apparently different modular structures. The first

A EGR1 interaction with DNA (three zinc fingers in complex)



B Schematic of physical interactions



Problem: implicit semantics prevent machine readability

C Represent network as a list of triples (URIs for individual databases)

Subject	Predicate	Object	Note
CID:23994	is_a	MI:0682	Zinc is a cofactor
CID:23994	MI:0407	CDD:pfam00096	zinc directly interacts w/ zinc finger domain
UniProt:P18146	is_a	GO:0003700	EGR1 is a transcription factor
craHsap:197014	is_a	SO:0000235	This motif is a transcription factor binding site
dom:P18146-d1	part_of	UniProt:P18146	first domain in protein
dom:P18146-d1	is_a	CDD:pfam00096	type of domain
...	...	...	...

D Explicit RDF representation (Turtle/N3 format)

```
#Define Namespaces
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix CID: <http://pubchem.ncbi.nlm.nih.gov/> .
@prefix CDD: <http://ncbi.nlm.nih.gov/Structure/cdd/> .
@prefix UniProt: <http://www.ncbi.nlm.nih.gov/entrez/> .
@prefix craHsap: <http://www.cisred.org/> .
@prefix MI: <http://www.psidev.info/> .
@prefix GO: <http://www.geneontology.org/> .
@prefix SO: <http://song.cvs.sourceforge.net/> .

#Begin List of Triples
CID:23994 is_a MI:0682 .
CID:23994 MI:0407 CDD:pfam00096 .
UniProt:P18146 is_a GO:0003700 .
craHsap:197014 is_a SO:0000235 .
dom:P18146-d1 part_of UniProt:P18146 .
dom:P18146-d1 is_a CDD:pfam00096 .

#...more triples below...
```

Figure 6: Network Ontology and RDF representation. Most current networks involve only one or two kinds of biological objects, such as proteins alone [63, 65, 150] or transcription factors and motifs [81]. In order to achieve the ambition of a reference network, however, a notation must be devised for dealing with many kinds of typed interactions. (**A**) As a motivating example, consider the interaction of EGRI with a transcription factor binding site, which involves three zinc finger domains and a zinc cofactor. (**B**) One possible schematic of this interaction is shown, where an individual protein with three domains (top layer) conditionally binds a DNA position (bottom layer) in the presence of zinc (middle layer). The problem is that it is not immediately obvious how to represent this in machine readable terms. (**C**) We believe that the solution lies in representing a network as a list of triples

attempts [61] applied arbitrary thresholds to the interactions derived from different assays and used the union or intersection of these sets as an integrated network. In some cases, such as large-scale yeast two hybrid data, the intersection was essentially the null set [62]. While the idea of combining different assays to reduce noise was a step in the right direction, the problem was that no clear method for weighting the confidence of different assays was available. As with unsupervised clustering, the underlying issue here was the lack of a 'true' set of curated modules to benchmark different assays against.

Data integration by supervised learning

Supervised normalization

The solution [13, 63–68] was to obtain a training set or 'gold standard' of known protein relationships. Different gold standards exist for different kinds of protein relationships; for example, positive examples of colocalized protein pairs can be calculated from MIPS data on protein localization [69], while positive examples of protein pairs in the same functional category can be generated from EcoCyc [70], Reactome [71], GO [72] or KEGG [73]. Negative examples can then be easily generated via random permutations of these positive labels [74]. Though simple, the permutation-based approach for generating negative examples has been shown to be superior to selecting a statistically biased subset of negative examples, such as proteins known to be in different subcellular localizations [74].

For a given gold standard, a useful predictor will separate positive from negative examples (Figure 2). This observed statistical separation can then be converted into a posterior probability by applying Bayes' Rule [65], allowing different predictors (uncurated data) to be compared in terms of their ability to recapitulate known biological labels (curated data). In the specific case of protein interaction prediction, a good predictor will recapitulate known labels by separating interacting protein pairs from non-interacting pairs (Figures 2 and 3).

Detection of corrupted data

One important application of this result is screening microarray experiments for corrupted data [65]. In addition to a battery of internal consistency checks [75, 76], a series of expression measurements can also be used to calculate a correlation matrix, which can then be compared to a training set. If coexpression correlations separate positive and negative training examples as in the lower left panel of Figure 2, the dataset contains at least some signal; if no separation is observed, problems may have occurred with some hybridizations.

Supervised integration

In addition to allowing comparison of different predictors and detecting corrupted data, a gold standard also enables us to perform data integration. In the context of protein interaction prediction, an array of association predictors is the input to a binary classifier function, which returns the integrated probability that two proteins are linked in the sense stipulated by the gold standard (Figure 3). When this binary classifier function is applied to predict interaction probabilities for all protein pairs in a genome, the result is an integrated probabilistic protein interaction network. Variants of this approach have been used to predict functional associations [13, 63, 65], physical contacts [77], synthetically lethal genetic interactions [68] and colocalizations [77, 78].

Importantly, this supervised learning framework for data integration is not limited to interaction

encoded in a 'Network Ontology'. This proposed Network Ontology is a meta-ontology that draws on established ontologies and controlled vocabularies. By combining these source vocabularies, the small set of interactions described in panel (**B**) can be described in terms of a set of unordered triples. Each triple represents a fact about the network, expressed as (subject, predicate, and object) tuple. In general, each member of the triple has its own canonical identifier. For example, the triple (CID:23994, MI:0407, CDD:pfam00096) indicates that zinc (CID:23994 in PubChem) physically interacts (MI:0407 in PSI-MI) with the zinc-finger domain (CDD:pfam00096 in the CDD). For simplicity, we have represented the 'is.a' and 'part.of' predicates as literals, but in general these should also be specified by URIs. For example, the subtleties regarding the Sequence Ontology's 'part.of' definition are treated during the discussion of extensional mereology operators in [151]. (**D**) The advantage of the triple-based representation of the network is that it corresponds to the RDF standard [152] of the W3C consortium. While RDF can be expressed as an XML file, the N3/Turtle notation [153] is far more compact and human readable. Shown is an example of a Turtle format encoding of the triplestore described in panel (**C**). After the preliminary enumeration of namespaces, each non-comment line corresponds to a single triple.

prediction (Figure 3), and has also been applied to direct prediction of protein function [79, 80] and transcription factor/DNA binding [81]. In fact, the concept of using supervised learning to systematize the data integration process has become popular in several other areas of bioinformatics, including gene finding [82], protein sequence alignment [83] and RNA secondary structure prediction [84, 85].

APPLICATIONS OF NETWORK MODELS

In the recent past, network analyses often concluded with a list of modules and an enumeration of topological statistics. Today, now that integrated networks are available for hundreds of organisms [86, 87] the trend is to make network analysis a starting point rather than an ending point, by developing tools for user-friendly network visualization, network-guided experimental validation and network alignment.

Experimental prioritization

Ultimately, an interaction network is a model of the cell, and a model is only useful to the extent that it successfully predicts experiments. In particular, one of the most important ways to leverage network data is not simply to analyze it, but to use it to understand what data to gather next.

One way to formulate this problem is in terms of an 'experiment recommender', which uses network context to prioritize experiments. For example, network context can be used to identify genes that are likely to be in pathways of interest [88]. Experiment recommenders of different kinds have also been used to determine rate constants [89], define metabolic topologies [90], determine disease genes [91] and discern causal structure in signaling pathways [9].

It is important to note that many such recommendation problems can be viewed as updates of an uncertain state variable, such as the GO category of a protein or the value of a rate constant. On a formal basis, this is highly similar to the Bayesian supervised learning model for data integration described in Figure 3, in which a prior gold standard is updated to produce a posterior distribution. There is thus a significant opportunity to unify the problems of data integration and experiment recommendation in a common Bayesian framework, where experiments are recommended, in order of their ability to reduce the uncertainty of state variables of interest.

Network alignment

Once multiple genome sequences became available, research attention naturally turned to the question of comparative genomics [92]. Similarly, the availability of several different kinds of networks from different sources and species has ignited interest in comparative *functional* genomics. Many questions are still open in this area: for example, can we enumerate an organism's inventory of modules much as we can enumerate its inventory of genes? Is it feasible to transfer module annotations from well-studied organisms to newly sequenced ones? And can we identify conserved modules of unknown function?

One promising way of answering such questions is through network alignment, which is a systems-biological analog of sequence alignment. Network alignment allows us to compare interaction networks between different species to find conserved modules. When comparing protein interaction networks, conserved modules are sets of proteins that have both conserved primary sequences and conserved pair-wise interactions between species. For example, we can apply network alignment to find all species with nitrate reduction systems similar to that of *Escherichia coli*, or to examine the extent to which the cell division apparatus is conserved across a set of microbes. A sample alignment is shown in Figure 4; the figure displays a putative DNA uptake and transformation module in which seven protein families across four species show a conserved pattern of functional association [93].

Network alignment has attracted much interest in recent years, beginning with manual alignments of metabolic pathways [94, 95], proceeding to precursors of network alignment guided by best bidirectional BLAST hits [60, 96, 97], and culminating in more recent graph-based formulations [98]. Recent alignment algorithms have introduced the ability to compare three networks at once [99] as well as simple models of network evolution [100]. We recently developed the Graemlin network aligner, which was the first program capable of identifying conserved functional modules across an arbitrary number of dense association networks. By using a number of

BLAST-like optimizations Graemlin's running time scaled linearly, rather than exponentially with the number of species [93].

Just as sequence alignment rests upon substitution matrices [101] and models of sequence evolution [102], we believe that it will be crucial to provide a principled foundation for network alignment by developing a detailed theory of network evolution [103, 104]. Moreover, just as fast algorithms for sequence alignment such as BLAST became ever more essential as sequence data accumulated, it seems clear that the utility of network alignment will rise in direct proportion to the quality of inferred interaction networks in different organisms.

Indeed, the pace of research in this area is accelerating with several papers published in the last few months [105–108]. Part of the reason for this interest is that many of the signal successes of bioinformatics have been concentrated in the area of alignment [109]. Even though the vast majority of objects in biology have not been directly characterized by experimentalists, information on objects which have good digital encodings, like sequences and structures, can easily propagated with an appropriate alignment tool. For example, we can characterize a protein in *Drosophila melanogaster* and immediately BLAST its digital representation to get some clue as to the function of that protein in other insects, or possibly even in humans or yeast.

Yet, the lack of digital representation means that many other interesting objects (like tissues or developmental hierarchies) are not yet easily 'aligned' between organisms. Currently, we resort to simple phylogenetic interpolation to reason that if organism X is phylogenetically equidistant between organism Y and organism Z, then its characteristics are intermediate between these two organisms. However, it is well known that gene trees are not the same as species trees [110–112], and that it is far more accurate to compare genes via sequence alignment. While the divergence of a network tree from the species tree is likely to be less than that of a gene tree (as a collection of genes will have lower sampling variance than an individual gene), nevertheless the same principle holds: the evolutionary history of a module is distinct from that of its host. The promise of network alignment, then, is that we may be to improve upon crude phylogenetic interpolation by directly comparing network models of higher-order processes (such as organs and developmental hierarchies) between species and individuals.

Network visualization

Large interaction datasets with thousands of nodes and edges are best visualized interactively rather than statically. Several tools for this purpose are now available, and can be divided into standalone applications, programming libraries and web applications.

Desktop tools

Among standalone programs, several options are available including Cytoscape [45], Osprey [113], Medusa [114] and Pajek [115]. Cytoscape is a popular choice with many features and plugins, but as it is written in Java it requires large amounts of memory to navigate dense networks. Osprey is similar in functionality, and is somewhat more responsive, but has a smaller user community. Medusa has several novel features, including support for multigraphs with multiple edges between a given pair of nodes. Pajek has many features for mathematical graph analysis but a comparatively steep learning curve.

Programming libraries

Data analysts often wish to dynamically generate network visualizations from within programs, and many libraries for this purpose are available. Cytoscape, mentioned earlier, has an API that can be called from within Java. The Boost Graph Library [116] and AT&T's Graphviz library [117] are open source C+ libraries which have bindings for many different programming languages, including R, Python and Perl.

Online network browsers

Several rich web applications for network visualization have been described in recent years, including STRING [87], PubGene [118], iHOP [119], PSTIING [55] and the Stanford Network Browser [86]. STRING provides several different kinds of interaction predictions between genes for many sequenced genomes. STRING, PubGene and iHOP all allow browsing of literature co-occurrence networks. PSTIING is a powerful data browser that is particularly useful for analysts looking for new datasets to integrate. Finally, the Stanford Network Browser provides access to integrated interaction networks for all sequenced microbes, as well as interfaces for network alignment [93]

and experimental target generation via a protein recommender.

TOWARDS REFERENCE NETWORKS

Now that high-throughput data collection is de rigeur, and algorithms for network integration and comparison have been described, we believe that a feasible near-term goal for systems biology is the construction of static 'reference networks' for key model organisms. It is important to define precisely what is meant by this term. These reference networks should integrate multiple data types (Figure 3), incorporate explicit models of uncertainty, and include ontologically typed edges and nodes. However, as they are meant to represent the 'average cell' of a given organism near the median of the norm of reaction [120], they should not directly incorporate interactions which only occur during certain perturbations, at specific times or within particular cell types. Such conditional interactions should be modeled by superimposing tracks and layouts on the static reference network rather than incorporating conditional interactions directly into the reference network.

Reference networks should exclude conditional interactions

To appreciate why this restriction is useful, a comparison to genome sequencing is appropriate (Figure 5). The concept of a reference assembly is a fiction, but a useful fiction. The genome coils and uncoils [121], moves about the cell [122], is methylated and demethylated [123], varies substantially between individuals [124] and has non-trivial 3D structure [125]. Nevertheless, each of these phenomena can be visualized and analyzed by superimposing tracks upon the reference assembly, which represents a lowest common denominator of analysis. In particular, by separating the raw data (the reference assembly) from the metadata (the species-specific tracks and annotations), cross-species comparisons and genome alignments are enabled [92, 126].

Similarly, by keeping the building blocks of the reference network separate from the details of when or where they interact, we can enforce a separation between data and metadata that will permit powerful kinds of network visualizations and alignments (Figure 5). This is particularly valuable because network metadata is likely to accumulate in bits and pieces due to the prohibitive cost of compiling cross-sectional data on different network states (Figure 1). With respect to visualizing this metadata, the primary new feature in the network context is the availability of layouts in addition to tracks, which are particularly suitable for visualizing spatial or functional relationships (Figure 5B).

Reference networks must include ontological markup

One of the most important lessons learned from genome sequencing was the value of the Gene Ontology's systematic, machine-readable approach to categorizing function [56]. Before GO, it was impossible for a computer to discern that a protein annotated as an 'alcohol dehydrogenase' was a kind of oxidoreductase. We propose that a similar state of affairs is currently prevalent in systems biology, and believe that a Network Ontology for explicit ontological markup of reference networks will prove to be an essential tool (Figure 6).

We envision this Network Ontology as a meta-ontology that derives largely from existing ontologies, something like a more focused analog of the Unified Medical Language System [127] for systems biology. Such an ontology would allow rich kinds of logical and statistical reasoning to be applied in a network context, as exemplified by the Hybrow project [128]. Many of the terms for this Network Ontology can be derived from existing ontologies like the Gene and Sequence Ontology and from lists of canonical identifiers such as those available through Entrez Gene [129], UniProt [130], CDD [131] and PubChem [129]. There are also several available standards in the systems biology space [132] which can serve as building blocks for this project, including SBML [133], CellML [134], BioPax [135] and PSI-MI [136]. Of these ontologies, SBML and CellML are invaluable tools for detailed, time-dependent modeling but may be too granular for genomic scale networks. BioPax and PSI-MI are more appropriate; BioPax was originally developed for exchanging pathway data between databases such as KEGG and Ecocyc, and PSI-MI was built for describing the results of high throughput experiments [137].

By combining these source vocabularies, the Network Ontology will provide a unified framework for defining a reference network and its associated metadata, in terms of lists of triples (Figure 6). Each triple corresponds to a fact about the network, represented as a subject/predicate/object tuple of uniform resource identifiers (URIs).

Each URI represents a canonical identifier drawn from one of the established databases or ontologies. In addition to the vast number of ontological terms compiled by the members of the OBO foundry [138], good URIs currently exist for proteins via UniProt, domains via the CDD, genes via Entrez Gene and small molecules via PubChem. Canonical names are also emerging for ncRNAs [139] and regulatory motifs [140], though a consensus solution will remain elusive until NCBI or EBI launches a database.

Given a consensus set of URIs for biological objects, an explicitly typed reference network can then be naturally represented as a set of ontological triples, such as 'A physically_interacts_with B', or 'X is_a Y', in which canonical URIs are used for each member of the triple (Figure 6). This triple-based representation of a network corresponds to the RDF format of the World Wide Web Consortium [141]. Though originally developed for the Semantic Web (i.e. web page X links to web page Y), a list of triples (also known as a 'triplestore') is clearly also a natural representation for pathway and network information. Importantly, significant progress has already been made by the BioRDF working group [142] towards converting key biological databases into RDF format.

One of the principle advantages of representing network data as an RDF triplestore with canonical URIs for each member of the triple is that if everyone uses the same URIs, then facts produced by different providers can be integrated by forming the union of the two triple stores (though in practice statistical methods will be used to resolve any contradictory triples). Another advantage is that a network in RDF format with explicitly typed nodes and edges can be the subject of non-trivial queries based on the SPARQL query language [141], such as 'find all X's which are regulated by Y' or 'find all signal transduction paths between A and B'; a working example of this kind of query engine can be seen at Pathway Knowledge Base [143]. Finally, a network with explicitly marked nodes and edges suggests natural possibilities for data visualization and enables rich kinds of network alignment.

Reference networks will likely use a Bayesian formulation for update and integration

As depicted in Figure 3, the shared thread behind the supervised learning methods for network integration and protein function prediction is to (i) select a biological object (protein pair, gene pair, protein, etc.), (ii) calculate a list of desired labels and predictive features and (iii) use machine learning to compute a mapping between features and labels. Given sufficient labels and predictors, data on any kind of biological object can be integrated. Some machine learning algorithms, such as LARS [144], include an explicit feature selection step and give information on the leverage of a given predictor on a label of interest. As such they can be used to recommend which kinds of data to collect to reduce the uncertainty in a given label, thereby folding the process of experimental recommendation into the same rigorous Bayesian framework. Recent work [7] has shown the advantages of integrating statistical techniques for scoring interaction confidence with the process of data collection; in the long run such techniques will become as common to network determination as base-calling algorithms [145, 146] have become to sequence determination.

CONCLUSIONS

With hundreds of high-throughput datasets now available, it may seem surprising to note that the current investment in network determination is currently only about 1% of that invested in genome sequencing [147]. However, given the recent push for a Human Interactome Project [49, 147], this may soon change. From an informatics perspective, the prospect of acquiring orders of magnitude more network data in the immediate future will demand more sophisticated tools for data integration, network alignment and ontological markup, which will likely involve ideas similar to the 'reference networks' framework outlined earlier.

Key Points

- The availability of vast quantities of different types of interaction data is pushing systems biology away from unsupervised clustering and towards algorithms for data integration, network alignment and experimental prioritization.
- The supervised learning approach, in which high throughput data is compared against a small training set of curated knowledge, has proven to be the most fruitful data integration strategy to date.
- Supervised predictions of function and interaction from multiple datasets are more robust than those derived from individual datasets, and have provided a foundation for recent work on network alignment and systematic validation.

- Methods for data integration and network comparison will become even more important if the proposed Human Interactome Project becomes a reality.
- An ambitious but feasible challenge for systems biology now is to automatically infer rich reference networks for key model organisms, including *Homo sapiens*, and to show that these reference networks produce previously unattainable experimental predictions.

Funding

B.S.S. was funded by an NSF VIGRE postdoctoral fellowship (NSF grant EMSW21-VIGRE 0502385). E.A. was supported by DOE grant DE-FG02-05ER64136. J.A.F. was funded in part by a Stanford Graduate Fellowship, and S.B., A.F.N. and J.A.F. were supported by NSF grant EF-0312459, NIH grant UO1-HG003162, the NSF CAREER Award and the Alfred P. Sloan Fellowship.

References

1. Schena M, Shalon D, Heller R, *et al*. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996;**93**:10614–9.
2. Winzeler EA, Shoemaker DD, Astromoff A, *et al*. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 1999;**285**: 901–6.
3. Collins S, Miller K, Maas N, *et al*. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 2007; **446**(7137):806–10.
4. Foster L, de Hoog C, Zhang Y, *et al*. A mammalian organelle map by protein correlation profiling. *Cell* 2006; **125**:187–99.
5. Gavin A-C, Aloy P, Grandi P, *et al*. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006; **440**(7084):631–6.
6. Kim JK, Gabel HW, Kamath RS, *et al*. Functional genomic analysis of RNA interference in C. elegans. *Science* 2005; **308**:1164–7.
7. Krogan N, Cagney G, Yu H, *et al*. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006;**440**(7084):637–43.
8. Lamb J, Crawford ED, Peck D, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**: 1929–35.
9. Sachs K, Perez O, Pe'er D, *et al*. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005;**308**:523–9.
10. Hartwell LH, Hopfield JJ, Leibler S, *et al*. From molecular to modular cell biology. *Nature* 1999;**402**(6761 Suppl): C47–52.
11. Bornholdt S. Less is more in modeling large genetic networks. *Science* 2005;**310**:449–51.
12. Ideker T, Winslow LR, Lauffenburger DA. Bioengineering and systems biology. *Ann Biomed Eng* 2006;**34**:1226–33.
13. Jansen R, Yu H, Greenbaum D, *et al*. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003;**302**:449–53.
14. Duarte NC, Becker SA, Jamshidi N, *et al*. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* 2007; **104**(6):1777–82.
15. Aldridge BB, Burke JM, Lauffenburger DA, *et al*. Physicochemical modelling of cell signalling pathways. *Nat Cell Biol* 2006;**8**:1195–203.
16. Alon U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. (Chapman & Hall/CRC Mathematical and Computational Biology Series). Boca Raton: Chapman & Hall/CRC, 2006.
17. Wilkinson DJ. *Stochastic Modelling for Systems Biology*. Boca Raton: Chapman Hall/CRC, 2006.
18. Meinhardt H, de Boer P. Pattern formation in *Escherichia coli*: a model for the pole-to-pole oscillations of Min proteins and the localization of the division site. *PNAS* 2001;**98**:14202–7.
19. Alon U, Surette MG, Barkai N, *et al*. Robustness in bacterial chemotaxis. *Nature* 1999;**397**:168–71.
20. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 1998;**149**:1633–48.
21. Albeck J, Macbeath G, White F, *et al*. Collecting and organizing systematic sets of protein data. *Nat Rev Mol Cell Biol* 2006;**7**:803–12.
22. Famili I, Mahadevan R, Palsson B. k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J* 2005;**88**:1616–25.
23. Schubert W, Bonnekoh B, Pommer A, *et al*. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol* 2006;**24**:1270–8.
24. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**: 101–13.
25. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev* 2007;**21**:1010–24.
26. Pokholok DK, Zeitlinger J, Hannett NM, *et al*. Activated signal transduction kinases frequently occupy target genes. *Science* 2006;**313**:533–6.
27. Ptacek J, Snyder M. Charging it up: global analysis of protein phosphorylation. *Trends Genet* 2006;**22**:545–54.
28. Davidson EH, Rast JP, Oliveri P, *et al*. A genomic regulatory network for development. *Science* 2002;**295**:1669–78.
29. Wei CL, Wu Q, Vega VB, *et al*. A global map of p53 transcription-factor binding sites in the human genome. *Cell* 2006;**124**:207–19.
30. Covert MW, Knight EM, Reed JL, *et al*. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;**429**:92–6.
31. Schuldiner M, Collins S, Thompson N, *et al*. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 2005;**123**:507–19.
32. Tong AH, Evangelista M, Parsons AB, *et al*. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;**294**:2364–8.

33. Chen X, Wu JM, Hornischer K, *et al*. TiProD: the tissue-specific promoter database. *Nucleic Acids Res* 2006;**34**(Database issue):D104–7.
34. Laub MT, McAdams HH, Feldblyum T, *et al*. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 2000;**290**:2144–8.
35. Spellman PT, Sherlock G, Zhang MQ, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998;**9**:3273–97.
36. Pellegrini M, Marcotte EM, Thompson MJ, *et al*. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999;**96**:4285–8.
37. Srinivasan B, Caberoy N, Suen G, *et al*. Functional genome annotation through phylogenomic mapping. *Nat Biotechnol* 2005;**23**:691–8.
38. Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999;**96**:2896–901.
39. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* 2001;**14**: 609–14.
40. Dudley Ae, Janse D, Tanay A, *et al*. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol Systems Biol* 2005;**1**msb4100004-E4100001-msb4100004-E4100011.
41. Demello A. Control and detection of chemical reactions in microfluidic systems. *Nature* 2006;**442**:394–402.
42. Hansen C, Quake SR. Microfluidics in structural biology: smaller, faster em leader better. *Curr Opin Struct Biol* 2003;**13**: 538–44.
43. Giaever G, Chu AM, Ni L, *et al*. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002;**418**:387–91.
44. Hu Z, Mellor J, Wu J, *et al*. Towards zoomable mu maps of the cell. *Nat Biotechnol* 2007;**25**:547–54.
45. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
46. Kuhn RM, Karolchik D, Zweig AS, *et al*. The UCSC genome browser database: update 2007. *Nucleic Acids Res* 2007;**35**(Database issue):D668–73.
47. Lander ES, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
48. Venter JC, Adams MD, Myers EW, *et al*. The sequence of the human genome. *Science* 2001;**291**:1304–51.
49. Ideker T, Valencia A. Bioinformatics in the human interactome project. *Bioinformatics* 2006;**22**:2973–4.
50. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. *Nucleic Acids Res* 2006;**34**(Database issue): D504–6.
51. Matthiessen MW. BioWareDB: the biomedical software and database search engine. *Bioinformatics* 2003;**19**:2319–20.
52. Stark C, Breitkreutz BJ, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**(Database issue):D535–9.
53. Galperin MY. The molecular biology database collection: 2007 update. *Nucleic Acids Res* 2007;**35**(Database issue): D3–4.
54. Fox JA, McMillan S, Ouellette BF. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res* 2006; **34**(Web Server issue):W3–5.
55. Ng A, Bursteinas B, Gao Q, *et al*. pSTIING: a 'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. *Nucleic Acids Res* 2006; **34**(Database issue):D527–34.
56. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 2000;**25**:25–9.
57. Saric J, Jensen LJ, Ouzounova R, *et al*. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* 2005;**22**(6):645–50.
58. Altman RB, Raychaudhuri S. Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol* 2001;**11**:340–7.
59. Sherlock G. Analysis of large-scale gene expression data. *Curr Opin Immunol* 2000;**12**:201–5.
60. Stuart J, Segal E, Koller D, *et al*. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;**302**:249–55.
61. Tong AH, Drees B, Nardelli G, *et al*. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 2002;**295**:321–4.
62. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;**98**:4569–74.
63. Lee I, Date SV, Adai AT, *et al*. A probabilistic functional network of yeast genes. *Science* 2004;**306**:1555–8.
64. Lu LJ, Xia Y, Paccanaro A, *et al*. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res* 2005;**15**:945–53.
65. Srinivasan B, Novak A, Flannick J, *et al*. Integrated Protein Interaction Networks for 11 Microbes. In: *Proceedings of the Tenth Annual International Conference on Computational Molecular Biology* (RECOMB 2006). Venice, Italy 2006;1–14.
66. Tanay A, Sharan R, Kupiec M, *et al*. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* 2004;**101**:2981–6.
67. Troyanskaya OG, Dolinski K, Owen AB, *et al*. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci USA* 2003;**100**: 8348–53.
68. Wong SL, Zhang LV, Tong AH, *et al*. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* 2004;**101**:15682–7.
69. Mewes HW, Heumann K, Kaps A, *et al*. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 1999;**27**:44–8.
70. Karp P, Riley M, Saier M, *et al*. The EcoCyc Database. *Nucleic Acids Res* 2002;**30**:56–8.
71. Vastrik I, D'Eustachio P, Schmidt E, *et al*. Reactome: a knowledgebase of biological pathways and processes. *Genome Biol* 2007;**8**:R39.
72. Harris MA, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**(Database issue):D258–61.

73. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(Database issue):D354–7.
74. Ben-Hur A, Noble WS. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics* 2006;**7**(Suppl 1):S2.
75. Irizarry R, Warren D, Spencer F, *et al*. Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005;**2**: 345–50.
76. Woo Y, Affourtit J, Daigle S, *et al*. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech* 2004;**15**:276–84.
77. Qi Y, Bar-Joseph Z, Klein-Seetharaman J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: struct, Funct, Bioinf* 2006;**63**:490–500.
78. Jansen R, Lan N, Qian J, *et al*. Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics* 2002;**2**:71–81.
79. Han J-D, Bertin N, Hao T, *et al*. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 2004;**430**: 88–93.
80. Lu P, Szafron D, Greiner R, *et al*. PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Res* 2005; **33**(Database issue):D147–53.
81. Beyer A, Workman C, Hollunder J, *et al*. Integrated assessment and prediction of transcription factor binding. *PLoS Computat Biol* 2006;**2**:e70.
82. Ratsch G, Sonnenburg S, Srinivasan J, *et al*. Improving the *Caenorhabditis elegans* genome annotation using machine learning. *PLoS Comput Biol* 2007;**3**:e20.
83. Do C, Gross S, SB. CONTRAlign: Discriminative Training for Protein Sequence Alignment. *Proceedings of the Tenth Annual International Conference on Computational Molecular Biology* (RECOMB 2006). Venice, Italy 2006;160–4.
84. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 2006;**22**(14):e90–8.
85. Gruber AR, Neubock R, Hofacker IL, *et al*. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res* 2007;**35**(Web Server issue):W335–8.
86. Srinivasan B, Novak AF, Flannick JA, *et al*. The Stanford Network Browser. http://networks.stanford.edu/ (18 May 2007, date last accessed).
87. von Mering C, Jensen LJ, Kuhn M, *et al*. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 2007;**35**(Database issue): D358–62.
88. Owen A, Stuart J, Mach K, *et al*. A gene recommender algorithm to identify coexpressed genes in *C. elegans*. *Genome Res* 2003;**13**:1828–37.
89. Flaherty P, Jordan M, Arkin A. Robust Design of Biological Experiments. *Proceedings of the Neural Information Processing Symposium,* 2005.
90. Barrett C, Palsson B. Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS Computat Biol* 2006;**2**:e52.
91. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;**24**:537–44.
92. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.
93. Flannick J, Novak A, Srinivasan BS, *et al*. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res* 2006;**16**:1169–81.
94. Dandekar T, Schuster S, Snel B, *et al*. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J* 1999;**343**(Pt 1):115–24.
95. Forst CV, Schulten K. Phylogenetic analysis of metabolic pathways. *J Mol Evol* 2001;**52**:471–89.
96. Ogata H, Fujibuchi W, Goto S, *et al*. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res* 2000; **28**:4021–8.
97. Yu H, Luscombe NM, Lu HX, *et al*. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 2004;**14**:1107–18.
98. Kelley BP, Sharan R, Karp RM, *et al*. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci USA* 2003;**100**:11394–9.
99. Sharan R, Suthram S, Kelley RM, *et al*. From the cover: conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* 2005;**102**:1974–9.
100. Koyuturk M, Kim Y, Subramaniam S, *et al*. Detecting conserved interaction patterns in biological networks. *J Comput Biol* 2006;**13**:1299–322.
101. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. *Proteins* 1993;**17**:49–61.
102. Durbin R, Eddy S, Krogh A, *et al*. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1999.
103. Berg J, Lassig M. Cross-species analysis of biological networks by Bayesian alignment. *Proc Natl Acad Sci USA* 2006;**103**:10967–72.
104. Weitz J, Benfey P, Wingreen N. Evolution, interactions, and biological networks. *PLoS Biol* 2007;**5**:e11.
105. Li Z, Zhang S, Wang Y, *et al*. Alignment of molecular networks by integer quadratic programming. *Bioinformatics* 2007;**23**(13):1631–9.
106. Liang Z, Xu M, Teng M, *et al*. Comparison of protein interaction networks reveals species conservation and divergence. *BMC Bioinformatics* 2006;**7**:457.
107. Singh R, Xu J, Berger B. Pairwise Global Alignment of Protein Interaction Networks by Matching Neighborhood Topology. *Proceedings of the 11th Annual International Conference on Computational Molecular Biology,* (RECOMB 2007).
108. Stumpf MPH, Kelly WP, Thorne T, *et al*. Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol Evol* 2007;**22**(7):366–73.
109. Batzoglou S. The many faces of sequence alignment. *Brief Bioinform* 2005;**6**:6–22.
110. Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. *PLoS Genet* 2006;**2**(5):e68.
111. Nichols R. Gene trees and species trees are not the same. *Trends Ecol Evol* 2001;**16**:358–64.
112. Pamilo P, Nei M. Relationships between gene trees and species trees. *Mol Biol Evol* 1988;**5**:568–83.

113. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biol* 2003;**4**:R22.

114. Hooper SD, Bork P. Medusa: a simple tool for interaction graph analysis. *Bioinformatics* 2005;**21**:4432–3.

115. de Nooy W, Mrvar A, Batagelj V. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.

116. Siek J, Lee L, Lumsdaine A. The Boost Graph Library. http://www.boost.org/libs/graph/doc/index.html (18 May 2007, date last accessed).

117. Ellson J, North S. Graphviz: Graph Visualization Software. http://www.graphviz.org/ (18 May 2007, date last accessed).

118. Jenssen TK, Laegreid A, Komorowski J, *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;**28**:21–8.

119. Fernandez J, Hoffmann R, Valencia A. iHOP web services. *Nucleic Acids Res* 2007;**35**(Web Server issue):W21–6.

120. Lynch M, Walsh B. Genetics and Analysis of Quantitative Traits. *Sinauer Associates,* 1998.

121. Champoux JJ. DNA topoisomerases: structure, function, and mechanism. *Annu Rev Biochem* 2001;**70**:369–413.

122. Riddihough G. Chromosomes through space and time. *Science* 2003;**301**:779.

123. Weber M, Schubeler D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol* 2007;**19**(3):273–80.

124. Abecasis G, Tam P, Bustamante C, *et al*. Human Genome Variation 2006: emerging views on structural variation and large-scale SNP analysis. *Nat Genet* 2007;**39**:153–5.

125. SantaLucia J, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 2004;**33**: 415–40.

126. Brudno M, Do CB, Cooper GM, *et al*. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 2003;**13**:721–31.

127. Unified Medical Language System (UMLS) 2007. http://umlsks.nlm.nih.gov/ (18 May 2007, date last accessed).

128. Racunas SA, Shah NH, Albert I, Fedoroff NV. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 2004;**20**:257–64.

129. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2007;**35**(Database issue):D5–12.

130. Mulder NJ, Apweiler R, Attwood TK, *et al*. New developments in the InterPro database. *Nucleic Acids Res* 2007;**35**(Database issue):D224–8.

131. Marchler-Bauer A, Anderson JB, Derbyshire MK, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007;**35**(Database issue): D237–40.

132. Stromback L, Lambrix P. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics* 2005;**21**:4401–7.

133. Hucka M, Finney A, Sauro HM, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.

134. Nielsen P, Halstead M. The evolution of CellML. *Conf Proc IEEE Engl Med Biol Soc* 2004;**7**:5411–4.

135. Luciano JS. PAX of mind for pathway researchers. *Drug Discov Today* 2005;**10**:937–42.

136. Orchard S, Hermjakob H, Taylor CF, *et al*. Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17–20 April 2005). *Proteomics* 2005;**5**:3552–5.

137. Hermjakob H, Montecchi-Palazzi L, Bader G, *et al*. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol* 2004;**22**:177–83.

138. Rubin DL, Lewis SE, Mungall CJ, *et al*. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS* 2006;**10**:185–98.

139. Kin T, Yamada K, Terai G, *et al*. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res* 2007;**35**(Database issue):D145–8.

140. Robertson G, Bilenky M, Lin K, *et al*. cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* 2006;**34**(Database issue):D68–73.

141. SPARQL Query Language for RDF 2007. http://www.w3.org/TR/rdf-sparql-query/ (18 May 2007, date last accessed).

142. Stephens S. HCLSIG BioRDF Subgroup 2007. http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup (18 May 2007, date last accessed).

143. Kotecha N, Bruck K, Lu W, Shah N. *Pathway knowledge base: integrating BioPAX compliant data sources* 2007. http://esw.w3.org/topic/HCLS/ISWC/Workshop/Abstracts?action=AttachFile&do=get&target=Nigam_Shah_Presentation.pdf (18 May 2007, date last accessed).

144. Efron B, Hastie T, Johnstone I, *et al*. Least Angle Regression. *Ann Statist* 2004;**32**:407–99.

145. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**:186–94.

146. Ewing B, Hillier L, Wendl MC, *et al*. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998;**8**:175–85.

147. Vidal M. Time for a human interactome project? *The Scientist,* March 2006.

148. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

149. Graupner S, Wackernagel W. Identification and characterization of novel competence genes comA and exbB involved in natural genetic transformation of *Pseudomonas stutzeri*. *Res Microbiol* 2001;**152**:451–60.

150. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* 2004;**7**:535–45.

151. Eilbeck K, Lewis SE, Mungall CJ, *et al*. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* 2005;**6**:R44.

152. Resource Description Framework (RDF) 2007. www.w3.org/RDF/ (18 May 2007, date last accessed).

153. RDF Primer - Turtle Version 2007. http://www.dajobe.org/2004/01/turtle/ (18 May 2007, date last accessed).