Putting microarrays in a context: Integrated analysis of diverse biological data

Olga G. Troyanskaya

Abstract
In recent years, multiple types of high-throughput functional genomic data that facilitate rapid functional annotation of sequenced genomes have become available. Gene expression microarrays are the most commonly available source of such data. However, genomic data often sacrifice specificity for scale, yielding very large quantities of relatively lower-quality data than traditional experimental methods. Thus sophisticated analysis methods are necessary to make accurate functional interpretation of these large-scale data sets. This review presents an overview of recently developed methods that integrate the analysis of microarray data with sequence, interaction, localisation and literature data, and further outlines current challenges in the field. The focus of this review is on the use of such methods for gene function prediction, understanding of protein regulation and modelling of biological networks.

INTRODUCTION
The availability of complete genomic sequences of several eukaryotic organisms, including the human genome, has brought molecular biology into a new era of systematic functional understanding of cellular processes. The sequences themselves provide a wealth of information, but functional annotation is a necessary step toward comprehensive description of genetic systems of cellular controls. High-throughput functional technologies, such as genomic and soon proteomic microarrays, allow one to rapidly assess general functions and interactions of proteins in the cell. In addition to gene expression microarrays, other high-throughput experimental methods are generating increasing amounts of data. In yeast Saccharomyces cerevisiae, the most well-studied eukaryotic organism that is commonly used in computational and experimental genomic studies, these data sets include protein–protein interaction studies (affinity precipitation, two-hybrid techniques), synthetic rescue and lethality experiments, and microarray analysis. This increase in functional data is also reflected in the rise of multiple functional databases, especially for yeast, including the Biomolecular Interaction Network Database, the Database of Interacting Proteins, the Molecular Interactions Database, the General Repository for Interaction Datasets, the MIPS Comprehensive Yeast Genome Database and the model organism database for yeast – Saccharomyces Genome Database (SGD). While classical genetic and cell biology techniques continue to play an important role in the detailed understanding of cellular mechanisms, the combination of rapid generation and analysis of functional genomics data with targeted exploration by traditional methods will facilitate fast and accurate identification of causal genes and key pathways affected in cellular regulation, development, and in disease.

Thus, the key goal of these high-throughput data is rapid functional annotation of the sequenced genomes and understanding of gene regulation and biological networks. Even in yeast, the most well-studied eukaryote, 1,481
of 5,788 open reading frames (ORFs) are still unnamed, and functional annotation is unknown for 1,865 ORFs. High-throughput functional data, especially the large number of microarray data sets, are important for rapid functional annotation of these unknown genes, but it is important to recognise that high-throughput methods sacrifice specificity for scale in the quality to coverage trade-off, yielding too many false positives in the data sets. Recent work has highlighted this problem, showing that different cDNA microarrays exhibit between 10 and 30 per cent variation among corresponding microarray elements. For gene function annotation and biological network analysis, an increase in accuracy is essential, even if it comes at the cost of some sensitivity. This review presents an overview of computational methods that incorporate the abundant microarray data with other data sources for increased specificity in gene function prediction and in identification of biological networks. Recent progress in integrated analysis of heterogeneous data is outlined, presenting the methods in a rough order of increasing complexity of biological questions – from gene function prediction, through regulation, to biological networks. A general overview of the data integration tasks is presented in Figure 1.

**Diverse genomic data**

- Gene expression dataset 1
- Gene expression dataset 2
- Gene expression dataset 4
- Yeast two-hybrid dataset 1
- Yeast two-hybrid dataset 2
- Affinity precip dataset 1
- Affinity precip dataset 2
- Synthetic lethality dataset
- Synthetic rescue dataset
- Synthetic interaction dataset
- Transcription factor bio sites
- Location in the genome
- Literature data

**Gene-gene relationships based on integrated data**

**Experimental testing of predictions & feedback**

**Prediction of regulation and biological networks**

**Computational data integration**

**Figure 1:** Overview of integrated analysis of genomic data. Multiple gene expression data sets and diverse genomic data can be integrated by computational methods to create an integrated picture of functional relationships between genes. These integrated data can then be used to predict biological function or to aid in understanding of protein regulation and biological networks modelling. Alternatively, computational approaches for biological networks prediction can analyse diverse genomic data directly, without the intermediate integration step. Upon evaluation by cross-validation or based on a test set of labelled data, best novel predictions should be tested experimentally, and the results of these experiments can be used to improve performance of the methods.
GENE FUNCTION PREDICTION

Currently, gene expression microarray data sets are the most commonly available functional genomic data owing to their relatively low cost and easily accessible technology. At the time of publication, NCBI’s Gene Expression Omnibus database\textsuperscript{17} already contained over 650 gene expression data sets, 60 of which are yeast and 203 are human, and other databases provide additional gene expression data. These data can be used to identify groups of coexpressed genes, and such groups, through the principle of ‘guilt by association’, can facilitate function prediction for unknown proteins and identification of regulatory elements. However, while gene coexpression data are an excellent tool for hypothesis generation, microarray data alone often lack the degree of specificity needed to make accurate biological conclusions. For such purposes, an increase in accuracy is needed, even if it comes at the cost of some sensitivity. This improvement in specificity can be achieved through incorporation of other data sources in an integrated analysis of gene expression data. These additional data sources include other high-throughput functional data (eg protein–protein interactions, genetic interaction data, localisation information), DNA and protein sequence data, published literature and phylogenetic information.

Improving microarray analysis with other genomic data

Bioinformatics methods for effective integration of high-throughput heterogeneous data can provide the improvement in specificity necessary for accurate gene function annotation and network analysis based on high-throughput data.\textsuperscript{8,9,34,35} While the exact amount of overlap and correlation among functional data sets is unclear,\textsuperscript{32,36–38} data integration has been shown to increase the accuracy of gene function prediction compared to a single high-throughput method.\textsuperscript{31,34,39–43} Specifically, studies demonstrated that using more than one type of functional data for predictions increased accuracy\textsuperscript{31} and that integrating more heterogeneous information increases the number of protein–protein interactions correctly identified,\textsuperscript{42} leading to better prediction of function for unknown proteins. This potential of data integration recently led to development of several computational methods for integrated analysis of microarray data with other data sources.

A simple scheme for increasing accuracy in function prediction based on heterogeneous data is to consider the intersection of interaction maps for different high-throughput data sets.\textsuperscript{44} While this scheme reduces the false positives, it has the drawback that the lowest-sensitivity data set will limit sensitivity of the entire analysis. As published large-scale interaction studies are not comprehensive even in model organisms, this strict sensitivity limitation is too restrictive for large-scale and general function prediction. Several other groups suggested approaches that provide increased sensitivity of function prediction from the intersection scheme above. In the first study of this type, Marcotte et al. predicted a number of potential protein functions for \textit{S. cerevisiae} based on a heuristic combination of different types of data.\textsuperscript{34,39} In another early study, Schwikowski et al. assigned putative protein function based on the number of interactions an unknown protein has with proteins from different functional categories.\textsuperscript{40} These studies demonstrated the potential of integrated data analysis, but they combine the information from different sources in a heuristic fashion, where confidence levels for protein–protein links are defined on a case-by-case basis. This approach is successful in these studies and served as a clear proof of concept, but it may be hard to generalise to new data sets, data types or other organisms because each approach is developed with specific data and application goal in mind and therefore lacks a general scheme or representation.
A more general method was developed by Clare and King, who introduced a rule-based method in which heuristics are learned based on heterogeneous data sources and known functional predictions. These heuristics are then applied to genes with unknown function to predict function. This study uses a modified C4.5 decision tree algorithm, and includes sequence, phenotype, expression and predicted secondary structure data. In a different approach, Karaoz et al. combined interactions and expression data by creating a weighted graph of protein–protein interactions with the weight between two genes derived from coexpression values of these genes in one gene expression data set. They then used a variant of discrete-state Hopfield network to assign function for unknown proteins, based on known annotations in the Gene Ontology.

Probabilistic integration of heterogeneous data
Recently, several computational methods have been suggested that combine data sets in a confidence-dependent manner. The advantage of such statistical approaches is that they enable general data integration and can easily adapt to new data sources. In addition, because these methods are probabilistic, their outputs can be filtered by the confidence or probability cut-off to a desired level of sensitivity and specificity (estimated based on the cross-validation trials or a test data set).

In a general methodology based on support vector machines, Lanckriet et al. have combined interactions, expression and sequence data by representing each input as a separate kernel. The weighted optimised combination of these kernels was then used to recognise membrane and ribosomal proteins as well as other general classes of proteins. This method is general and can also readily provide information, encoded in the kernel weights, on the extent to which each data source contributes to the final prediction. One disadvantage of such discriminative approaches is that a separate classifier is generally built for each functional category, thereby making it possible to only predict general functional categories (e.g. metabolism) because of lack of training data for more specific functions. Methodologies that first perform general data integration, creating a general graph of functional relationships, and then predict function based on such a graph, can alleviate this problem. For example, Troyanskaya et al. used a Bayesian network-based method for general integrated analysis of functional genomic data. They then predicted function for each unknown gene based on significant over-representation of known proteins of particular function in the unknown gene’s neighbourhood in the graph. In an alternative approach, Zhang et al. predicted co-complexed protein pairs with probabilistic decision trees based on expression and proteomics data.

Including prior knowledge through biological literature
In addition to high-throughput experimental methods, traditional experimental techniques have generated volumes of biological knowledge in the past decades. Results of such experiments are often substantially more accurate than large-scale functional genomic data, and many of their conclusions have been verified by multiple techniques. This knowledge is encoded in the wealth of biological literature, which, if properly analysed, may provide the strongest aid yet for the analysis of high-throughput data. For example, Raychaudhuri et al. use biomedical abstracts to resolve boundaries of hierarchical clusters of gene expression patterns and to recognise clusters that are most functionally coherent. Unfortunately, current work in this area focuses on analysis of keywords or article abstracts, largely because full-text literature mining is restricted by the lack of availability of full-text articles copyrighted to biomedical journals.
In addition to original literature, increasing sources of human-curated databases of structured biological knowledge are available. Probably of most influential is the Gene Ontology – an acyclic directed graph of biological terms divided into three parts: biological process, cellular location and molecular function. Gene Ontology terms are being used to annotate genes in different organisms, and these annotations often serve as the ‘gold standard’ or training data for microarray analysis and gene function prediction methods. In addition to gene function, multiple databases aim to encode knowledge about metabolic and regulatory pathways in different organisms, for example the MetaCyc and KEGG pathway databases. These are also very valuable resources for training and evaluation of computational analysis methods. Hanisch et al., for example, used biological networks as an integrated part of their clustering algorithm – with a single distance metric derived from both metabolic networks (from the KEGG database) and gene expression data.

USING MICROARRAYS TO DECIPHER GENE REGULATION

Gene expression data provide insight not only into gene function, but also into regulatory processes in the cell. In fact, very early in microarray analysis several groups designed methods for identification of potential transcription factor binding sites in the upstream sequences of coexpressed genes, for example. The general approach is to cluster gene expression patterns and then identify motifs or motif combinations common to each cluster. Bussemaker et al. developed a method that does not require clustering and can identify statistically significant motifs based on a single genome-wide set of expression values. However, motif discovery methods cannot on their own identify which transcription factor binds each particular motif, and therefore stop short from identifying regulatory modules (sets of coexpressed genes regulated by sets of transcription factors). The recently developed chromatin immunoprecipitation microarray (ChIP) technology can connect specific transcription factors to a large number of binding sites. This technique can identify direct binding of a specific protein complex to DNA on whole-genome scale and thus is complementary to gene expression microarrays. Integrated analysis of ChIP and gene expression microarrays can identify coregulated groups of genes, their regulators and the corresponding transcription factor binding sites with higher accuracy than analysis of either data type alone. An iterative approach suggested by Bar-Joseph et al., for example, improves clustering of gene expression microarray data by using ChIP microarray data to identify combinations of regulators. Another method developed by Kato et al. identifies over-represented motif combinations found upstream from strongly coexpressed genes, and associates these motifs with transcription factors. Segal et al. used a Bayesian framework for identifying modules based on known regulatory proteins and gene expression data. All of these methods, by identifying groups of coexpressed and coregulated genes and determining their regulators, identify small components of regulatory circuits of the cell.

INTEGRATED ANALYSIS OF BIOLOGICAL NETWORKS

Possibly some of the most interesting questions of present-day computational functional genomics arise in the area of biological networks prediction, where the goal is to decipher all patterns of regulation in the cell. Creating network models involves, explicitly or implicitly, solving every one of the above-described problems: gene function prediction, understanding of protein–protein interactions and identification of regulatory relationships. Although
multiple studies have attempted to estimate gene networks from microarray data alone, gene expression is usually not sufficient for accurate network modelling because of its limited scope (only transcriptional regulation is represented in gene expression microarray data sets, and they cover a limited number of conditions) and its high noise levels. Integrated analysis of multiple types of high-throughput data is essential for effective prediction of accurate biological networks.

Increasing number of studies on modelling biological networks based on integrated data are being published. Hartemink et al. reduced noise in regulatory network models by using localisation data to influence the prior of their Bayesian network model, in which gene expression influenced the model likelihood. However, such a model would still miss non-transcriptional regulation that is often due to physical interactions between proteins. To address this issue, several groups used protein–protein interaction data in addition to gene expression data sets in constructing probabilistic network models. Tanay et al. also included growth phenotype and transcription factor binding data, in addition to gene expression and protein–protein interactions. They used a biclustering technique to identify statistically significant modules based on the diverse data sets, then constructed biological networks based on transcription factor binding profiles and their correspondence to modules.

OPEN PROBLEMS IN DATA INTEGRATION

This review has outlined how integrated analysis of microarray data with other genomic data sources can increase prediction accuracy and provide a coherent view of functional information derived from diverse data types. Integrated methods can be based on formal probabilistic reasoning and can generate predictions based on heterogeneous data sources, and some are generalisable to new data sources as they become available. Although several promising probabilistic methods for integrated analysis have been developed, the problem of general data integration for both gene function prediction and pathway modelling is still not fully solved. No truly general and robust method that can be routinely applied to noisy, heterogeneous data has yet been developed. Additionally, the majority of methods have been demonstrated only in baker’s yeast, as multicellular organisms present a host of additional challenges for data integration.

One very promising direction in functional analysis of microarray data is integration of data from multiple organisms. Recently, several groups have started using coexpression information from homologous genes in several species to increase specificity of functional relationships identified from gene expression experiments (eg Stuart et al., McCarroll et al.). Such comparative genomics techniques, on their own or combined data integration methods described in this review, will undoubtedly contribute to functional annotation and modelling of biological networks.

It is also important to note that computational methods are always limited by the coverage and quality of experimental data they use. Public availability of high-quality, high-throughput data sets is therefore essential for rapid functional annotation. Further experimental validation of computational predictions by traditional laboratory techniques is ideal for validation and for improvement of the computational methodology. Such validation can be accomplished through collaborations with biological researchers and through open publication of predictions in the form easily accessible to biologists.

Development of accurate data integration methods for functional genomics relies on labelled data for training and validation, for example genes with known functions or known biological pathways. Such data, generated
Public availability of genomic data and of curated biological information are critical for development of effective data integration and modelling methods.

by traditional biological methods, are often scarce and for the most part represented in biological literature in the free-text format that cannot be readily used for automatic training or validation. One very effective solution to this problem is human curation, employed by several databases (eg *Saccharomyces* database27). However, curation is costly and thus currently limited. Therefore, accurate computational analysis of biomedical literature to extract biological relationships that can be used as ‘gold standard’ data is an area of great importance that presents many natural language processing challenges.

**CONCLUSION**

Key challenges in present-day molecular biology are the functional annotation of unknown genes within sequenced genomes and determining protein interactions and regulation in biological networks. Traditional experimental methods are too slow and labour-intensive to accomplish these tasks on the genomic scale in the near future. Therefore we must rely on high-throughput techniques along with computational analysis to direct more traditional experimentation. In the past, computational techniques in functional genomics have focused primarily on gene expression microarray data. But integrated analysis techniques for diverse biological data have emerged as more large-scale functional data have become available. Future development of more accurate integrative methodologies and their expansion to multicellular organisms complemented by further development of high-throughput experimental technologies will be critical for complete functional annotation of model organisms and human genomes.

**References**

* Papers of particular interest published within the period of this review.

** Papers of extreme interest published within the period of this review.


This paper describes the GRID repository, which currently includes databases of functional genomic data for yeast, fly and worm. The databases provide search and download interfaces for physical and genomic interaction data sets, as well as other data collected from multiple publications. This is an excellent source of data for data integration in model organisms.


Model organism databases serve as the central portal for information about the model organism. SGD, the model organism database for yeast, provides summary of knowledge about each gene in yeast, as well as links to most yeast studies and high-throughput data sources. The database staff curates yeast literature and annotates genomic features with data from appropriate publications. This is an invaluable resource for computational researchers working on high-throughput data analysis for yeast.


The investigators developed a Bayesian system for integration of diverse functional genomic data, including gene expression microarrays, interactions, sequences and localisation data. The priors in the Bayesian network were formally assessed from experts in yeast biology. This is the only system that incorporates multiple analyses of microarray data as well as multiple data types. The system was applied to gene function prediction and performed substantially better than any of the input methods. This study provided novel function predictions for unknown yeast proteins, and also demonstrated how data integration can help improve curated annotations for known genes.


37. Kemmeren, P., van Berkum, N. L., Vila, J.


The authors present an SVM-based framework for integration of diverse data sources and apply this framework to identification of membrane and ribosomal proteins. In this method, similarity relationships between pairs of proteins within each data set are represented with a separate kernel function. These kernel functions are then combined optimally by use of semidefinite programming to reduce the task to a convex optimization problem. The relative weights of each kernel in the optimal linear combination provide a way to assess how informative each data type is for each prediction task.


This study presents a computational method that uses abstracts from biomedical literature to improve integrated clustering analysis of gene expression data. The authors evaluate functional coherence of a cluster by an information-theoretic score measure calculated from word-based similarity of article abstracts associated with each gene in the cluster. They set hierarchal cluster boundaries by maximizing this score, so that the functional coherence of the clusters is maximized. The method is applied to yeast and fly data sets, and in both cases it automatically identified biologically relevant clusters.


This study suggests a clustering approach for microarray data that integrates prior knowledge of biological networks into the clustering itself. The authors derive a distance metric that is influenced by both the extent of coexpression of pairs of genes and by the proximity of these genes in the network. This method allows the network structure to directly influence the clustering process. Unfortunately, this, and any other such network-based approach, is inherently largely limited to metabolic networks as a limited number of regulatory pathways are well known.

This study presents a probabilistic method that identifies regulatory modules based on gene expression data and a precompiled set of potential regulatory genes. The methodology, based on probabilistic graphical models, is an iterative procedure that petitions genes into modules and defines the conditions under which each type of regulation occurs for each module. The authors applied their method to a yeast stress response data set supplemented by a set of 466 yeast transcription factors and potential signaling proteins. They identified multiple functionally coherent regulatory modules, and confirmed several of them experimentally. This study demonstrates the power of combining cutting-edge computational approaches with biological insight and experimental testing: an innovative computational method is applied to an interesting biological problem, followed by informed interpretation of novel biological results and their experimental testing.
This group demonstrated modular organisation of the yeast molecular network by identifying modules (sets of similarly behaving genes) based on heterogeneous genomic data. The investigators represented expression, interactions, phenotype and regulation data as a bipartite graph, then used a biclustering algorithm to identify statistically significant modules. In addition to exploring the architecture of the yeast regulatory network, the authors also provides multiple new functional annotations for unknown yeast genes.