

# **Integration and analysis of diverse genomic data**

**Olga G. Troyanskaya**

Department of Computer Science &  
Lewis-Sigler Institute for Integrative Genomics  
Princeton University, USA

## **Tutorial Summary**

In the recent years, multiple types of high-throughput functional genomic data have become available that facilitate rapid functional annotation and pathway modeling in the sequenced genomes. Gene expression microarrays are the most commonly available source of such data. However, genomic data sacrifice specificity for scale compared to traditional experimental methods, yielding large quantities of relatively lower quality measurements. This problem has generated much interest in bioinformatics in the past two years, as sophisticated computational methods are necessary for accurate functional interpretation of these large-scale datasets. This tutorial will present an overview of recently developed methods for integrated analysis of functional genomic data and outline current challenges in the field. The focus will be on the development and use of such methods for gene function prediction, understanding of protein regulation, and modeling of biological networks.

## **Tutorial Outline**

- Goals of data integration
- Overview of available experimental data
- Evaluation of data/method accuracy
- Overview of computational methodology
- Data representation for integration
- Application of data integration

This tutorial will be of interest to computational researchers interested in contributing to the field of data integration and analysis of heterogeneous data and to biologists with some computational background who are interested in using the methods on their experimental data and understanding their properties and limitations.

## **Tutorial level**

Introductory to intermediate. This tutorial will serve as a thorough introduction to data integration in functional genomics, but some advanced issues will also be introduced.

## **Prior knowledge required**

This tutorial will be self-contained and assume no prior background in the field of data integration or biological data analysis. No specific computational or biological background will be assumed, and the audience may include computer scientists, statisticians, bioinformaticians, and computationally savvy biologists. The audience should be familiar with basic biological concepts (e.g. regulation, transcription, etc) and basic computation (probability).

All concepts will be introduced on an intuitive level, so a biologist or a computer scientist will be comfortable with the material. Building on this introductory material, state-of-the-art methods for data integration will be introduced with special emphasis on assumptions, limitations, and strengths of each method. Finally, open problems in the field will be discussed.

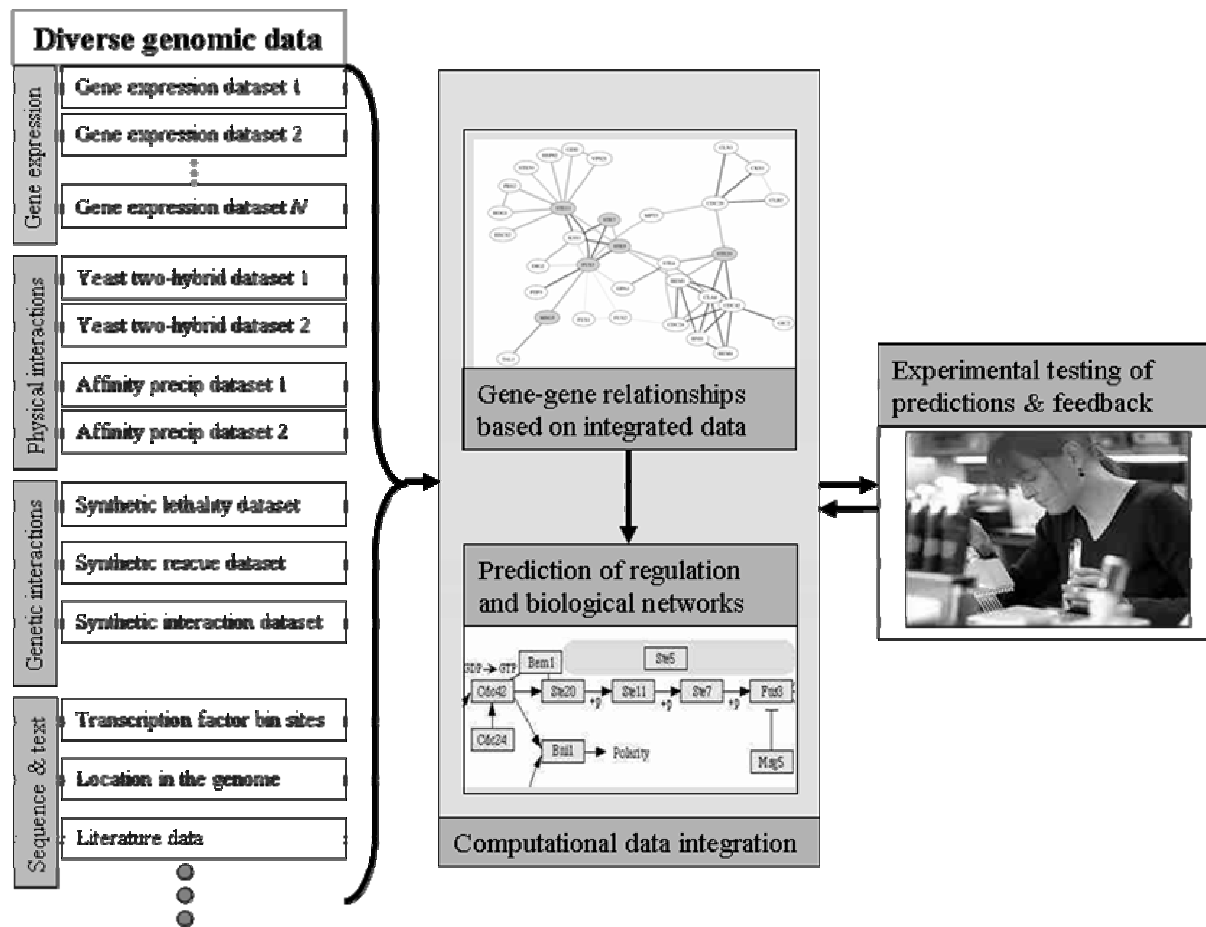
## Introduction

The availability of complete genomic sequences of several eukaryotic organisms, including the human genome (1-6), 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 (7-9). High-throughput functional technologies, such as genomic (10, 11) and soon proteomic microarrays (12-16), allow one to rapidly assess general functions and interactions of proteins in the cell. In addition to gene expression microarrays (17), 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 datasets include protein-protein interaction studies (affinity precipitation (18), two-hybrid techniques (19)), synthetic rescue (20) and lethality (20, 21) experiments, and microarray analysis (10, 11). This increase in functional data is also reflected in the rise of multiple functional databases, especially for yeast, including: the Biomolecular Interaction Network Database (22), the Database of Interacting Proteins (23), the Molecular Interactions Database (24), the General Repository for Interaction Datasets (25<sup>•</sup>), the MIPS Comprehensive Yeast Genome Database (26), and the model organism database for yeast—*Saccharomyces* Genome Database (SGD) (27<sup>•</sup>). 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, 1481 of 5788 open reading frames (ORFs) are still unnamed, and functional annotation is unknown for 1865 ORFs. High-throughput functional data, especially the large number of microarray datasets, are important for rapid functional annotation of these unknown genes, but it is important to recognize that high-throughput methods sacrifice specificity for scale in the quality to coverage tradeoff, yielding to many false positives in the datasets (8, 28-32). Recent work has highlighted this problem, showing that different cDNA microarrays exhibit between 10 and 30 percent variation among corresponding microarray elements (33). For gene function annotation and biological network analysis, an increase in accuracy is essential, even if it comes at the cost of some sensitivity (30). 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. We outline recent progress in integrated analysis of heterogeneous data, presenting the methods in a rough order of increasing complexity of biological questions – from gene function prediction, to regulation, to biological networks. A general overview of the data integration tasks is presented in Figure 1.

**Figure 1. Overview of integrated analysis of genomic data.** Multiple gene expression datasets 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 modeling. Alternatively, computational approaches for biological networks prediction can analyze diverse genomic data directly, without the intermediate integration step. Upon evaluation by cross-validation or based on a test set of labeled 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 datasets are the most commonly available functional genomic data due to their relatively low cost and easily accessible technology. At the time of publication, NCBI's Gene Expression Omnibus database (17) already contained over 650 gene expression datasets, sixty of which are yeast and 203 are human datasets, 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 (e.g. protein-protein interactions, genetic interaction data, localization 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 (8, 9, 34, 35). While the exact amount of

overlap and correlation among functional datasets is unclear (32, 36-38), data integration has been shown to increase the accuracy of gene function prediction compared to a single high-throughput method (31, 34, 39-43). Specifically, studies demonstrated that using more than one type of functional data for predictions increased accuracy (31) and that integrating more heterogeneous information increases the number of protein-protein interactions correctly identified (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 datasets (44). While this scheme reduces the false positives, it has the drawback that the lowest-sensitivity dataset 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 *S. cerevisiae* based on a heuristic combination of different types of data (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 (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 generalize to new datasets, 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 *et al.*, who introduced a rule-based method in which heuristics are learned based on heterogeneous data sources and known functional predictions (45). 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 dataset (46). They then used a variant of discrete-state Hopfield network to assign function for unknown proteins, based on known annotations in the Gene Ontology (47).

### **Probabilistic integration of heterogeneous data**

Recently, several computational methods have been suggested that combine datasets 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 cutoff 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.* has combined interactions, expression, and sequence data by representing each input as a separate kernel. The weighted optimized combination of these kernels was then used to recognize membrane

and ribosomal proteins (48<sup>•</sup>) as well as other general classes of proteins (49). 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 graph, can alleviate this problem (Figure 1). For example, Troyanskaya *et al.* used a Bayesian network-based method for general integrated analysis of functional genomic data (35<sup>•</sup>). They then predicted function for each unknown gene based on significant over-representation of known proteins of particular function in the unknown gene's neighborhood 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 (50).

### **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 analyzed, 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 recognize clusters that are most functionally coherent (51<sup>•</sup>). 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 (47). 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 (52). 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 (53, 54). 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 (55<sup>•</sup>).

### **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 (56-59). 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 (58).

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 (60). Another method developed by Kato *et al.* identifies over-represented motif combinations found upstream from strongly co-expressed genes, and associates these motifs with transcription factors (61). Segal *et al.* used a Bayesian framework for identifying modules based on known regulatory proteins and gene expression data (62<sup>••</sup>). 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 modeling because of its limited scope (only transcriptional regulation is represented in gene expression microarray datasets, 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 modeling biological networks based on integrated data are being published. Hartemink *et al.* reduced noise in regulatory network models by using localization data to influence the prior of their Bayesian network model, in which gene expression influenced the model likelihood (63). However, such 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 interactions data in addition to gene expression datasets in constructing probabilistic network models (64). Tanay *et al.* also included growth phenotype and transcription factor binding data, in addition to gene expression and protein-protein interactions (65<sup>••</sup>). 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 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 generalizable 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 modeling 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 multi-cellular 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 co-expression information from homologous genes in several species to increase specificity of functional relationships identified from gene expression experiments (e.g. 66, 67). Such comparative genomics techniques, on their own or combined data integration methods described in this review, will undoubtedly contribute to functional annotation and modeling 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 datasets 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 labeled data for training and validation, for example genes with known functions or known biological pathways. Such data, generated by traditional biological methods, is 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 (e.g. 27). 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 labor-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 multi-cellular organisms complemented by further development of high-throughput experimental technologies will be critical for complete functional annotation of model organisms and human genomes.

## **Annotated references**



- – papers of particular interest published within the period of this review
- – papers of extreme interest published within the period of this review

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 Genes. *Science* 1996;274(5287):546-567.
2. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* 2002;415(6874):871-80.
3. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of *Drosophila melanogaster*. *Science* 2000;287(5461):2185-95.
4. Consortium TCeS. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998;282(5396):2012-8.
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860-921.
6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291(5507):1304-51.
7. Kitano H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr Genet* 2002;41(1):1-10.
8. Steinmetz LM, Deutschbauer AM. Gene function on a genomic scale. *J Chromatogr B Analyt Technol Biomed Life Sci* 2002;782(1-2):151-63.
9. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001;2:343-72.
10. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet* 1999;21(1 Suppl):20-4.
11. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270(5235):467-70.
12. Cahill DJ, Nordhoff E. Protein arrays and their role in proteomics. *Adv Biochem Eng Biotechnol* 2003;83:177-87.
13. Sydor JR, Nock S. Protein expression profiling arrays: tools for the multiplexed high-throughput analysis of proteins. *Proteome Sci* 2003;1(1):3.
14. Oleinikov AV, Gray MD, Zhao J, Montgomery DD, Ghindilis AL, Dill K. Self-assembling protein arrays using electronic semiconductor microchips and in vitro translation. *J Proteome Res* 2003;2(3):313-9.
15. Huang RP. Protein arrays, an excellent tool in biomedical research. *Front Biosci* 2003;8:d559-76.
16. Cutler P. Protein arrays: the current state-of-the-art. *Proteomics* 2003;3(1):3-18.
17. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30(1):207-10.
18. Larsson PO, Mosbach K. Affinity precipitation of enzymes. *FEBS Lett* 1979;98(2):333-8.
19. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature* 1989;340(6230):245-6.
20. Novick P, Osmond BC, Botstein D. Suppressors of yeast actin mutations. *Genetics* 1989;121(4):659-74.
21. Bender A, Pringle JR. Use of a screen for synthetic lethal and multicopy suppressor mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae*. *Mol Cell Biol* 1991;11(3):1295-305.
22. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003;31(1):248-50.

23. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002;30(1):303-5.
24. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. *FEBS Lett* 2002;513(1):135-40.
- 25\*. Breitkreutz BJ, Stark C, Tyers M. The GRID: The General Repository for Interaction Datasets. *Genome Biol* 2003;4(3).

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

26. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002;30(1):31-4.
- 27\*. Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, et al. Saccharomyces Genome Database. *Methods Enzymol* 2002;350:329-46.

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.

28. Grunenfelder B, Winzeler EA. Treasures and traps in genome-wide data sets: case examples from yeast. *Nat Rev Genet* 2002;3(9):653-61.
29. Chen Y, Xu D. Computational analyses of high-throughput protein-protein interaction data. *Curr Protein Pept Sci* 2003;4(3):159-81.
30. Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T, Boone C. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends Cell Biol* 2003;13(7):344-56.
31. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417(6887):399-403.
32. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 2002;1(5):349-56.
33. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, et al. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res* 2001;29(8):E41-1.
34. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999;285(5428):751-3.

- 35\*. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc Natl Acad Sci U S A* 2003;100(14):8348-53.
- The investigators developed a Bayesian system for integration of diverse functional genomic data, including gene expression microarrays, interactions, sequences, and localization 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.

36. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* 2002;18(10):529-36.
  37. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* 2002;9(5):1133-43.
  38. Werner-Washburne M, Wylie B, Boyack K, Fuge E, Galbraith J, Weber J, et al. Comparative analysis of multiple genome-scale data sets. *Genome Res* 2002;12(10):1564-73.
  39. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999;402(6757):83-6.
  40. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nat Biotechnol* 2000;18(12):1257-61.
  41. Bader GD, Hogue CW. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* 2002;20(10):991-7.
  42. Gerstein M, Lan N, Jansen R. Proteomics. Integrating interactomes. *Science* 2002;295(5553):284-7.
  43. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001;29(4):482-6.
  44. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 2002;295(5553):321-4.
  45. Clare A, King RD. Predicting gene function in *Saccharomyces cerevisiae*. *Bioinformatics* 2003;19 Suppl 2:II42-II49.
  46. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A* 2004;101(9):2888-93.
  47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25(1):25-9.
  - 48\*. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics* 2004.
- 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 dataset 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.
49. Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS. Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput* 2004:300-11.
  50. Zhang LV, Wong SL, King OD, Roth FP. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 2004;5(1):38.
  - 51\*. Raychaudhuri S, Chang JT, Imam F, Altman RB. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res* 2003;31(15):4553-60.

This study presents a computational method that uses abstracts from biomedical literature to improve hierarchical 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 hierarchical cluster boundaries by maximizing this score, so that the functional coherence of the clusters is maximized. The method is applied to yeast and fly datasets, and in both cases it automatically identified biologically relevant clusters.

52. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, et al. *Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)*. *Nucleic Acids Res* 2002;30(1):69-72.

53. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, et al. *MetaCyc: a multiorganism database of metabolic pathways and enzymes*. *Nucleic Acids Res* 2004;32 Database issue:D438-42.

54. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. *The KEGG resource for deciphering the genome*. *Nucleic Acids Res* 2004;32 Database issue:D277-80.

55\*. Hanisch D, Zien A, Zimmer R, Lengauer T. *Co-clustering of biological networks and gene expression data*. *Bioinformatics* 2002;18 Suppl 1:S145-54.

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 co-expression 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 networks-based approach, is inherently largely limited to metabolic networks as a limited number of regulatory pathways is well-known.

56. Roth FP, Hughes JD, Estep PW, Church GM. *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation*. *Nat Biotechnol* 1998;16(10):939-45.

57. Zhu Z, Pilpel Y, Church GM. *Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm*. *J Mol Biol* 2002;318(1):71-81.

58. Bussemaker HJ, Li H, Siggia ED. *Regulatory element detection using correlation with expression*. *Nat Genet* 2001;27(2):167-71.

59. Liu X, Brutlag DL, Liu JS. *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes*. *Pac Symp Biocomput* 2001:127-38.

60. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, et al. *Computational discovery of gene modules and regulatory networks*. *Nat Biotechnol* 2003;21(11):1337-42.

61. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ. *Identifying combinatorial regulation of transcription factors and binding motifs*. *Genome Biol* 2004;5(8):R56.

62\*. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data*. *Nat Genet* 2003;34(2):166-76.


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 partitions 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.

63. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput* 2002:437-49.
64. Nariai N, Kim S, Imoto S, Miyano S. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac Symp Biocomput* 2004:336-47.
- 65<sup>••</sup>. Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci U S A* 2004;101(9):2981-6.
- This group demonstrated modular organization 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.
66. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003; 302(5643):249-55.
67. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet.* 2004; 36(2):197-204.



14th Annual International Conference On Intelligent Systems For Molecular Biology



***Integration and Analysis of Diverse Genomic Data***

Olga Troyanskaya, PhD

Assistant Professor  
Lewis-Sigler Institute for Integrative Genomics  
& Department of Computer Science  
Princeton University, USA

I/MB 2006, Fortaleza, Brazil, August 6-10, 2006

## Outline of this tutorial

- Goals of data integration
- Overview of available experimental data
- Evaluation of data/method accuracy
- Overview of computational methodology
- Data representation for integration
- Application of data integration



58

## Goals & challenges of data integration

- Explosion of genomic data, but no equivalent explosion of biological information
- Why?
  - Data are noisy
  - Datasets are incomplete
  - Data are heterogeneous
- Effective data integration can lead to better biological predictions and faster growth of biological information



59

## Experimental data

- Coexpression
- Genetic association
- Physical association
- Protein arrays
- Localization
- Sequence
- Structure
- Literature

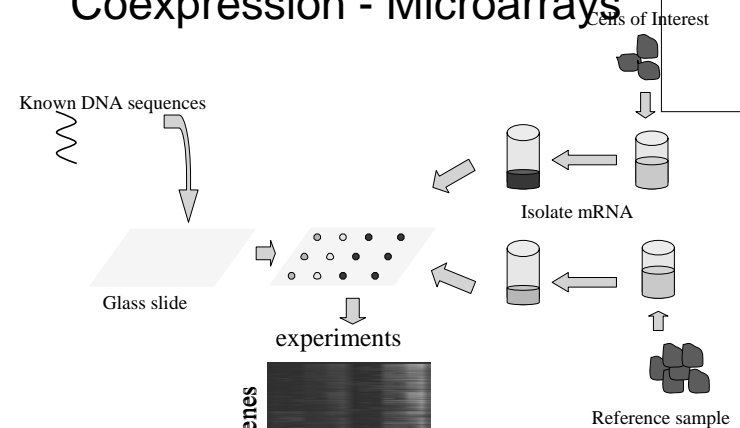


60

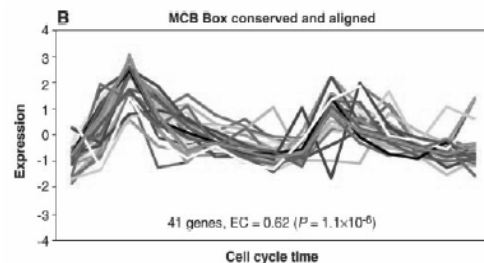
## Coexpression

- Coexpressed genes (microarrays)
- Chromatin IP on microarrays (ChIP on chip)

## Coexpression - Microarrays



## Co-regulated genes are co-expressed

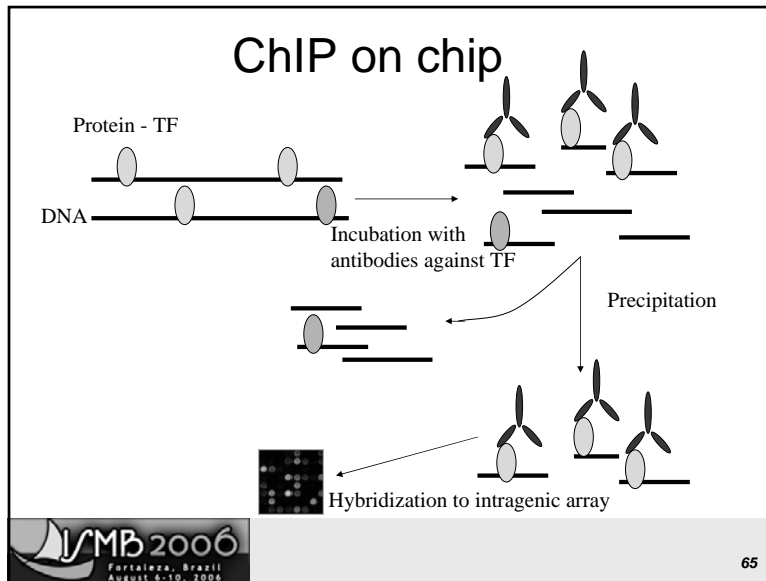


Expression profiles of 53 genes in *S. cerevisiae* genome that contain the exact match to an MCB box in their promoters (profiles normalized by mean & variance).

## Identifying TF factor binding sites directly – “ChIP on Chip”

- Array-based method for identification of binding sites of known TFs
- Each array corresponds to one TF





## Genetic interactions

- Synthetic lethality
- Synthetic interaction

**IUMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

66

## Synthetic lethality

- When gene A is deleted and B is still present, the cell is viable
- When gene B is deleted and A is still present, the cell is viable
- When both genes A and B are deleted, is the cell viable or not? If the cell is viable, then the genes are not functionally linked. If the cell is inviable, then the genes ARE functionally linked.

A-B+

Alive

A+B-

Alive

A-B-

Dead!

Synthetic lethality

**IUMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

67

## Synthetic interaction

- When gene A is deleted and B is still present, the cell is wild-type
- When gene B is deleted and A is still present, the cell is wild-type
- When both genes A and B are deleted, does this induce a non-wild-type phenotype? If yes (e.g. slow growth), then genes A and B have synthetic relationship.
- Note: if A-B- grows like wild-type, there still may be a different phenotype under which A and B have synthetic relationship =>
  - Negative results here don't mean much in general, but mean something specific to phenotypes!

A-B+

Wild-type growth

A+B-

Wild-type growth

A-B-

Slow growth

Synthetic interaction

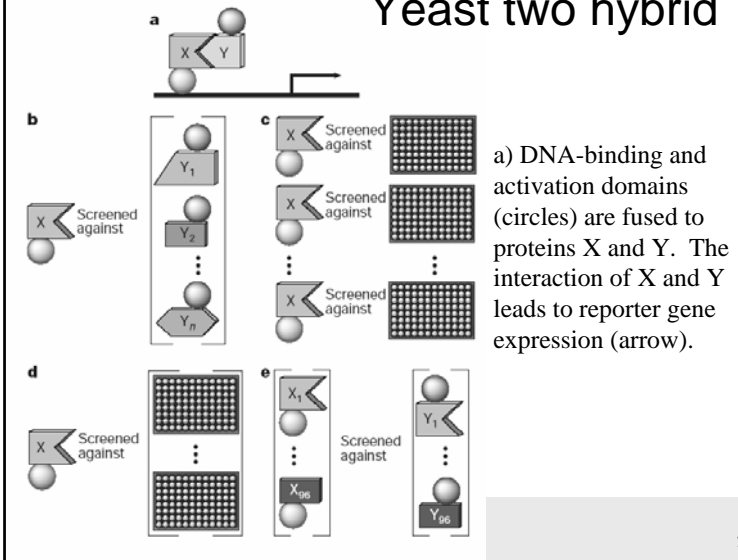
**IUMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

68

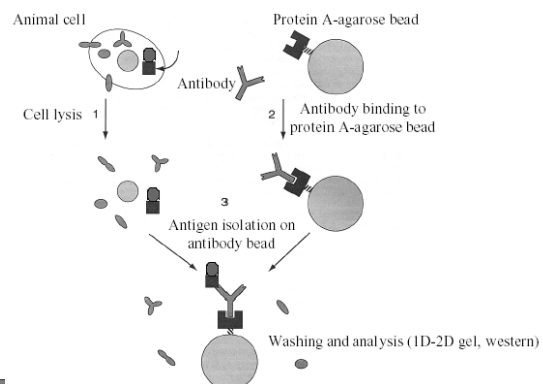
## Physical interactions

- Yeast two hybrid
- Co-IP precipitation
- FRET
- Protein arrays (can also test **molecular** function directly)

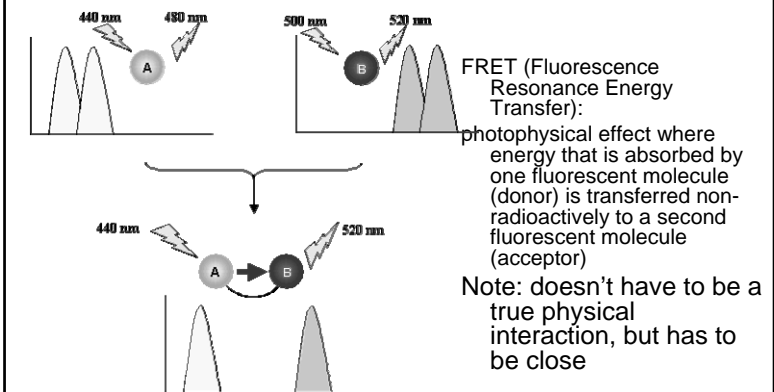
## Yeast two hybrid



## Co-IP



## FRET



# Protein arrays

**a**

Antibody  
Antigen  
Aptamer  
Allergen

Serum probes  
Cell lysates  
Living cells

Protein expression level  
Protein profiling  
Diagnostics

**b**

Protein  
Protein  
Peptide  
Peptide

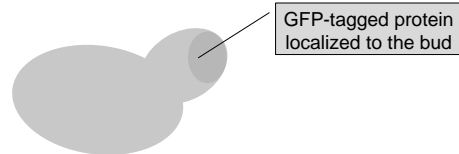
Protein probes  
Nucleic acid probes  
Drug probes  
Enzymes

Protein binding properties  
Pathway building  
Drug discovery  
Post-translational modification

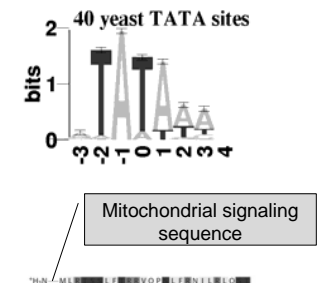
73

- Protein localization
- Sequence-based data
- Structure-based data
- Biomedical literature
- Public databases

- Proteins are tagged and their localization studied
- Protein that are co-localized may be more likely to have functional relationships
- Not all co-localizations are created equal:
  - Co-localization to the cytoplasm means very little
  - Co-localization to the nucleolus means more
- Localization may change depending on experimental conditions



- TF binding site predictions
- Homologues data
- Motifs (e.g. mito signal peptide targeting protein to mitochondria)



## Structure data

- Structural motifs
- Predicted functional binding sites (based on structure)
- Structural similarity to known proteins with specific function



77

## Biomedical Literature

- NLP-based prediction of relationships
  - Name co-occurrence in abstracts
  - Detecting specific types of relationships (e.g. geneA activates geneB)
- Curated literature
  - Ontologies (Gene Ontology, KEGG, MIPS)
  - Independent curation efforts by interaction databases



78

## Public databases

- Interactions data is often available through public databases
- Some databases are dataset-specific (e.g. O'Shea's lab co-localization DB)
- Some are general collections of data
  - Of some types: GRID, KEGG...
  - For one organism: SGD, FlyBase...



79

## Interaction coverage - yeast

- Interaction coverage is uneven
- Different biological processes can be better represented by different data types
- Some high-throughput studies actually focus on specific processes



80

## Evaluation of accuracy

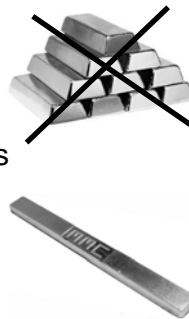
- Gene Ontology
- MIPS – Munich Information center for Protein Sequences
- KEGG - Kyoto Encyclopedia of Genes and Genomes

## Evaluation – the basics

- Any experimental or computational method needs to be evaluated
- Evaluation requires a reasonable number of answers
- Evaluation method depends on what question was asked
- Most current data integration efforts focus on one of the following questions:
  - prediction of interactions between proteins
  - prediction of gene function
  - prediction of pathways

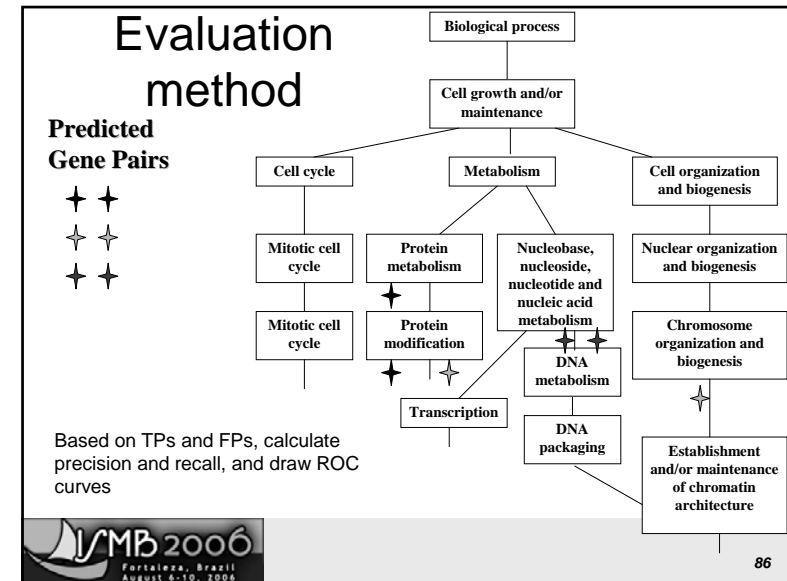
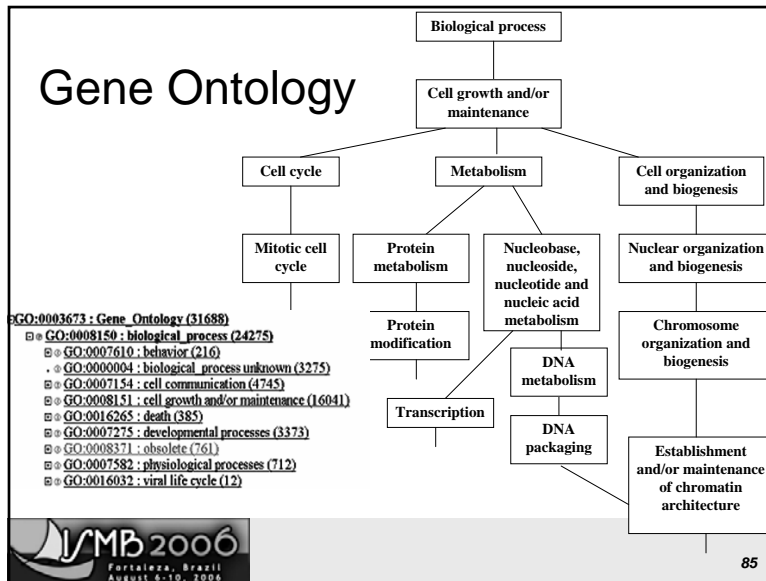
## “Gold” Standards

- Expert-curated assignments of genes to functional groups, complexes, or pathways
- Gene Ontology
- KEGG - Kyoto Encyclopedia of Genes and Genomes
- MIPS – Munich Information center for Protein Sequences
- Far from “Gold”, more like Pewter...



## Gene Ontology

- A loosely hierarchical (a DAG) organization of biological concepts
- Actually, three ontologies:
  - Biological process
  - Molecular function
  - Cellular component
- Pros:
  - Relative well annotated for many organisms
  - Provides varying levels of specificity
  - DAG structure gives a sense of relationships between nodes
- Cons:
  - Annotation coverage varies by process and organism
  - DAG structure makes it challenging to decide which nodes to use (biological process node is too general, for example)



## MIPS

- Loosely hierarchical (hierarchy not as deep as GO)
- In between KEGG and GO in terms of both specificity and coverage
- Pros:
  - Hierarchical
  - Hierarchy less deep, makes somewhat easier to choose appropriate nodes for evaluation
- Cons:
  - Annotation can be not as complete as GO (e.g. for yeast)

## KEGG

- Pathway-based...sort of
- Very specific coverage of metabolism, some regulatory pathways, and some other functional groups
- Pros: specificity
- Cons:
  - specificity (proteins that most biologists would consider related can belong to different pathways in KEGG)
  - Low coverage

## Important evaluation “footnotes”

- None of the “gold” standards is guaranteed to be fully correct, thus some TPs may not be right
- None of the “gold” standards is complete, so many of the FPs may be novel discoveries
- Gold standards don’t fully agree with each other
  - careful to not fit the standard to the data
- However, **comparative** evaluation is reasonable, and the numbers are likely to be close (though too conservative)

## Computational Methodology: an overview

- Machine learning methods
  - Training and evaluation
  - Bayes nets
  - Decision trees
  - Support Vector Machines
- Heuristic methods

## Machine learning methods

- Automatically learn to make accurate predictions based on past observations
- Most methods require both positive and negative training data
- Generative vs. discriminative methods

## Why Use Machine Learning?

### Advantages:

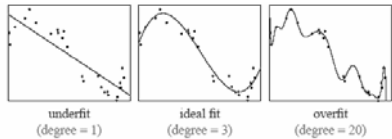
- Often much more accurate than human-crafted rules (since data driven)
- Humans often incapable of expressing what they know (e.g., rules of English, or how to recognize letters), but can easily classify examples
- Don't need a human expert or programmer
- Flexible — can apply to any learning task
- Cheap — can use in applications requiring many classifiers (e.g., one per function, one per data type, ...)

### Disadvantages

- need a lot of labeled data
  - Biology doesn't have much labeled data
  - Very few negatives
- error prone— usually impossible to get perfect accuracy

## Training and testing machine learning methods

- Separate training and test sets
- Crossvalidation
- Boosting
- Important to avoid overfitting (e.g. fitting points with a polynomial)



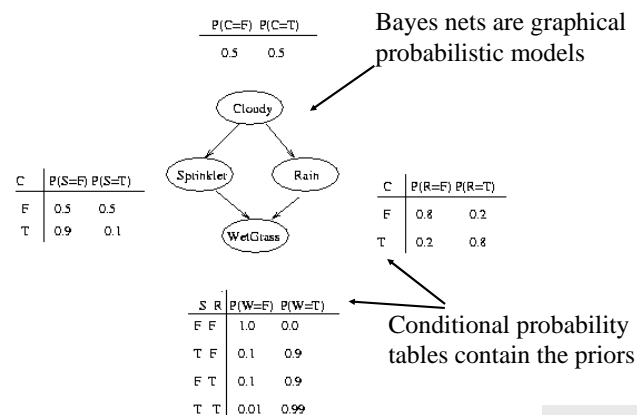
## Bayesian networks

- Graphical probabilistic models
- Can represent prior knowledge/belief
- Can be learnt from data or constructed by experts in the field
- Reasoning based on the Bayes rule

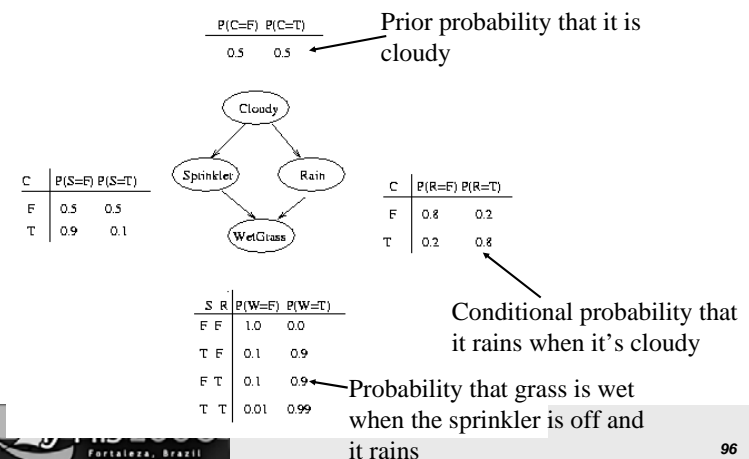
$$P(R = r | e) = \frac{P(e | R = r)P(R = r)}{P(e)}$$

$$P(e) = \sum P(e | R = r)P(R = r) \quad \text{Marginal likelihood}$$

## The sprinkler Bayes net



## The sprinkler Bayes net





## Inference $P(B|A) = \frac{P(A \cap B)}{P(B)}$

Observe: grass is wet

Two possible causes: either it is raining, or the sprinkler is on. Which is more likely? Use Bayes' rule to compute the posterior probability of each explanation.

$$\Pr(S = 1|W = 1) = \frac{\Pr(S = 1, W = 1)}{\Pr(W = 1)} = \frac{\sum_{c,r} \Pr(C = c, S = 1, R = r, W = 1)}{\Pr(W = 1)} = 0.2781/0.6471 = 0.430$$

$$\Pr(R = 1|W = 1) = \frac{\Pr(R = 1, W = 1)}{\Pr(W = 1)} = \frac{\sum_{c,s} \Pr(C = c, S = s, R = 1, W = 1)}{\Pr(W = 1)} = 0.4581/0.6471 = 0.708$$

where

$$\Pr(W = 1) = \sum_{c,r,s} \Pr(C = c, S = s, R = r, W = 1) = 0.6471$$

is a normalizing constant (probability (likelihood) of the data).



that the grass is wet because it is hood ratio is  $0.7079/0.4298 = 1.647$ .

97

## Learning Bayesian networks

- Two learning problems: structure + CPTs

Structure	Observability	Method
Known	Full	Maximum Likelihood Estimation
Known	Partial	EM (or gradient ascent)
Unknown	Full	Search through model space
Unknown	Partial	EM + search through model space

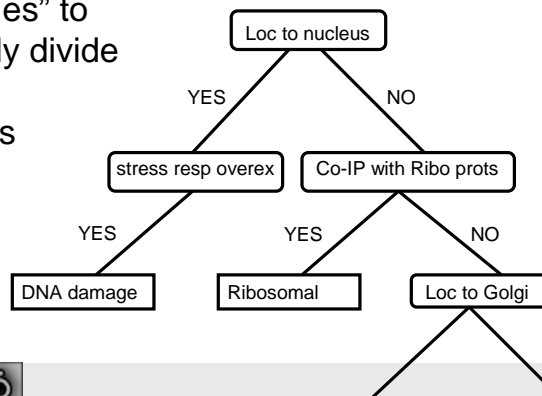
Due either to missing data or to hidden nodes



98

## Decision trees

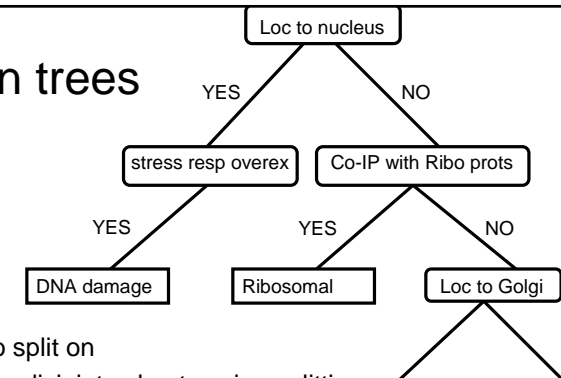
- Learn "rules" to recursively divide data into subgroups



99

## Decision trees

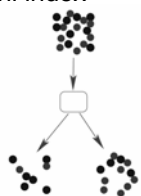
- Choose rule to split on
- Divide data into disjoint subsets using splitting rule
- Repeat recursively for each subset
- Stop when leaves are <almost> pure



100

## Splitting rules

- Best rules lead to greatest increase in purity
- Purity can be measured by
  - Decrease in entropy:
  - Gini index



$p_+$  – frac of positive examples

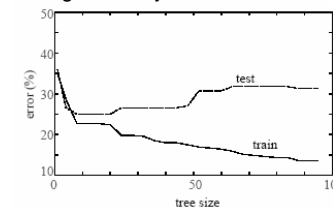
$p_-$  – frac of negative examples

$Entropy = -p_+ \ln p_+ - p_- \ln p_-$

$Gini\ index = p_+ p_-$

## Overfitting?

- trees must be big enough to fit training data (so that “true” patterns are fully captured)
- BUT: trees that are too big may overfit (capture noise or spurious patterns in the data)
- Significant problem: can’t tell best tree size from training error
- Usually grow the tree to maximize training accuracy, then prune back

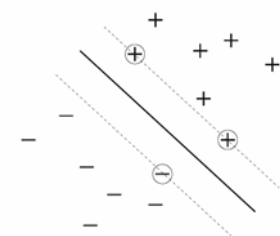


## Decision tree conclusions

- best known:
  - C4.5 (Quinlan)
  - CART (Breiman, Friedman, Olshen & Stone)
- Pros:
  - very fast to train and evaluate
  - relatively easy to interpret
- Cons:
  - accuracy often not state-of-the-art

## Support Vector Machines

- given linearly separable linearly separable data
- margin margin = distance to separating hyperplane
- choose hyperplane that maximizes minimum margin intuitively:
- want to separate +’s from -’s as much as possible
- margin = measure of confidence
- So SVMs maximize the margin



## What if data is not linearly separable?

- map into higher dimensional space in which data becomes linearly separable
- Can be done efficiently using kernels
- Pros:
  - fast algorithms now available
  - state-of-the-art accuracy
  - power and flexibility from kernels
  - theoretical justification
- Cons:
  - Not so simple to program
  - Discriminative methods require to learn a classifier for each question (e.g. each functional group)

## Heuristic methods

- Rule-based methods (e.g. predict interaction whenever more than 2 data types call it)
- Can be quite accurate and useful
- Can be hard to create good rules/heuristics
- Hard to generalize to new data types etc
- Heuristics can be combined with probabilistic evaluation to lead to effective methods
- Need extensive evaluation for accuracy and generality (same for ml methods)

## Data representation for integration

- Pair-wise representation
- Vector-based

## Data representation challenge

- Genomic data are heterogeneous
- To integrate data, it must be represented in a coherent way
- A closely related challenge is database integration (won't be discussed here)

## Pair-wise representation for gene groupings

<b>Cluster 1</b>					
Gene A		Gene	Gene	Gene	Gene
Gene B		A 1	B 1	C 1	D 1
Gene C	Gene	1	1	1	
<b>Method 2</b>					
Gene A	Gene	1			1
Gene D	Gene	1			

Matrix doesn't have to be binary – e.g. each value could be 0...1

## Vector-based representation for gene groupings

<b>Cluster 1</b>					
Gene A		Exp 1	Exp 2	Exp 3	Gene
Gene B					A 1
Gene C	Gene	0.1	0.7	-2.3	1
<b>Method 2</b>					
Gene A	Gene	4	3	0.2	1
Gene B	Gene	0.2	0.6	-1	1
Gene D	Gene	-1.3	0.4	2	0

## Data representation challenges

- Any data representation currently causes data loss
- Effective data representation can depend on the integration task (pathway vs. function prediction)
- Need to be careful of data representation – if critical part of data is not propagated through the process, even a great data integration method may not be effective (esp. important for continuous data e.g. microarrays)

## Applications of data integration (and some examples)

- Function prediction
  - Based on single data type
  - Based on integrated data
- Prediction of regulatory modules
- Regulatory networks prediction

## Function prediction based on one type of data



113

## The Rosetta Stone method

General concept

Rosetta Stone in organism 1: [A]---[B]  
 Protein A in organism 2: [A]  
 Protein B in organism 2: [B]

Top sequence = fused domain that's homologous to two separate seqs from another species

*C. elegans*  
 Ade 5,7,8: [Ade5]---[Ade7]---[Ade8]  
 Yeast Pur2: [Pur2]  
 Yeast Pur3: [Pur3]  
  
*E. coli* TrpC: [TrpC]  
 Yeast TrpG: [TrpG]  
 Yeast TrpF: [TrpF]



Eisenberg et al. Nature 405 11423

## neighbors

Observed gene locations



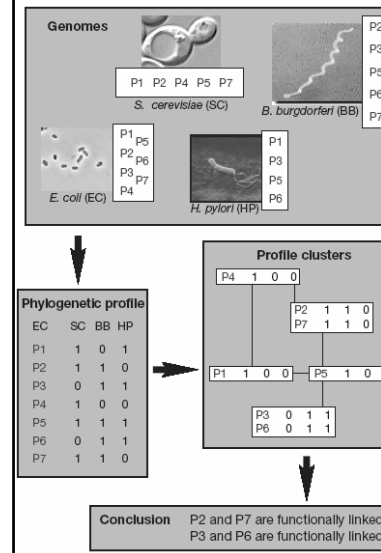
Inferred functional linkage

If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.

Eisenberg et al. Nature 405 11423

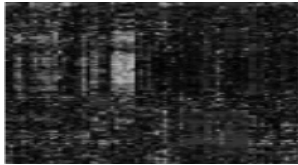
## Phylogenetic profiles method

Proteins are considered functionally linked if they share phylogenetic profiles (presence and absence in genomes). Proteins do not have to be homologous by sequence.



Eisenberg et al. Nature 405 11423

## Annotation assignment based on co-expression clusters



If enrichment for genes of a specific biological process, can claim unknowns are also involved in that process.

Prob of  $x$  out of  $n$  annotations assigned to the same GO term by chance

$P(x \text{ or more of } n \text{ genes being annotated to a particular term})$

$$\sum_{j=x}^n \left( \frac{n!}{j!(n-j)!} \right) \times p^j \times (1-p)^{n-j}$$

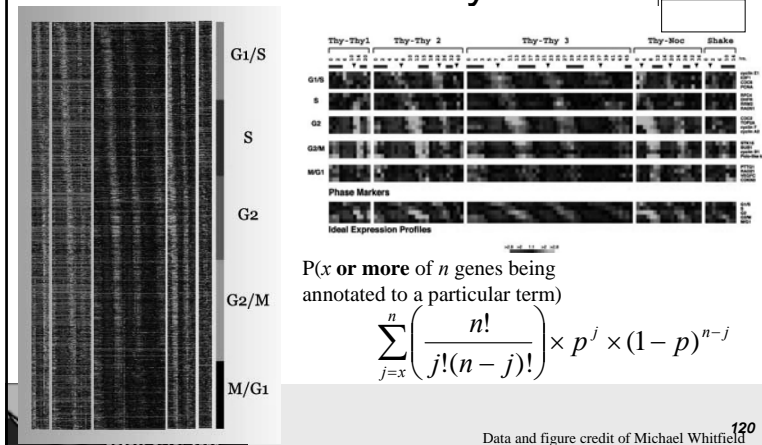
Num of permutations of  $x$  of  $n$  genes having the annotation

## Data integration for gene function prediction

## “Guilt by association” principle

- If gene  $a$  acts similar to genes  $\{b, c, d\}$  in a set of experiment
- And genes  $\{b, c, d\}$  all function in biological process  $P$
- Then by “guilt by association” gene  $a$  also functions in biological process  $P$

## “Guilt by association in microarrays”



## Proof-of-Principle of data integration: intersection-based integration

- Early methods looked for intersection or union of multiple data types
- For example, Marcotte et al. 1999
  - First paper proposing a data integration
  - A heuristic-based method for finding intersection
  - Allows to identify potential functions for a number of proteins
  - But:
    - doesn't take into account relative accuracy/coverage of methods
    - Intersection dramatically decreases coverage
    - Hard to generalize to new data in an effective way

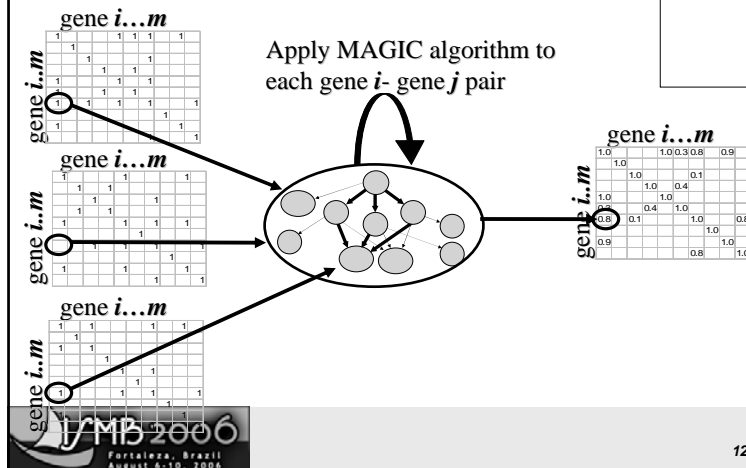
## General integration

- Probabilistic methods
  - Bayesian
  - Graph algorithms-based
- Decision tree methods
- Support vector machines
- Methods based on biomedical literature
  - Curated data – Gene Ontology
  - NLP of biomedical literature

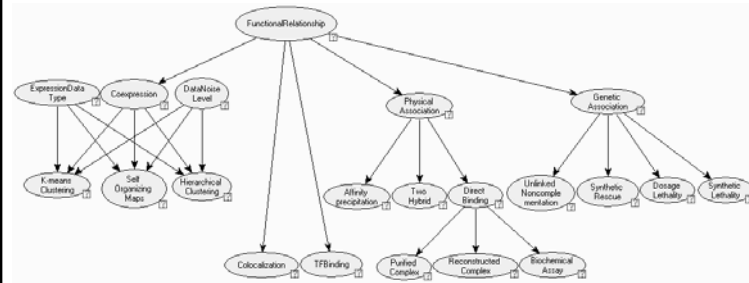
## Bayesian methods

- Several Bayesian methods proposed
  - Troyanskaya et. al 2003
  - Gerstein et. al 2003
  - Etc...
- Pros:
  - Probabilistic
  - Easy to tell which experimental sources contribute more to predictions
  - General data integration
  - Don't have to train a classifier for each functional group => even small functional groups can often be classified correctly
- Cons:
  - Not a discriminative approach, so may lose power
- Also an example of setting up the problem and evaluation for gene function prediction

## MAGIC: Multi-source Association of Genes by Integration of Clusters



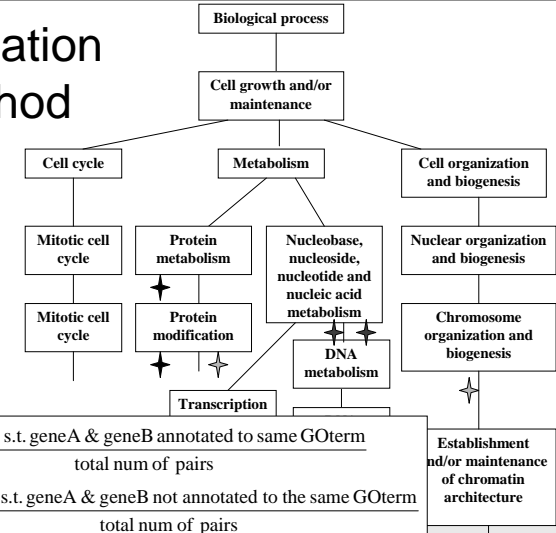
## MAGIC Bayesian network



## Evaluation method

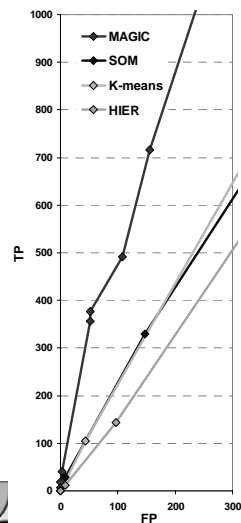
Predicted  
Gene Pairs

++  
+  
++

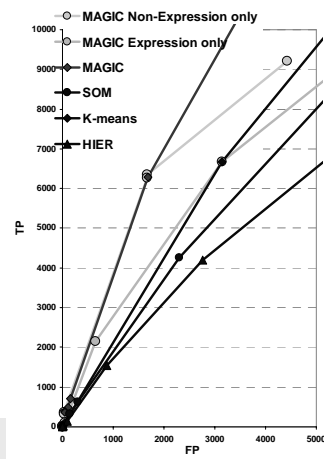


$$\text{proportionTP} = \frac{\text{num pairs s.t. geneA \& geneB annotated to same GOterm}}{\text{total num of pairs}}$$

$$\text{proportionFP} = \frac{\text{num pairs s.t. geneA \& geneB not annotated to the same GOterm}}{\text{total num of pairs}}$$

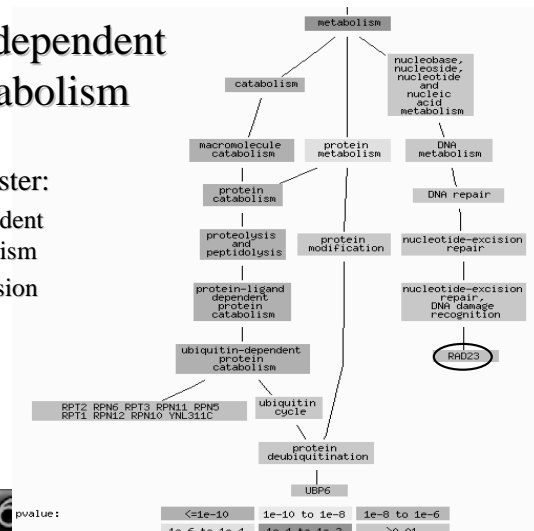


MAGIC performs better than input methods over a range of FP levels



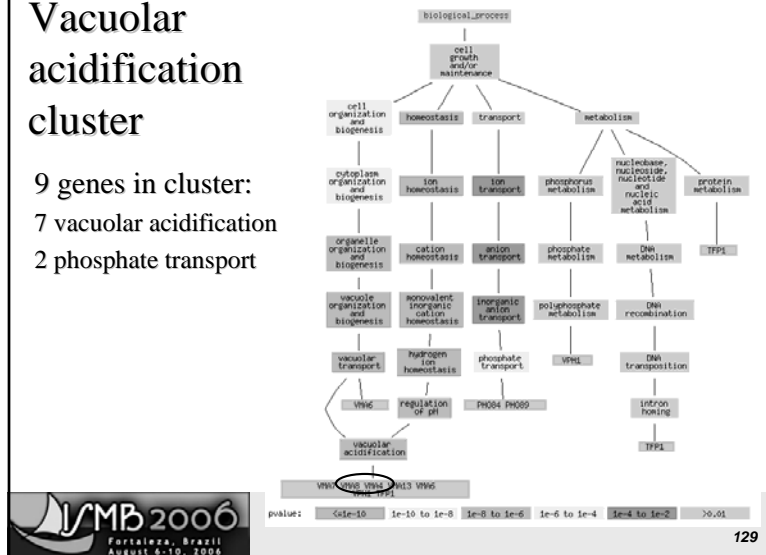
## Ubiquitin-dependent protein catabolism

10 genes in cluster:  
9 ubiquitin-dependent protein catabolism  
1 nucleotide-excision repair



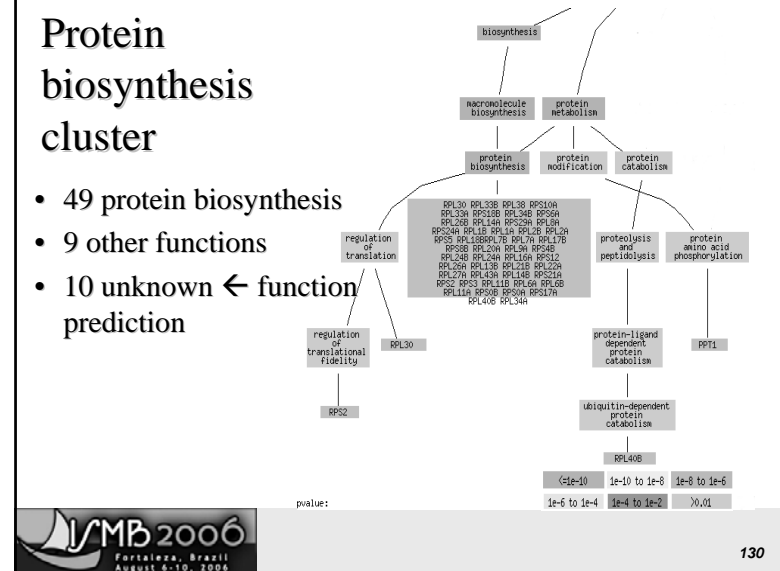


9 genes in cluster:  
7 vacuolar acidification  
2 phosphate transport



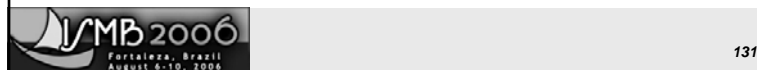
Protein  
biosynthesis  
cluster

- 49 protein biosynthesis
- 9 other functions
- 10 unknown ← function prediction



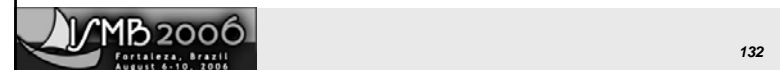
## Decision-tree-based methods

- Clare and King 2003
  - Heuristics learned based on diverse data and known functional annotations
  - Uses a modified C4.5 decision tree algorithm
- Zhang et al 2004
  - Predicted co-complexed protein pairs using probabilistic decision trees
  - Uses expression and proteomic data



## SVM-based methods

- Lanckriet et al 2004
- combined interactions, expression and sequence data by representing each input as a separate kernel
- Weighted optimised combination of these kernels used to recognize membrane and ribosomal proteins
- Pros:
  - General
  - can tell the extent to which each data source contributes to final prediction (encoded in the kernel weights)
- Cons:
  - separate classifier is built for each functional category => only possible to predict general functional categories (eg metabolism) because of lack of training data for more specific functions.

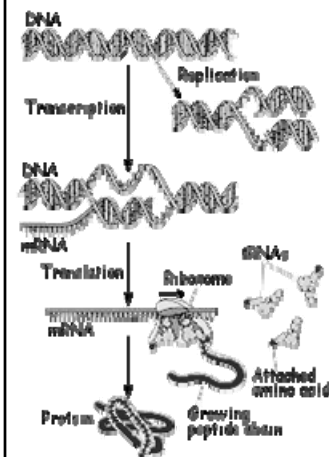


## Other approaches

- Many other approaches, for example:
- Karaoz et al. 2004
  - Combined interactions and expression data by creating a weighted protein-protein interactions graph
  - $\text{Weight}(gA, gB) \sim \text{coexp}(gA, gB)$
  - Function for unknown genes assigned based on a variant of discrete-state Hopfield network

## Data integration to study gene regulation

- Regulation and how it works
- Identifying motifs based on GE and sequence data
- Predicting regulatory modules (a case study)

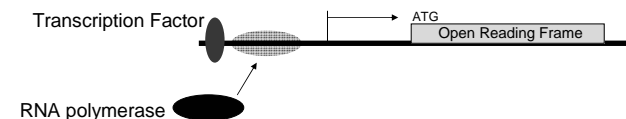


## Opportunities for gene regulation

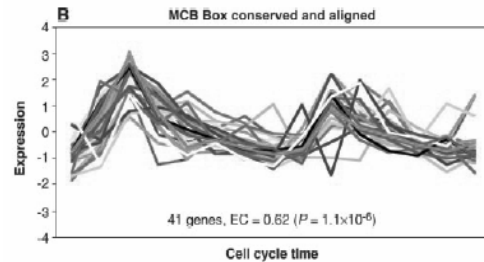
- Opening of DNA duplex
- Transcription
- mRNA stability
- Translation
- Protein stability
- Protein modification

## Transcriptional regulation

- Thought to be the most used
- Does not waste intermediate products (mRNA, protein, etc)
- But transcriptional regulation is slow, and thus may not be used in cases when fast, transient regulation is necessarily



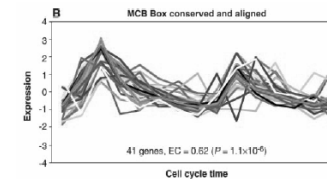
## Co-regulated genes are co-expressed



Expression profiles of 53 genes in *S. cerevisiae* genome that contain the exact match to an MCB box in their promoters (profiles normalized by mean & variance).

## Integration of expression with sequence for motif discovery

- Identify sequence motifs or motif combinations common to each group of co-expressed genes



→ ACGCGT

## Regulatory motif discovery from Gene Expression data

- Identify sets of co-regulated genes from microarrays
  - Unsupervised analysis - clustering
  - Supervised analysis
- Identify common motifs in regulatory regions of co-regulated genes
  - Combinatorial methods (enumeration with tricks)
  - Probabilistic methods (EM, Gibbs Sampling – a special case of MCMC)

## Gibbs Sampling

- Given:
  - $x^1, \dots, x^N$  sequences
  - motif length  $K$ ,
  - background  $B$ ,
- Find:
  - Model  $M$
  - Locations  $a_1, \dots, a_N$  in  $x^1, \dots, x^N$

Maximizing log-odds likelihood ratio:

$$\sum_{i=1}^N \sum_{k=1}^K \log \frac{M(k, x_{a_i+k}^i)}{B(x_{a_i+k}^i)}$$

## Gibbs Sampling (2)

- AlignACE: first statistical motif finder
- BioProspector: more recent, faster algorithm with higher accuracy

### Algorithm (sketch):

#### 1. Initialization:

- Select random locations in sequences  $x^1, \dots, x^N$
- Compute an initial model  $M$  from these locations

#### 2. Sampling Iterations:

- Remove one sequence  $x^i$
- Recalculate model
- Pick a new location of motif in  $x^i$  according to probability the location is a motif occurrence



141

## Gibbs Sampling (3)

### Initialization:

- Select random locations  $a_1, \dots, a_N$  in  $x^1, \dots, x^N$
- For these locations, compute  $M$ :

$$M_{kj} = \frac{1}{N} \sum_{i=1}^N (x_{a_i+k} = j)$$

- That is,  $M_{kj}$  is the number of occurrences of letter  $j$  in motif position  $k$ , over the total



142

## Gibbs Sampling (4)

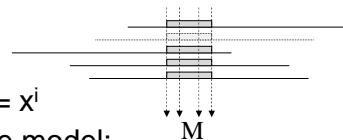
### Predictive Update:

- Select a sequence  $x = x^i$
- Remove  $x^i$ , recompute model:

Again,  $M_{kj}$  is the proportion of occurrences of letter  $j$  in motif position  $k$

$$M_{kj} = \frac{1}{(N-1) + B} (\beta_j + \sum_{s=1, s \neq i}^N (x_{a_s+k} = j))$$

where  $\beta_j$  are pseudocounts to avoid 0s,  
and  $B = \sum_j \beta_j$



143

## Gibbs Sampling (5)

### Sampling:

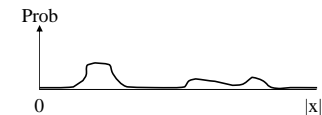
For every  $K$ -long word  $x_j, \dots, x_{j+K-1}$  in  $x$ :

$$Q_j = P(\text{word} \mid \text{motif}) = M(1, x_j) \times \dots \times M(K, x_{j+K-1})$$

$$P_i = P(\text{word} \mid \text{background}) = B(x_j) \times \dots \times B(x_{j+K-1})$$

Let  $A_j = \frac{Q_j / P_j}{\sum_{j=1}^{|x|-K+1} Q_j / P_j}$

Represents weights for sampling  
(words more different from  
background get higher weight)



How "overrepresented"  
this word is in motif vs.  
background

Sample a random new position  $a_i$  according to the probabilities  
 $A_1, \dots, A_{|x|-K+1}$  (new location for the motif)



144

## Gibbs Sampling (6)

Running Gibbs Sampling:

1. Initialize
2. Run until convergence
3. Repeat 1,2 several times, report common motifs

## Advantages / Disadvantages

- Very similar to EM (essentially EM's stochastic analog)

### Advantages:

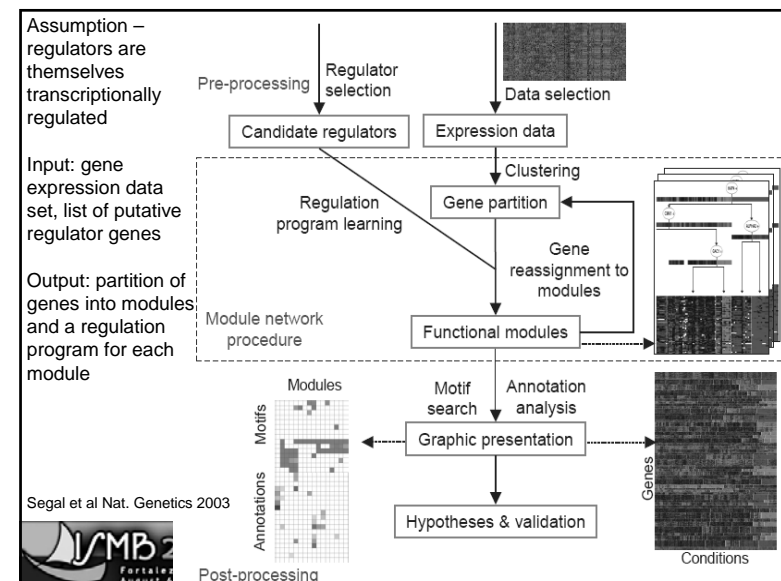
- Easier to implement
- Less dependent on initial parameters
- Less likely to converge to local minima than EM
- More versatile, easier to enhance with heuristics

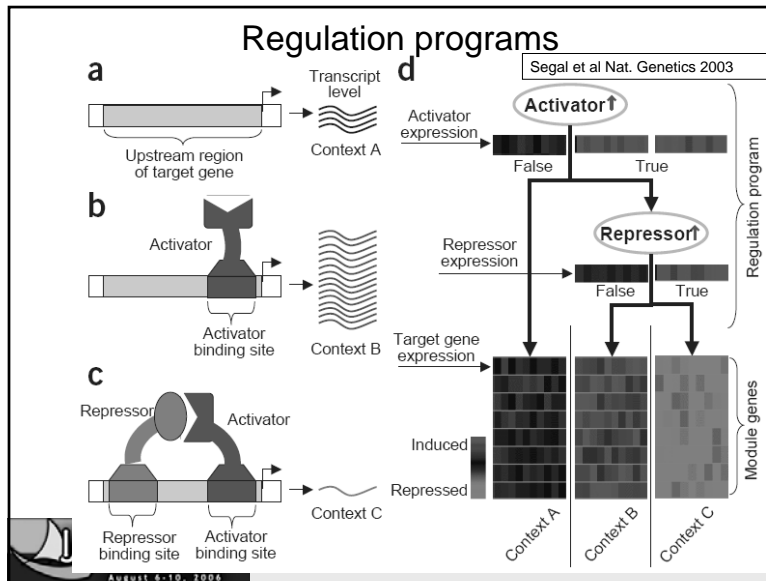
### Disadvantages:

- More dependent on all sequences to exhibit the motif
- Less systematic search of initial parameter space (doesn't converge to point estimate like EM)

## Regulatory modules

- Regulatory Module – set of genes that are co-regulated by a shared regulation program
- Knowing motifs from coexpression doesn't guarantee where the TF actually binds, or what the protein the TF is...
- Can use ChIP data in addition to GE data to identify regulatory modules
- Alternatively, can combined known regulators data with GE data to identify regulatory modules – e.g. Segal et. al Nat. Genetics 2003





## Defining regulation programs

- Regulation program specifies:
  - Set of contexts (rules describing behavior of genes in modules e.g. upregulation)
  - Response of modules in each context
- Contexts organized in regression tree
  - Decision nodes are regulators
  - Each path to a leaf defines a context using texts on the path
  - Contexts effectively specify sets of arrays
  - Context model: normal distribution over the expression of the module's genes in these arrays (mean, variance stored in corresponding leaf)
  - Small variance => tight regulation

## Learning Module Networks with EM

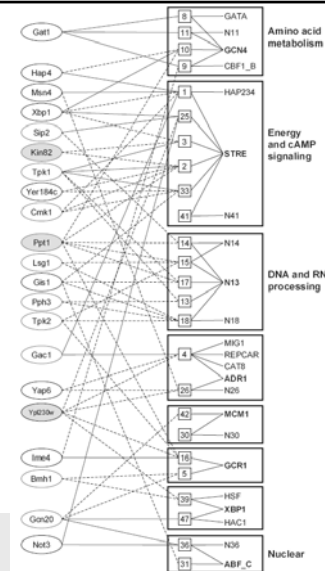
- E-step** - Given  $g$ 's inferred regulation program, find module that best predicts  $g$ 's behavior
  - For each gene  $g$ :
    - Calculate:
 
$$P(g | \text{regulatory\_program}) = \prod_{\text{arrays}} p(g_j | \text{array}, \text{context})$$
 context  $c$  of array  $j$  is defined as  $N(\mu_c, \sigma_c)$
    - Reassign gene  $g$  to the program that gives highest  $P(g | \text{regulatory\_program})$
- M-step** - Given partition of genes into modules, learn best regulation program (tree) through combinatorial search of trees
  - Tree grown from root to leaves, for each regulatory node:
    - Choose query that best partitions gene expression into two distinct distributions
    - Stop when no such split exists

## EM learning details

- Initialize with clusters
- Converges after 23 iterations to the 50 modules (initial assignments changed for 49% of genes)

## Predicted Modules

- Regulators (in ovals) are connected to modules (numbered squares)
  - Red line – regulation supported in literature, dashed line - inferred
- Module groups (boxes) share common motifs and sometimes common function
- Yellow regulators tested experimentally

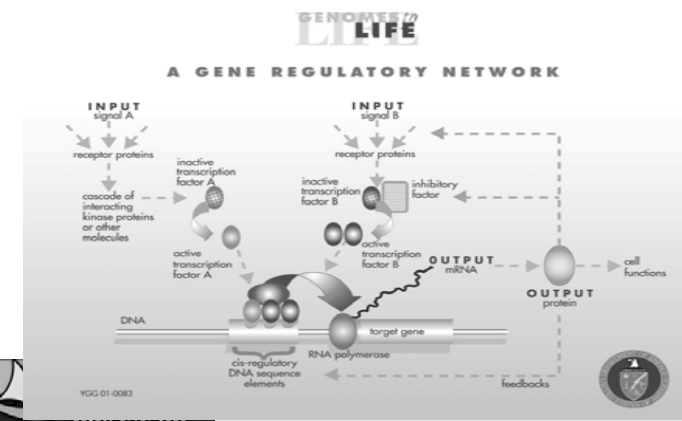


## Limitations

- Requires a set of putative transcription factors as input and a predefined number of modules
- Cannot find regulators whose expression does not change sufficiently for detection
- Cannot identify multiple regulators that participate in a regulatory even, will only identify one of them
- Can mistakenly identify a gene as regulator because it is highly predictive of a module either b/c it's a member of the module or by chance (gene has to be a member of the putative regulator set)
- Will not identify regulatory events specific to regulator and its target
- Can only handle non-overlapping modules – a gene can belong to only one module (other methods address this problem)

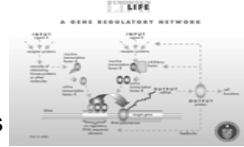
## Analysis & modeling of biological networks based on diverse data

## Biological networks – combinations of many modules

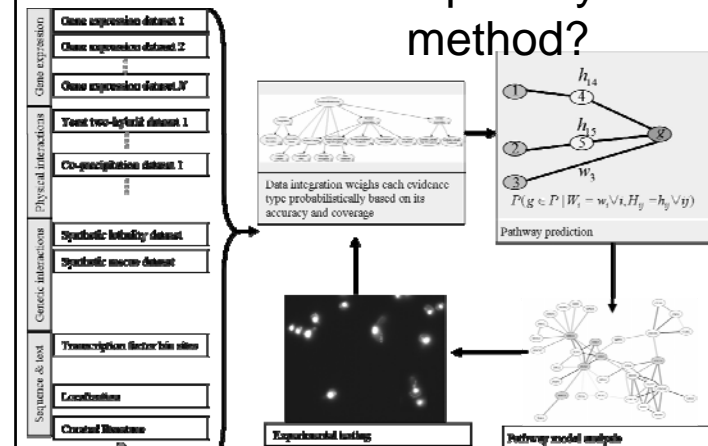


## Challenges in modeling of biological networks

- Scale of the problem:
  - networks are collections of many regulatory modules
  - Need more experimental data
  - Need more training data
- Biology is complex
  - Many different types of interactions
  - Cellular compartments may play a role
- May need a step-wise process that integrates experimentation



## Perhaps a hybrid method?



## Open problems in data integration

- No truly general & robust method for data integration available
- Data sharing still a challenge
- Integration of data from multiple organisms a promising field
- Need more experimental data
- Need better/more gold standards

## Some hopes

- Data should be available in full, with full descriptions of experiments
- Computational methods should be available for use, and their algorithms clearly explained in publications
- Clear and comprehensive evaluations, using at least GO, which currently is the most complete curated annotation (at least for yeast)



## Acknowledgements

- The Function group
- Everyone who let me use their images, as noted on each slide
- Agencies who fund us:



## Thank you! ... Questions?

Laboratory of Bioinformatics and Functional Genomics

- All complete references provided in the handout
- [ogt@cs.princeton.edu](mailto:ogt@cs.princeton.edu)
- [function.princeton.edu/](http://function.princeton.edu/)



162