

Chapter 21

JOINT LEARNING USING MULTIPLE TYPES OF DATA AND KNOWLEDGE

Zan Huang, Hua Su, and Hsinchun Chen

Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721

Chapter Overview

This chapter discusses joint learning research in biomedical domains. A brief review of the field of joint learning research is given, with emphases on the large-scale data and knowledge resources used for learning and the central biological questions involved. Two representative joint learning case studies are presented with algorithmic details. The two case studies involved two representative joint learning tasks, protein function classification and regulatory network learning, and two important algorithmic frameworks for joint learning, the kernel-based framework and probabilistic graphical models. A wide range of biological data and existing knowledge was also involved in these two studies.

Keywords

data mining; machine learning; joint learning; data and knowledge; algorithm

1. INTRODUCTION

The past decade has witnessed an explosive growth of biomedical research attributed to technological advances and the completion of genome projects for a variety of species. Genomic and proteomic technologies such as DNA microarray, genome-wide two-hybrid screening, and high-throughput mass spectrometry have generated an enormous amount of data. These data contain information that may provide a global picture of cellular composition, structure, functionality, and responses to environmental changes. Thus, data-driven discovery of functional features of genes and proteins and the complex network of cellular processes becomes a critical task for functional genomics research. On the other hand, advances in information technologies and their applications in biomedical research also resulted in a boom of biomedical ontologies that store meta-level and background biomedical knowledge as well as repositories of digitized literature text and human-curated biomedical findings. We use the word "knowledge" to refer to these resources in this chapter where there is no ambiguous interpretation based on the context. The common feature of these resources is that they are based on the biomedical literature and the common understanding of the biomedical domain, while biomedical data are obtained directly from large-scale experiments or measurements. Traditional data analysis methods are typically not sufficient for learning sophisticated data patterns from complex, large-scale biological data and knowledge with a variety of representations and granularity levels. The general lack of appropriate data analysis tools for exploiting these rich biological data resources has resulted in a boom in application and development of data mining and machine learning technologies in biomedical research in recent years.

Biomedical data mining and machine learning research is largely driven by the biological data and knowledge that is available. Researchers develop new or adapt existing machine learning algorithms to exploit special characteristics of the biological data to answer particular biological questions. Many special characteristics of biological data are associated with the inherent complexity of biological systems. A biological process such as gene regulation involves a complex network of genes, RNAs, proteins, and small molecules. Such a regulation process involves multiple layers of molecular interactions. A single type of experimental data typically reveals only a certain aspect of the underlying biological phenomena. Microarray data, for instance, delineates the global gene expression patterns in a cell and has been largely used to infer genetic regulatory networks (De Jong, 2002; Friedman et al., 2000). However, since microarray data only provides measurements at the transcriptional level and cannot reflect other aspects of

gene regulation (Jansen et al., 2002), microarray-based network learning models may infer incorrect gene regulatory relations. In addition to this inherent data deficiency, high-throughput datasets also often contain errors and noise arising from imperfections of the experimental technology which further impedes learning effectiveness.

To overcome the limitations of learning from a single type of biological data, many recent studies have been focusing on learning methods from multiple types of data. Results learned from a combination of different types of data are likely to lead to a more coherent model by consolidating information on various aspects of the biological process. Moreover, the effects of data noise in learning results will be dramatically reduced, assuming that technological errors across different datasets are largely independent and the probability that an error is supported by more than one type of data is small (Friedman, 2004).

A natural approach to exploit multiple types of data that are related to a common biological question is to combine the results learned from different types of data. We may combine the results in different ways such as accepting the learning results only when learned from all data types. More interestingly, and supposedly with promises of delivering better learning results, multiple types of data can be analyzed simultaneously under a unified framework. In the bioinformatics community such a general learning problem is called a *joint learning* problem, defined as the process of combining different types of data and learning knowledge from them in a single framework or algorithm (Hartemink and Segal, 2004).

This chapter familiarizes readers with the field of joint learning by providing a comprehensive review of the literature and two case studies. In Section 2 we review existing joint learning research using multiple types of biological data and knowledge, with a focus on the data and knowledge employed in these studies and the target biological questions. In Sections 3 and 4 we present two case studies on joint learning from multiple types of data and knowledge, with discussions on algorithmic details. The first case study describes a kernel-based computation framework to achieve optimal combination of various types of data ranging from protein-protein interactions to protein sequences to gene expression data for the problem of predicting protein functional classifications. The second study deals with learning regulatory networks from experimental data and existing biological knowledge, particularly the gene expression data and known genetic interactions. Both studies showcase the promises and challenges of joint learning research for the biomedical domain. We conclude the chapter in Section 5 by summarizing key insights, challenges, and future directions.

2. OVERVIEW OF THE FIELD

Joint learning has attracted much attention in the bioinformatics community because biologists need to learn knowledge from different types of data. Research in this area has been largely driven by data and knowledge that are made available by a wide range of recently developed technologies. Accordingly, our review first summarizes the data and knowledge resources available for joint learning and the biological questions that joint learning approaches try to address. We then review the major modeling techniques and their findings for joint learning from multiple types of data and from data and knowledge, respectively.

2.1 Large-scale Biological Data and Knowledge Resources

The data resources used in previous joint learning research can be classified into experimental data and biological knowledge. Here we refer to *data* as original experimental measurements or results, such as DNA and protein sequences, gene expression profiles, and protein-protein interaction measured by yeast two-hybrid screening, among others. In contrast, *knowledge* refers to human-curated research findings recorded in well-structured databases or documented in biomedical literature.

Biological data used by previous joint learning research include, but are not limited to, the following major categories:

(1) *Nucleotide and protein sequences*: Sequences of genes and proteins are typically represented by a string of characters written in a certain alphabet. Other sequence-related information includes length of genes or operons, the intergene distances, organization of introns and exons, and functional motif of proteins, etc. GenBank (Benson et al., 2004), as a part of the international nucleotide sequence database collaboration, is a comprehensive sequence database that contains publicly available DNA sequences for different organisms. It also provides information on taxonomy, genome, mapping, protein structure and domain, as well as the biomedical literature associated with a particular sequence. Swiss-Prot, TrEMBL (Boeckmann et al., 2003), and PIR (Wu et al., 2004) are the three major protein sequence databases, independently operating until the end of 2003. Recently they have been merged to form UniProt (Apweiler et al., 2004), a universal protein sequence and annotation database.

(2) *Protein structural information*: The 3-D protein structure is important because it determines the function of a protein. The 3-D structure of a protein can be defined by the coordinates of its crystal structure, which may suggest the location of catalytic sites of an enzyme or interaction sites by

which a protein interacts with other molecules. The Protein Data Bank (PDB) is the primary source of information on the 3D structure of proteins and other macromolecules (Bourne et al., 2004).

(3) *Gene and protein expression profiles*: Large-scale expression profiling of genes and proteins benefits from the development of DNA microarray and proteomic technologies. Microarray technology makes use of the sequence resources created by the genome projects, allowing exploration of expression patterns of thousands of genes at one time. The typical output of a microarray experiment is a matrix of gene expression measurements with rows designating genes and columns designating samples, conditions or patients, etc. There are dozens of publicly accessible repositories of microarray data, including the Stanford Microarray Database (Gollub et al., 2003) and ArrayExpress at EBI (European Bioinformatics Institute) (Brazma et al., 2003). Protein expression data has also been generated due to technological advances in biochemistry such as high resolution 2-D polyacrylamide gel electrophoresis (PAGE), mass spectrometry, and more recently, protein array technology. OPD (Open Proteomic Database) provides mass spectrometry-based data.

(4) *Protein-DNA binding data*: The interaction between protein and DNA molecules plays an essential role in gene regulation, hence is an important source of information on genetic regulatory networks. Protein-DNA binding data can be represented as binary relationships between a protein and a DNA molecule. BIND (Bader et al., 2003) and aMAZE (Lemer et al., 2004), among others, provide this type of information.

(5) *Protein-protein interaction data*: Protein-protein interactions provide information on cellular communication, signal transduction, and gene regulation. These interactions can also be represented by binary relationships between two proteins. The Database of Interacting Proteins (DIP) (Salwinski et al., 2004) is a database that documents experimentally determined protein-protein interactions. The MINT database (Zanzoni et al., 2002) stores experimentally verified protein interactions for mammalian proteomes. BIND also contains protein-protein interaction data.

Human-curated research findings are also a useful resource for joint learning. These knowledge sources may include human-encoded databases, ontologies, and biomedical literature text. Human-encoded databases and ontologies, sometimes called knowledge bases, have been created with pathway relations manually extracted from the literature to support keyword search, pathway visualization, and other data analysis and concept exploration tasks. These include databases for regulatory pathways, e.g., KEGG (Kanehisa et al., 2002) and TRANSPATH (Krull et al., 2003), and databases for metabolic pathways, e.g. KEGG (Kanehisa et al., 2002 and BioCyc {Karp, 2002 #61}). There are several large repositories of literature

focused specifically on biomedicine. Hosted by the U.S. National Library of Medicine (NLM), PubMed is one of the most comprehensive repositories of biomedical literature, containing about 14 million references on life sciences or biomedicine. PubMed has been used in many previous studies that analyze textual documents to support efficient searching and browsing and to discover new knowledge.

The URLs of all the resources mentioned here are listed in the Online Resources section of this chapter. For a more complete list and updates we refer readers to the journal of *Nucleic Acids Research*, which dedicates the first issue of each year to biological databases.

The reviewed biological data and knowledge contain rich information about cellular composition and organization. Most joint learning research focuses on tasks that help elucidate cellular processes using data from multiple information resources. One example is identification of motifs, a short sequence of amino acids or nucleotides that form the contact interfaces between two interacting protein or DNA molecules, using gene expression, transcription binding site, or protein interaction data, etc. (Li et al., 2004; Prakash et al., 2004; Takusagawa and Gifford, 2004). Another important task is classification of proteins and prediction of their function using protein sequence, protein interaction, and gene expression data (De Hoon et al., 2004; Eskin and Agichtein, 2004; Lanckriet et al., 2004). Gene regulatory networks analysis is also a popular area where joint learning approaches have been used, in which experimental data and existing biological knowledge are both involved (Chrisman et al., 2003; Hartemink et al., 2002; Imoto et al., 2003; Nariai et al., 2004; Segal et al., 2002; Segal et al., 2003; Tamada et al., 2003; Yoo et al., 2002). In the next two sub-sections, we review the previous studies on learning from multiple types of data and knowledge.

2.2 Joint Learning Using Multiple Types of Data

In this section we review previous joint learning studies using multiple types of biological data in three major learning tasks: learning regulatory networks, functional classification of genes and proteins, and motif identification. Most previous research in joint learning has been focused on these three learning tasks. As the field of joint learning is a fast-growing field that is still in its early stages of development, we can expect to see more studies targeting a broad range of biological questions in the next several years.

2.2.1 Learning Regulatory Networks

In recent years there has been great interest in the analysis of gene expression reflected in microarray data, especially in learning gene regulatory networks (De Jong, 2002; Friedman, 2004; Friedman et al., 2000; Pe'er et al., 2001; Yoo et al., 2002). However, regulatory networks learned from gene expression data alone have several limitations. Practically, the repeatability of microarray measurements is often impeded by the instability of mRNA molecules and the fact that expression changes occur within minutes following certain triggers (Emmert-Buck et al., 2000; Model et al., 2001). Biologically, microarray mainly reflects gene regulation at the transcriptional level, while other types of regulation such as protein interaction do not necessarily correlate with it (Jansen et al., 2002). RNA samples used in microarray are pooled from a large number of cells and variation among cells is usually ignored (Chu et al., 2003). Last but not least, since microarray data have the problem of high dimensionality of input space (genes) compared to the small number of available samples, the statistical power of the learned network is thus often questionable (Husmeier, 2003; Somorjai et al., 2003).

Joint learning research in this area is largely motivated by the aforementioned deficiencies of the gene expression data. Several recent studies have shown the benefit of utilizing multiple types of data for inference of more accurate regulatory networks. Hartemink et al. (Hartemink et al., 2002) used genomic location and expression data to infer genetic regulatory networks in yeast. The genomic location of the binding site of a particular transcription factor was used as prior knowledge for network inference from microarray data using a Bayesian framework. They found that the location data was complementary to expression data and hence improved network learning. Both Segal (Segal et al., 2002) and Nariai (Nariai et al., 2004) used protein interaction and gene expression data jointly to learn genetic networks in yeast. In Segal et al.'s study, the two types of data were integrated in a unified probabilistic model where genes were partitioned into clusters based on their co-expression pattern and existence of interaction between their protein products. They found that the jointly learned model matched much better with known functional gene groups and protein complexes. Nariai et al. used a Bayesian framework to integrate the two types of data by incorporating protein-protein interaction into network learning. If two proteins interact or form a complex, the genes encoding these proteins were combined as one node in the network. They claimed that by including protein complexes, the gene network was more accurately estimated from microarray data. Transcription factor binding motif has also been used in combination with the expression data. Tamada et al. (Tamada et

al., 2003) reported more accurate learning results using transcription factor binding motifs found in the promoters of genes to refine networks learned from expression data.

2.2.2 Functional Classification of Genes and Proteins

Annotating or predicting the functional class of genes and proteins is another crucial task of functional genomics. Prediction of protein function is traditionally conducted based on sequence homology between the unknown protein and proteins with known functions. Since the function of proteins is also reflected by other factors including cellular localization, expression levels under certain conditions, and interaction with other proteins, joint learning from all these types of data should provide a more complete and detailed picture of their function or functional classes.

Lanckriet et al. (Lanckriet et al., 2004) proposed a kernel-based framework to combine multiple types of data and successfully predicted the functional classes of yeast protein using an extended support vector machine (SVM) algorithm. They used five types of data including the domain structure of each protein, protein-protein interactions, genetic interactions, co-participation of a protein complex, and cell-cycle gene expression measurements. Tsuda and Noble (Tsuda and Noble, 2004) went one step further, refining the kernel representations of heterogeneous data types to improve the classification performance. Their study was used to represent protein-protein interaction and metabolic network data and demonstrated improved classification accuracy over previously used kernels. The kernel-based approach has attracted much research interest recently. Many studies have focused on developing more appropriate kernels that can represent the most relevant information of particular biological data types. Kernel-based joint learning approaches are discussed in detail in the case studies in Section 3.

Joint learning techniques have also been applied to classification of other features of genomes. For example, De Hoon et al. (De Hoon et al., 2004) used operon length, intergene distance, and gene expression information to predict the operon structure of the bacterium *Bacillus subtilis* genome. Expressing the operon prediction by each type of data as a Bayesian probability, they combined them into a Bayesian classifier, which achieved higher accuracy than those classifiers based on a single type of data.

2.2.3 Motif Discovery

Motifs are short segments of nucleotide or amino acid sequence on a DNA or protein molecule that have a particular function in molecular

interaction or catalysis. Identifying motifs in DNA and proteins helps elucidate the function of macromolecules. Motifs can be inferred from sequence data, molecular interactions, and expression data. The rationale for the association between motif discovery and expression data is that if two genes are co-regulated as detected by expression measurements, they are more likely to share the same TF binding motif. Joint learning from these types of data has shown promising results. Takusagawa & Gifford (Takusagawa and Gifford, 2004) conducted TF binding motif discovery in yeast with a unified probabilistic model. Information of TF binding from chromatin immunoprecipitation microarray experiments, together with yeast genome sequence and a statistical measure of the intergenic length, were used. They chose to learn from negative intergenic sequence, i.e. sequences that are not bound to TFs, to avoid the false positive noise, which increased the accuracy of motif prediction.

To identify motifs of protein-protein interaction, Li et al. (Li et al., 2004) used a sequential approach to learn binding motif pairs from the 3-D structure of protein complexes and sequences of interacting proteins. A binding motif pair was defined as a pair of motifs each derived from one side of the binding protein sequences. They first extracted maximal contact sequence segment pairs from the complex structural data, and then grouped the sequence data of interacting proteins using the segment pairs as templates. An iterative refinement process was also taken to enable the derivation of significant binding motif pairs.

2.3 Joint Learning Using Data and Knowledge

Besides joint learning from different types of data, another type of joint learning study is the inclusion of knowledge into the learning process. Knowledge from either documented literature or human experts provides guidance for knowledge discovery in very high dimensional data and has the potential to significantly speed up the mining process and deliver more accurate learning results. Only a few previous studies fall into this category. These studies mainly exploited domain knowledge to perform joint learning with different types of biological data. Recently, learning jointly from biomedical literature text and genomic data also started to draw significant attention. In this section we review the limited literature on joint learning with domain knowledge and text mining results. In the next section we present a detailed case study on joint learning from domain knowledge, text mining results, and experimental data.

In learning regulatory networks, Chrisman et al. (Chrisman et al., 2003) incorporated biological knowledge and expression data using a Bayesian framework. Background knowledge, i.e., known regulatory relations

between genes, was encoded as pre-specified regulatory relations in the regulatory network model. The remaining part of the network is inferred with experimental data given such partially fixed network structure. Including knowledge dramatically increased the statistical power of the learned network models, especially with a small sample size, which is typical for expression data. In a similar study, Imoto et al. (Imoto et al., 2003) incorporated protein-protein interaction, protein-DNA interactions, transcriptional factor binding site information, and knowledge from existing literature into a Bayesian framework to learn a regulatory network from microarray data. Their approach involves combining two regulatory network models, one based on the microarray expression data, and the other based on prior knowledge. They tuned the combination of the two models to find the optimal balance between knowledge and microarray data. Monte Carlo simulation and experimental data both showed the effectiveness of this approach.

One example study that incorporates text mining results and experimental data was Iossifov et al.'s study of protein interaction network inference (Iossifov et al., 2004). They used a unified probabilistic model to integrate the interactions observed from a yeast two-hybrid experiment and interactions documented in literature, which were automatically extracted by an information extraction system. Inference of the protein-protein interaction networks was performed with a Markov Chain Monte Carlo technique. However, this study was largely based on simulated data. Validity of this approach with real data is still to be tested.

Our review emphasizes the data sources and learning tasks addressed in previous work but only briefly describes specific learning techniques. We recommend readers look up the specific studies mentioned in our review for algorithmic details. With the two case studies we present in the next section, we cover two important algorithmic frameworks that have been the foundation for a large portion of the previous joint learning studies: the kernel-based framework and the probabilistic graphical models. In addition, we list several important studies in the suggested readings section.

In the next two sections we present two carefully selected joint learning cases that represent the major research areas and analytical techniques in the field. The two studies represent two major learning tasks (predicting functional classification of proteins and learning gene regulatory networks), two major types of joint learning study (learning from multiple types of data and learning from data and knowledge), and two important algorithmic frameworks for joint learning (the kernel-based framework and probabilistic graphical models).

3. KERNEL-BASED DATA FUSION OF MULTIPLE TYPES OF DATA

We discuss a principled framework based on kernel-based approaches for combining multiple types of biological data for classification problems proposed by Lanckriet et al. (Lanckriet et al., 2004). This study exemplifies how to achieve synergies among multiple types of biological data for an important learning problem, protein function prediction. Their proposed framework is also important in that it provides a mechanism for optimal combination of data with heterogeneous representations under a generic computational framework.

3.1 Protein Function Prediction

The function of an unannotated protein can be predicted based on multiple sources of information given the set of proteins with known functions. For example, it may be predicted based on an observed similarity between the sequences of the unannotated protein and proteins of known functions. The unknown protein may have functional relationships with other proteins similar to those of an annotated protein. The functional relationship between two proteins can also be inferred if they occur in fused form in some other organism, if they co-occur in multiple species, if their corresponding mRNAs share similar expression patterns, or if the proteins interact with one another.

It was pointed out by Lanckriet et al. (Lanckriet et al., 2004) that the comparison and fusion of these different data should produce a more detailed and useful representation for protein interactions. Machine learning algorithms working on such a combined data representation have the potential to provide better protein function prediction results.

3.2 Kernel-based Protein Function Prediction

The computational framework we discuss here relies on the use of kernel-based statistical learning methods that have proven very useful in bioinformatics. Under such methods data are represented by means of a kernel function, which defines similarities between pairs of genes, proteins, etc. Defined specifically for different types of data, the kernel functions implicitly capture various aspects of the underlying biological machinery. At the same time, these kernel functions provide the mapping between heterogeneous biological data and a common similarity representation. This common similarity representation then provides the foundation for

principled approaches for optimal combination of data with originally different representations.

- Kernel Methods

Under a kernel-based method, data items are embedded into a vector/feature space (typically different from the natural data presentation) and are analyzed in this space to identify data patterns. Typically nonlinear projections are performed such that nonlinear data patterns under the original data representations will appear as linear patterns in the projected feature space. Figure 1 shows an example of such a projection f with which the originally nonlinear separation between the two types of data items becomes a linear one in the projected space.

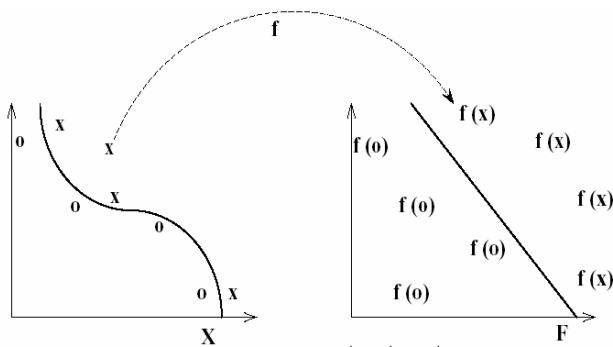


Figure 21-1. An example of data projection

Under the kernel-based approach, the projection $f(\mathbf{x})$ is specified implicitly using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \langle f(\mathbf{x}_i), f(\mathbf{x}_j) \rangle$. The benefit of such an implicit specification of projection is that to identify data patterns that only involve inner products of the data items, such as the task of similarity-based clustering, one does not need to have the explicit representation of the mapping f . It suffices to be able to evaluate the kernel function, which is often much easier than computing the coordinates of the points explicitly. Thus, quite flexible kernel functions can be applied to search for the nonlinear patterns among data items without even knowing the nature of the projected feature space. For a finite set of data items we do not even require the exact specification of the kernel function itself. All we need is a square *kernel matrix* $K = (k_{ij})$, each entry $k_{ij} = \langle f(\mathbf{x}_i), f(\mathbf{x}_j) \rangle$ is the inner products of the projected data points. This kernel matrix can be interpreted as one that describes a particular notion of similarity between data items. We will see examples of these kernel matrices later in this section.

Support vector machine (SVM) (Vapnik, 1995) is an important application of kernel-based methods to binary classification problems. In this

study we employ this algorithm to perform protein function prediction under the kernel method framework. The *l*-norm soft margin support vector machine is used in this study, which forms a linear discriminant boundary in the feature space, $g(\mathbf{x}) = \mathbf{w}^T f(\mathbf{x}) + b$. Given a labeled sample $S_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, \mathbf{w} and b are optimized to maximize the distance between the positive and negative class, allowing misclassifications:

$$\min_{\mathbf{w}, b, \xi} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^n \xi_i \quad (1)$$

$$y_i (\langle \mathbf{w}, f(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, n$$

where C is a regularization parameter, trading off error against distance. This formulation leads to the dual problem described below, for which an efficient algorithm is available (Platt, 1998).

$$\max W(\boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (2)$$

$$s.t. \sum_{i=1}^n y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, \dots, n$$

3.2.1 Kernel-based Joint Learning

Given multiple related data sets (e.g., gene expression, protein sequence, and protein-protein interaction data) one can specify a particular meaningful notion of similarity between proteins for each data set to form the corresponding kernel matrix. For example, for the gene expression data, one can use the vector similarity function to compute the similarity between each pair of genes (proteins) based on their expression profiles (expression level measurements across all experimental samples). For protein-protein interaction data we can form a protein-protein interaction network representing proteins as vertices and interactions among proteins as edges. We can then use, for example, the shortest path length between two proteins (the number of edges on the shortest path connecting the two proteins) as the kernel matrix entries. Using this approach heterogeneous data are cast into the common format of kernel matrices. With appropriate notions of similarities founded on domain knowledge, these kernel matrices could contain a significant portion of the relevant data patterns.

The common representation of kernel matrices of different data types enables us to develop principled approaches to search for the optimal combination of these data for learning tasks. An intuitive approach is linear combination. Given a set of kernels or kernel matrices $\mathbf{K} = \{K_1, \dots, K_m\}$ derived from multiple types of data, one can form a linear combination:

$$K = \sum_{i=1}^m \mu_i K_i \quad (3)$$

where μ_i 's are the combination coefficients. This combined kernel can be used to replace the single kernel in (2). Now the optimization problem is extended to not only find the optimal linear discriminant boundaries (α_i 's in (2)) but also to find the optimal combination of multiple datasets (μ_i 's in (3)). It is shown in (Lanckriet et al., 2002) that the problem of finding optimal α_i 's and μ_i 's reduces to a convex optimization problem known as a semidefinite program (SDP). Details on efficient algorithms for solving this optimization problem can be found in (Lanckriet et al., 2002; Vandenberghe and Boyd, 1996), while general search algorithms for optimization problems such as the genetic algorithms can also be employed.

3.2.2 Experimental Study

In Lanckriet et al.'s study, various types of data were combined for predicting functional classification associated with yeast proteins. They used the functional catalogue provided by the MIPS Comprehensive Yeast Genome Database (CYGD, <http://mips.gsf.de>). Using the top-level categories in the functional hierarchy a set of 3,588 proteins of 13 classes was identified. The prediction problem was then cast as 13 binary classification tasks, one for each functional class. Six types of data were used to produce input kernel matrices as shown in Table 1.

Table 21-1. Six types of protein data used in Lanckriet et al.'s study

#	Data	Similarity kernel definition $K(\mathbf{x}_i, \mathbf{x}_j)$
1	Domain structure of each protein summarized using the mapping provided by SwissPort v7.5 (http://us.expasy.org/sprot) from protein sequences to Pfam domains (http://pfam.wustl.edu). Each protein is characterized by a 4950-bit vector, in which each bit represents the presence or absence of one Pfam domain.	Inner products applied on the 4950-bit protein vectors.
2	Protein-protein interactions from CYGD	Random walk measure on a network of proteins linked by interactions (Kondor and Lafferty., 2002)
3	Genetic interactions from CYGD	Same as above
4	Co-participation in a protein complex as determined by tandem affinity purification (TAP) from CYDG	Same as above

continued

#	Data	Similarity kernel definition $K(\mathbf{x}_i, \mathbf{x}_j)$
5	77 cell cycle gene expression measurements per gene (Spellman et al., 10998)	Gaussian kernel defined on the expression profiles: $\exp(-\ \mathbf{x}_i - \mathbf{x}_j\ ^2/2\sigma)$
6	Protein sequences	Inner products derived from the Smith-Waterman pairwise sequence comparison algorithm (Smith and Waterman, 1981)

The reported results in Lanckriet et al.'s study showed that the combined kernel achieved the best performance in protein function classification. The data types that were assigned the largest combination weights in the optimal combined kernel include: protein sequence similarity, Pfam domain structure, and protein-protein interaction.

4. LEARNING REGULATORY NETWORKS USING MICROARRAY AND EXISTING KNOWLEDGE

In this section we focus on discussing joint learning from experimental data and existing knowledge. We present a case study on learning regulatory networks. This important learning problem has attracted substantial interest in bioinformatics research. We first provide some background knowledge on learning regulatory networks from microarray data and review the relevant literature in Section 3.2.1. We discuss in detail a computational framework for learning regulatory networks jointly from microarray data and known regulatory interactions in Section 3.2.2.

4.1 Learning Regulatory Networks Using Microarray

Recent advances in microarray technologies have made possible large-scale gene expression analyses based on simultaneous measurements of thousands of genes. Such high-throughput experimental data have initiated much recent research on large-scale gene expression data analysis. Various data mining techniques (e.g., clustering and classification) have been employed to uncover the biological functions of genes from microarray data. Recently, these techniques have included a reverse engineering approach to extracting gene regulatory networks from microarray data.

The general objective of gene regulatory network analysis is to extract pronounced gene regulatory features (e.g., activation and inhibition) and to reveal the structure of the transcriptional gene regulation processes by examining gene expression patterns. A simple thought experiment illustrates

the essence of the approach (Friedman, 2004). If the expression level of gene A is regulated by proteins B and C , then A 's expression level is a function of the joint activity levels of B and C . In most current biological datasets, however, protein activity levels are not available. Thus we resort to using expression levels of genes as a proxy for the activity levels of the proteins they encode. This is a problematic assumption, as the expression level of a gene does not always correlate with the activity level of the protein it encodes. Because of this fundamental data problem, the objective of regulatory network reconstruction is not to find the exact complete regulatory network corresponding to the underlying regulatory processes. We are rather looking for approximate networks and potential regulatory relations or features as hypotheses for further experiments which could lead to new discoveries.

Treating the expression levels of genes as functions of expression levels of other genes, we can rely on the variations of gene expression levels across different samples to apply reverse engineering techniques for constructing the network of regulatory relations among those genes. Taking as input a matrix of gene expression measurements, with genes and samples as the two dimensions, previous studies have proposed various network learning approaches: pair-wise comparison, differential equation estimation, and Bayesian network learning, among others.

Bayesian network learning has been the most commonly used approach for learning regulatory networks in recent years. In a Bayesian network framework, the expression level of each gene is modeled as a random variable whose value depends stochastically on other random variables that correspond to expression levels of other genes. It fits the regulatory network learning problem particularly well because of the stochastic nature of the gene expression dependencies resulted from the variability in underlying biology and measurement noise. We provide a brief description of the Bayesian network learning framework, as it is the basic learning framework adopted in the case study to be presented in this section.

Many studies have shown that regulatory networks learned from microarray data have the potential to help researchers propose and evaluate new hypotheses (De Jong, 2002; Friedman et al., 2000; Pe'er et al., 2001; Yoo et al., 2002). However, regulatory networks learned from gene expression data alone have inherent deficiencies. Biologically, the learned gene networks only provide a partial picture of the complex signaling and regulation processes. Many other important factors, including proteins and small molecules, are hidden from observation in microarray data. Technically, the small number of samples available in microarray data makes it difficult to infer statistically robust network structures. Many researchers have proposed the incorporation of other types of experimental data and existing biological knowledge to improve the statistical power of

regulatory network learning. Many previous studies included in our review in Section 2 have shown that combining microarray data with a wide range of genomic and proteomic data, such as genomic location data (Gerber et al., 2003; Hartemink et al., 2002; Segal et al., 2002), DNA sequence data (Imoto et al., 2004; Segal et al., 2002; Tamada et al., 2003), known genetic interactions (Chrisman et al., 2003; Imoto et al., 2003), and protein-protein interactions (Imoto et al., 2004; Nariai et al., 2004; Segal et al., 2003), resulted in more accurate network models.

Bayesian networks are a type of graphical models that represent probabilistic relationships among variables of interest. For a finite set $X = \{X_1, \dots, X_n\}$ of random variables, a Bayesian network $B = \langle G, \Theta \rangle$ contains a qualitative component G , which is an acyclic graph that encodes the Markov assumption that each variable (represented as a vertex in the graph) is independent of its non-descendants, given its parents and a quantitative component Θ that represents the set of parameters characterizing the conditional probability distribution. When using Bayesian networks to represent gene networks, a gene is regarded as a random variable X_i , and a relationship between a gene and its parents is represented by the conditional probability.

The problem of learning a Bayesian network (gene network) can be stated as follows. Given a microarray dataset $D = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$ of N independent instances of X , find a network $B = \langle G, \Theta \rangle$ that best matches D . The common approach to this problem is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data and to search for the optimal network according to this score. A commonly used scoring function is the Bayesian scoring metric:

$$\text{Score}(G: D) = \log P(G | D) = \log P(D | G) + \log P(G) + C \quad (1)$$

where C is a constant independent of G and $P(D | G)$ is the marginal likelihood which averages the probability of the data over all possible parameter assignments to G :

$$P(D | G) = \int P(D | G, \Theta) P(\Theta | G) d\Theta \quad (2)$$

The Bayesian scoring metric has been commonly used in the gene network learning literature (Friedman, et al., 2000; Hartemink, et al., 2002). The particular choice of priors $P(G)$ and $P(G | \Theta)$ for each G determines the exact Bayesian metric score. We followed the standard Bayesian network learning procedure using *BDe* priors (Friedman, et al., 2000; Heckerman, et al., 1995) and greedy hill-climbing search (Friedman, et al., 2000; Nariai, et al., 2004).

Figure 21-2. Bayesian networks for learning regulatory networks

4.2 Joint Learning Using Known Genetic Interactions

Among the various types of information that have been used by researchers to enhance regulatory network learning, known genetic interactions, either prepared by domain experts or collected from biological knowledge bases, probably provide the most reliable information for learning. In this section we present one of our previous studies as an example of computational approaches for learning regulatory networks using known genetic interactions and microarray data.

4.2.1 Microarray Data and Genetic Interactions

Before getting into the details of the computational approach, we first describe the data and knowledge involved in our study.

Microarray data: We used a *Homo sapiens* dataset provided by the Arizona Cancer Center. It contains expression measurements of 33 samples (11 cell lines, 2 with wild-type p53 and 9 with mutated p53, under 3 treatments with respect to introduction of exogenous p53) on a platform of 5,306 human genes. Based on discussions with the domain scientists who conducted the experiments, we selected a threshold value of 1 in logarithmic (base 2) scale and used the 200 genes with greatest variations for the network analysis. We refer to these 200 genes as *informative genes* as the specific microarray data only provides information involving regulatory relations among these genes.

Known genetic interactions: Known regulatory relations from two knowledge sources are used: gene and protein interactions found in the CSNDB database (Takai-Igarashi and Kaminuma, 1999) and regulatory relations automatically extracted from biomedical abstracts using a biological relation parser (Leroy and Chen, 2002) (refer to previous Chapters for details on biological relation parsing from literature). From CSNDB, 1,188 gene expression and cell signaling relations and 3,511 entities have been included. From 23,234 MedLine p53-related abstracts, our biological relation parser automatically extracted 1,903 relations with at least one side matched to the 200 genes in the microarray data. We analyzed the activation (1,398) or inhibition (387) relations in this study.

4.2.2 Joint Learning Approaches

The known genetic interactions were typically used to provide the ground truth regarding a (typically small) portion of the gene network to be inferred. Having part of the network guaranteed to be correct would ensure a network model of generally improved accuracy (Chrisman et al., 2003; Imoto et al.,

2003). However, this simple approach of joint learning is largely limited by the small number of known genetic interactions among a particular set of genes that show pronounced gene expression variations in a microarray dataset (we refer to this problem as the *limited overlap problem*).

The following statistics further illustrate this limited overlap problem. After gene name normalization we formed a set of known genetic interactions (contains only activation and inhibition relations) from the CSNDB database and automatically extracted from MedLine abstracts. This set consisted of 1,586 relations involving 1,212 biological entities. Only 20 of the 200 informative genes appeared in this set (these 20 genes are referred to as *overlapping* genes thereafter). In addition, there was only 1 known interaction with both ends matching the 200 informative genes. This data clearly demonstrates the limited overlap problem which limits the usefulness of approaches that directly leverage existing knowledge.

Under the direct approach of utilizing the existing knowledge for learning, the known interactions are treated as separate relations rather than a connected regulatory network that is used by biologists to make assessments and hypotheses. To be consistent with such human usages of known genetic interactions, the computational approach needs to place individual known interactions into a connected network as well. The computational approaches we present below are based on this intuition to better leverage the set of known genetic interactions for regulatory network learning.

We formed a regulatory network based on the set of known interactions and refer to this network as the *knowledge network*. We focused on analyzing the portion of the knowledge network that was reachable from the 20 overlapping genes. The resulting network consisted of 308 biological entities and 576 relations. This network serves as the basis for the following discussions.

A. Reasoning on the Knowledge Network: Qualitative Probabilistic Networks

The knowledge network is qualitative in nature, as contrasted with quantitative models such as kinetics and differential equations that provide detailed exact biological knowledge. In the biological domain there is a large amount of qualitative information emerging from functional genomic and proteomic studies. To be able to exploit such knowledge, qualitative reasoning (Forbus, 1984; Kuipers, 1986) has been applied to support simulation of qualitative information.

Qualitative reasoning has been shown to be a powerful method in the domains of medicine (Kuipers and Kassirer, 1987) and qualitative physics (Weld and de Kleer, 1990). Several studies in the bioinformatics literature have investigated qualitative reasoning and simulation of biological systems

(Heidtke and Schulze-Kremer, 1998; Karp and Mavrovouniotis, 1994; Kazic, 1993; McAdams and Shapiro, 1995; Meyers and Friedland, 1984). A rich set of qualitative reasoning tools has been adopted, including qualitative differential equations (Heidtke and Schulze-Kremer, 1998), frame-based process representation and execution (Karp, 1993), and deductive technologies (Prolog) (Kazic, 1993).

To be consistent with the probabilistic nature of gene regulatory processes and to provide reasoning capabilities that infer transitive relations and simulate gene expression levels, we employ a well-developed formalism, *qualitative probabilistic networks*. Qualitative probabilistic networks (QPNs) were designed by M. P. Wellman (Wellman, 1990) as qualitative abstractions of probabilistic networks. They have the same graphical structure as their quantitative counterparts, but instead of quantifying the probabilistic relationships between variables using conditional probabilities, they summarize the relationships in qualitative signs. The variables represent the genes and proteins while links with qualitative signs represent the inhibition/activation regulatory relations.

Given a set of biological entities $Y = \{Y_1, \dots, Y_J\}$ in the knowledge network, a QPN model contains an acyclic graph structure $G_{qual} = (V(G_{qual}), A(G_{qual}))$. The nodes, $V(G_{qual})$, represent the biological entities. The arcs, $A(G_{qual})$, represent the activation/inhibition relationships. A QPN uses a set of signs assigned to the arcs, $S^p(Y_i, Y_j)$, $p \in \{+, -, 0, ?\}$, to represent qualitative influences between two nodes. A *positive qualitative influence*, for example, of a node Y_i on its (immediate) successor Y_j , denoted $S^+(Y_i, Y_j)$, expresses that observing higher values of Y_i makes higher values of Y_j more likely, regardless of any other direct influence on Y_j . A *negative qualitative influence*, denoted by S^- , and a *zero qualitative influence*, denoted by S^0 , is analogously defined. If the influence of node Y_i on node Y_j is not monotonic or if it is unknown, we say that it is ambiguous, denoted $S^?(Y_i, Y_j)$.

The set of influences represented in a QPN exhibits various convenient and useful properties (Wellman, 1990). The *transitivity* and *composition* properties are central to the QPN reasoning used in this study. The property of transitivity asserts that qualitative influences along a chain that specifies at most one incoming arc for each node combine into a single net influence whose sign is given by the \otimes -operator from Table 2. The property of composition asserts that multiple qualitative influences between two nodes along parallel chains combine into a single net influence whose sign is given by the \oplus -operator.

For probabilistic inference with a QPN, an elegant algorithm is available, designed by Druzdzel and Henrion (Druzdzel and Henrion, 1993; Henrion and Druzdzel, 1991). The basic idea of the algorithm is to trace the effect of observing a node's value upon the probabilities of the values of all other

nodes in the network by message passing between neighboring nodes. In essence, this sign-propagation algorithm computes the sign of influence along the active trails between the observed node and all other nodes.

Table 21-2. The \otimes -operator and \oplus -operator for combining signs

\otimes	+	-	0	?	\oplus	+	-	0	?
+	+	-	0	?	+	+	?	+	?
-	-	+	0	?	-	?	-	-	?
0	0	0	0	0	0	+	-	0	?
?	?	?	0	?	?	?	?	?	?

Given: A qualitative probabilistic network and an evidence node e .

Output: Sign of the influence of e on each node in the network.

Data structures:

{ In each of the nodes
 sign ch ; // sign of change
 sign evs ; // sign of evidential support }

Main program:

for each node n in the network do $ch := '0'$;
 Propagate-Sign ($\phi, e, n, '+'$);

Recursive procedure for sign propagation:

{ $trail$ // visited nodes,
 $from$ // sender of the message,
 to // recipient of the message,
 $sign$ // sign of influence from $from$ to to }
 Propagate-Sign ($trail, from, to, sign$)

begin

if $to.ch = sign \oplus to.ch$ then exit; // exit if already made the update
 $to.ch := sign \oplus to.ch$; // update the sign of to
 $trail := trail \cup to$; // add to to the set of visited nodes

for each n in the Markov blanket* of to do

begin

$s :=$ sign of the arc; // direct or intercausal
 $sn := n.ch$; // current sign of n
 if the arc to n is active and $n \notin trail$ and $sn \neq to.ch \otimes s$ then
 Propagate-Sign ($trail, to, n, to.ch \otimes s$)

end

end

*The Markov blanket of a variable A includes the parents of A, the children of A and the variables sharing a child with A.

Figure 21-3. The qualitative sign propagation algorithm

For each node a sign is determined indicating the direction of change in the node's probabilities occasioned by the new observation given all previous ones. Initially, all node signs equal '0'. For the newly observed node, an appropriate sign is entered, that is, either a '+' for the observed

value ‘over-expressed’ or a ‘-’ for the value ‘under-expressed’. The node updates its sign and subsequently sends a message to each neighbor that is not independent of the observed node. The sign of each message becomes the sign product (\otimes) of its previous sign and the sign of the arc it traverses. Each message keeps a list of the nodes it has visited and its origin so it can avoid visiting any node more than once. Each message travels on one evidential trail. Each node, on receiving a message, updates its own sign with the sign sum (\oplus) of itself and the sign of the message. Then it passes a copy of the message to all unvisited neighbors that need to update their signs. A sketch of the sign propagation algorithm is presented in Figure 3.

With this sign propagation algorithm, we can achieve the following two basic reasoning capabilities:

(1) Deriving transitive relations: to assess the transitive influence of entity a to entity b , we assign a positive sign to node a (over-expression in a) and propagate the sign using the presented algorithm. The resulting sign of entity b specifies the transitive regulatory relation between a and b (+ for activation, - for inhibition, 0 for no effect, and ? for an uncertain relation). Because of the symmetric property of the sign product (\otimes) and sign sum (\oplus) operations, if we assigned a negative sign to node a , the sign that b will receive after propagation will be - for activation, + for inhibition, 0 for no effect, and ? for an uncertain relation. Thus the transitive relation between a and b is not dependent on what sign we assign to a .

(2) Simulating the expression levels of hidden biological entities based on expression levels of observed genes: given the signs of a set of nodes in the network, we can propagate the sign of each node throughout the entire network and aggregate the multiple effects on the nodes with unknown signs to obtain the simulated signs of these nodes under this condition.

The knowledge network obtained from the literature typically needs refinement before it can be represented as a QPN. We might get multiple (and contradictory) relations between two biological entities from the literature. In our current study our network representation does not include a relation between these two biological entities. For such relations to be included, expert validation is required to resolve the contradiction. Cycles could also appear in the knowledge network from the literature, which violate the acyclic characteristics of a QPN. Although such cycles rarely occurred in our experiments, principled approaches need to be developed to resolve this issue in future research.

B. Joint Learning Algorithms Based on Bayesian Networks and Qualitative Probabilistic Networks

Leveraging the two reasoning capabilities of the QPN formalism, we propose two joint learning approaches to combining known genetic

interactions with microarray experiment data for learning regulatory networks.

Transitive Effect Approach: The transitive effect approach derives transitive influences among the overlapping genes based on the knowledge network. The derived transitive influences are then incorporated into the Bayesian network learning process as known relations. In the current study we have employed a simple approach to incorporate the derived transitive relations. These relations were treated as if they were direct relations documented in the literature. More advanced incorporation approaches could be developed to better accommodate the stochastic and unreliable nature of the derived transitive relations.

Synthesized Expression Approach: The synthesized expression approach simulates the “expression” data of the biological entities in the knowledge network based on the observed expression levels of the genes in the overlap set. The sign propagation algorithm is used to derive the synthesized observations. For each observation, the overlapping genes are assigned the values corresponding to their observed expression levels in the microarray data (+ for over-expression, – for under-expression, and 0 for no change in expression level). The sign propagation algorithm is used to determine the sign of the remaining nodes in the knowledge network by aggregating the influences from each of the nodes in the overlap set. The derived signs of these nodes give their synthesized expression data for that specific sample. Since the synthesized expression data are derived directly from the propagated signs based on the QPN model, the derived synthesized expression data conform exactly to the QPN model from the existing knowledge. This set of synthesized expression data is then combined with the original expression data to form an enhanced gene (biological entity) expression dataset. Bayesian network learning is then applied on this enhanced data set with the relations in the knowledge network as given.

4.2.3 Experimental Results

In the experimental study we compared the learning results obtained using only the observed expression data to results obtained through learning with the transitive effect and synthesized expression approaches. The approach of directly incorporating known interactions was not considered because of the limited overlap problem exhibited in our data. Evaluation of the learned genetic regulatory relations is always a weak point of this type of research due to the lack of a set of good new hypotheses as the benchmark for comparison. In our study we employed domain experts to assess the interestingness of the learned regulatory relations as potential hypotheses. The evaluation results showed that both joint learning approaches had

improved the quality of learned regulatory networks as compared to learning without any prior knowledge. In particular, compared with learning without prior knowledge, the transitive effect approach generated a small number of additional relations surrounding the overlapping genes with high percentages of interesting relations, while the synthesized expression approach enabled the incorporation of large-scale existing knowledge into the learning process and generation of large numbers of additional, reasonably accurate regulatory relations.

5. CONCLUSIONS AND DISCUSSION

In this chapter we reviewed the field of joint learning research in biomedical domains. We presented two case studies on two representative joint learning tasks, protein function classification and regulatory network learning. These two studies involved a wide range of biological data and existing knowledge and two important algorithmic frameworks for joint learning, the kernel-based framework and probabilistic graphical models. Throughout our review and the two case studies we have been emphasizing the importance of data representation for joint learning research. To a certain extent, identifying the appropriate computable data representation is the most critical step in conducting joint learning research. With properly chosen data representations, joint learning problems may be transformed into formulations that can be solved by existing algorithms.

One critical challenge for joint learning research is to develop meaningful and scalable evaluation methods. Many learning tasks in joint learning are difficult to evaluate due to the lack of appropriate comparison benchmarks. Biological experiments that are specifically designed for data mining and machine learning purposes may need to be conducted to provide biologically-founded validation. The joint learning community also needs to address the general critiques of the biological foundation for combining various types of data and knowledge for learning. Typical machine learning validation approaches do not provide explanations for the learning performance improvements that result from the use of different types of data. This kind of explanation is crucial if biologists are to understand and utilize the learning results. Close collaboration with domain experts and formal studies justifying the underlying joint learning processes are needed.

6. ACKNOWLEDGEMENTS

The authors are supported by the grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: A Toolkit for Gene Pathway Analysis." We thank the National Library of Medicine, the Gene Ontology Consortium, and the Hugo Nomenclature Committee for making the ontologies available to researchers.

REFERENCES

- Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., and Boeckmann, B. (2004). "UniProt: The Universal Protein Knowledgebase," *Nucleic Acids Research* 32, D115-D119.
- Bader, G. D., Betel, D. and Hogue, C. W. V. (2003). "BIND: The Biomolecular Interaction Network Database," *Nucleic Acids Research* 31, 248-250.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2004). "'GenBank: Update,'" *Nucleic Acids Research* 32, D23-D26.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C. and Estreicher, A. (2003). "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003," *Nucleic Acids Research* 31, 365-370.
- Bourne, P. E., Address, K. J., Bluhm, W. F. and Chen, L. (2004). "The Distribution and Query Systems of the RCSB Protein Data Bank," *Nucleic Acids Research* 32, D223-D225.
- Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M. and Al., E. (2003). "ArrayExpress: Public Repository For Microarray Gene Expression Data at the EBI," *Nucleic Acids Research* 31, 68-71.
- Chrisman, L., Langley, P., Bay, S. and Pohorille, A. (2003). "Incorporating Biological Knowledge into Evaluation of Causal Regulatory Hypotheses," in *Pacific Symposium on Biocomputing*, Pp. 128-139.
- Chu, T., Glymour, C., Scheines, R. and Spirtes, P. (2003). "A Statistical Problem for Inference to Regulatory Structure from Associations of Gene Expression Measurements with Microarrays," *Bioinformatics* 19, 1147-52.
- De Hoon, M. J. L., Imoto, S., Kobayashi, K., Ogasawara, N. and Miyano, S. (2004). "Predicting the Operon Structure of Bacillus Subtilis Using Operon Length, Intergene Distance, and Gene Expression Information," in *Pacific Symposium on Biocomputing*, Pp. 276-287.
- De Jong, H. (2002). "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review," *Journal of Computational Biology* 9, 67-103.
- Druzdzel, M. J. and Henrion, M. (1993). "Efficient Reasoning in Qualitative Probabilistic Networks," in *Eleventh National Conference on Artificial Intelligence*, 548-553.
- Emmert-Buck, M. R., Strausberg, R. L., Krizman, D. B., Bonaldo, M. F. and Al., E. (2000). "Molecular Profiling of Clinical Tissue Specimens: Feasibility and Applications," *American Journal of Pathology*, 156, 1109-1115.
- Eskin, E. and Agichtein, E. (2004). "Combining Text Mining and Sequence Analysis to Discover Protein Functional Regions," in *Pacific Symposium on Biocomputing*, Pp. 288-299.
- Forbus, K. D. (1984). "Qualitative Process Theory," *Artificial Intelligence* 24, 85-168.
- Friedman, N. (2004). "Inferring Cellular Networks Using Probabilistic Graphical Models," *Science* 303, 799-805.

- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). "Using Bayesian Network to Analyze Expression Data," *Journal of Computational Biology* 7, 601-620.
- Gerber, G. K., Joseph, Z.-B., Lee, T. I., Robert, F., Gordon, D. B., Fraenkel, E., Simon, I., Jaakkola, T. S., Young, R. A. and Gifford, D. K. (2003). "Computational Discovery of Gene Modules and Regulatory Networks," in *11th International Conference on Intelligent Systems For Molecular Biology*.
- Gollub, J., Ball, C. A., Binkley, G., Sherlock, G. and Al., E. (2003). "The Stanford Microarray Database: Data Access and Quality Assessment Tools," *Nucleic Acids Research* 31, 94-96.
- Hartemink, A. and Segal, E. (2004). "Session Introduction," in *Pacific Symposium on Biocomputing*, Pp. 262-263.
- Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2002). "Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models," in *Pacific Symposium on Biocomputing*, Pp. 437-449.
- Heckerman, D., Geiger, D. and Chickering, D. H. (1995). "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning* 20, 197-243.
- Heidtke, K. R. and Schulze-Kremer, S. (1998). "Design and implementation of a Qualitative Simulation Model of Lambda Phage infection," *Bioinformatics* 14, 81-91.
- Henrion, M. and Druzdzel, M. J. (1991). "Qualitative Propagation and Scenario-based Approaches to Explanation in Probabilistic Reasoning," *Sixth Conference on Uncertainty in Artificial Intelligence*, Pp. 17-32.
- Husmeier, D. (2003). "Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks," *Bioinformatics* 19, 2271-2282.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2003). "Estimating Gene Networks by Bayesian Networks from Microarrays and Biological Knowledge," in *11th International Conference on Intelligent Systems For Molecular Biology*.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004). "Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks," *Journal of Bioinformatics and Computational Biology* 2, 77-98.
- Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J. S., White, K. P. and Rzhetsky, A. (2004). "Probabilistic Inference of Molecular Networks from Noisy Data Sources," *Bioinformatics* 20, 1205-13.
- Jansen, R., Greenbaum, D. and Gerstein, M. (2002). "Relating Whole-genome Expression Data with Protein-protein Interactions," *Genome Research* 12, 37-46.
- Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002). "The KEGG Databases at GenomeNet," *Nucleic Acids Research* 30, 42-46.
- Karp, P. D. (1993). "A Qualitative Biochemistry and Its Application to the Tryptophan Operon," in Hunter, L. (Ed), *Artificial Intelligence and Molecular Biology*, AAAI Press, Pp. 289-324.
- Karp, P. D. and Mavrovouniotis, M. M. (1994). "Representing, Analyzing, and Synthesizing Biochemical Pathways," *IEEE Expert* 9, 11-22.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C., & Gama-Castro, S. (2002). "The EcoCyc Database," *Nucleic Acids Research*, 30, 56-58.
- Kazic, T. (1993). "Reasoning About Biochemical Compounds and Processes," in *Second International Conference on Bioinformatics, Supercomputing and the Human Genome Project. Singapore*, Pp. 35-49.
- Kondor, R. I. and Lafferty, J. (2002). "Diffusion Kernels on Graphs and Other Discrete Input Spaces," in *International Conference on Machine Learning*, Pp. 315-322.
- Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A. and Wingender, E. (2003).

- "TRANSPATH: An Integrated Database on Signal Transduction and a Tool for Array Analysis," *Nucleic Acids Res.* 31, 97-100.
- Kuipers, B. (1986). "Qualitative Simulation," *Artificial Intelligence* 29, 289-338.
- Kuipers, B. and Kassirer, J. (1987). "Knowledge Acquisition by Analysis of Verbatim Protocols," in Kidd, A. (Ed), *Knowledge Acquisition For Expert Systems*, Plenum, Pp. 289-338.
- Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E. and Jordan, M. I. (2002). "Learning the Kernel Matrix with Semi-definite Programming," in *19th International Conference on Machine Learning*, Pp. 323-330.
- Lanckriet, G. R. G., Deng, M., Cristianini, N., Jordan, M. I. and Noble, W. S. (2004). "Kernel-based Data Fusion and Its Application to Protein Function Prediction in Yeast," in *Pacific Symposium on Biocomputing*, Pp. 300-311.
- Lemer, C., Antezana, E., Couche, F., Fays, F. and Al., E. (2004). "The AMAZE LightBench: A Web Interface to a Relational Database of Cellular Processes," *Nucleic Acids Research* 32, D443-D448.
- Leroy, G. and Chen, H. (2002). "Filling Preposition-based Templates to Capture Information from Medical Abstracts," in *Pacific Symposium on Biocomputing*, Pp. 350-361.
- Li, H., Li, J., Tan, S. H. and Ng, S.-K. (2004). "Discovery of Binding Motif Pairs from Protein Complex Structural Data and Protein Interaction Sequence Data," in *Pacific Symposium on Biocomputing*, Pp. 312-323.
- McAdams, H. H. and Shapiro, L. (1995). "Circuit Simulation of Genetic Networks," *Science* 269.
- Meyers, S. and Friedland, P. (1984). "Knowledge Based Simulation of Genetic Regulation in Bacteriophage Lambda," *Nucleic Acids Research* 12, 1-9.
- Model, F., Adorjan, P., Olek, A. and Piepenbrock, C. (2001). "Feature Selection for DNA Methylation Based Cancer Classification," *Bioinformatics* 17, 157-164.
- Nariai, N., Kim, S., Imoto, S. and Miyano, S. (2004). "Using Protein-protein Interactions for Refining Gene Networks Estimated from Microarray Data by Bayesian Networks," *Pacific Symposium on Biocomputing*, Pp. 336-347.
- Pe'er, D., Regev, A., Elidan, G. and Friedman, N. (2001). "Inferring Subnetworks from Perturbed Expression Profiles," *Bioinformatics* 17, S215-24.
- Platt, J. C. (1998). "Fast Training of Support Vector Machines Using Sequential Minimum Optimization," in Schölkopf, B., Burges, C., and Smola, A. (Ed), *Advances in Kernel Methods- Support Vector Learning*, MIT Press, Pp. 185-08.
- Prakash, A., Blanchette, M., Sinha, S. and Tompa, M. (2004). "Motif Discovery in Heterogeneous Sequence Data," in *Pacific Symposium on Biocomputing*, Pp. 348-359.
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004). "The Database of Interacting Proteins: 2004 Update," *Nucleic Acids Research* 32, D449-D451.
- Segal, E., Barash, Y., Simon, I., Friedman, N. and Koller, D. (2002). "From Promoter Sequence to Expression: A Probabilistic Framework," in *6th International Conference on Research in Computational Molecular Biology*.
- Segal, E., Wang, H. and Koller, D. (2003). "Discovering Molecular Pathways from Protein Interaction and Gene Expression Data," *Bioinformatics* 19, i264-i272.
- Smith, T. F. and Waterman, M. S. (1981). "Identification of Common Molecular Subsequences," *Journal of Molecular Biology* 147, 195-197.
- Somorjai, R. L., Dolenko, B. and Baumgartner, R. (2003). "Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions," *Bioinformatics* 19, 1484-91.
- Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein,

- D. and Futcher, B. (1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell* 9, 3, 273-297.
- Takai-Igarashi, T. and Kaminuma, T. (1999). "A Pathway Finding System for the Cell Signaling Networks Database," *Silico Biology* 1, 129-146.
- Takusagawa, K. T. and Gifford, D. K. (2004). "Negative Information For Motif Discovery," in *Pacific Symposium on Biocomputing*, Pp. 360-371.
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. and Miyano, S. (2003). "Estimating Gene Networks from Gene Expression Data by Combining Bayesian Network Model with Promoter Element Detection," *Bioinformatics* 19, II227-II236.
- Tsuda, K. and Noble, W. S. (2004). "Learning Kernels from Biological Networks by Maximizing Entropy," *Bioinformatics* 20, I326-I333.
- Vandenbergh, L. and Boyd, S. (1996). "Semidefinite Programming," *SIAM Review* 38, 49-95.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag.
- Weld, D. S. and De Kleer, J. (1990). *Readings in Qualitative Reasoning About Physical Systems*. Morgan Kaufmann.
- Wellman, M. P. (1990). "Fundamental Concepts of Qualitative Probabilistic Networks," *Artificial Intelligence* 44, 257-303.
- Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L. and Natale, D. A. (2004). "PIRSF: Family Classification System At the Protein Information Resource," *Nucleic Acids Research* 32, D112-D114.
- Yoo, C., Thorsson, V. and Cooper, G. F. (2002). "Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data," in *Pacific Symposium on Biocomputing*, Pp. 498-509.
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002). "MINT: A Molecular INteraction Database," *FEBS Letters* 513, 135-140.

SUGGESTED READINGS

- Baldi, P. and S. Brunak. 2001. *Bioinformatics: The Machine Learning Approach*, The MIT Press, Cambridge.
A good introductory book on application of machine learning to biology research.
- Buntine, W. 1996. "A guide to the literature on learning probabilistic networks from data," *IEEE Transactions on Knowledge and Data Engineering*, 8(2), 195-210.
An early review paper on regulatory network analysis.
- Cheng, J., R. Greiner, J. Kelly, D. A. Bell and W. Liu. 2002. "Learning Bayesian networks from data: an information-theory based approach," *The Artificial Intelligence Journal*, 137, 43-90.
This paper provides a complete presentation of the information-theory based dependency analysis algorithm for learning Bayesian networks.
- Chrisman, L., P. Langley, S. Bay and A. Pohorille. 2003. "Incorporating biological knowledge into evaluation of causal regulatory hypotheses," In the Proceedings of Pacific Symposium on Biocomputing, 8, 128-139.
This is one of the early studies that combine biological knowledge with gene expression data to infer regulatory networks.

- De Jong, H. 2002. "Modeling and simulation of genetic regulatory systems: a literature review," *Journal of Computational Biology*, 9, 67-103.
This is a comprehensive review of quantitative methods for modeling genetic regulatory networks.
- Friedman, N. 2004. "Inferring cellular networks using probabilistic graphical models," *Science*, 303(5659), 799-805.
This is an important recent paper summarizing the major types of regulatory network analysis using probabilistic graphical models.
- Hartemink, A. J., D. K. Gifford, T. S. Jaakkola and R. A. Young. 2002. "Combining location and expression data for principled discovery of genetic regulatory network models," In the Proceedings of Pacific Symposium on Biocomputing, 7, 437-449.
This is one of the early studies on joint learning from multiple types of genomic data for regulatory network analysis.
- Imoto, S., T. Higuchi, T. Goto, K. Tashiro, S. Kuhara and S. Miyano. 2004. "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," *Journal of Bioinformatics and Computational Biology*, 2(1), 77-98.
This is a recent study that combines relatively large-scale biological knowledge from human-curated database with gene expression data for learning regulatory networks.
- Lanckriet, G. R. G., M. Deng, N. Cristianini, M. I. Jordan and W. S. Noble. 2004. "Kernel-based data fusion and Its application to protein function prediction in yeast," In the Proceedings of Pacific Symposium on Biocomputing, 9, 300-311.
This paper proposed the kernel-based data fusion model to perform joint learning. One of the case studies in this chapter is based on this paper.
- Segal, E., H. Wang and D. Koller. 2003. "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics*, 19(Suppl: 1), i264-i272.
This is an early paper exploring joint learning with protein interaction data and gene expression data to enhance regulatory network learning.
- Speed, T. 2003. *Statistical Analysis of Gene Expression Microarray Data*, CRC Press.
An introductory book to gene expression data analysis.
- Tamada, Y., S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara and S. Miyano. 2003. "Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection," *Bioinformatics*, 19(Suppl 2), II227-II236.
Another joint learning paper exploring the combination of gene expression data and promoter elements.

ONLINE RESOURCES

DOE HGP Genomics Primers:

http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/index.shtml

Functional Genomics:

<http://www.functionalgenomics.org.uk/sections/resources/>

Gene Regulatory Networks:

<http://doegenomestolife.org/science/generegulatorynetwork.shtml>

On-line Bioinformatics Courses:

http://www.bioinformatik.de/cgi-bin/browse/Catalog/Research_and_Education/Online_Courses_and_Tutorials/

Pacific Symposium on Biocomputing:

<http://psb.stanford.edu/>

GenBank:

<http://www.ncbi.nlm.nih.gov/>

SwissProt:

<http://au.expasy.org/sprot>

TrEMBL:

<http://au.expasy.org/sprot/>

PIR (Protein Information Resource):

<http://pir.georgetown.edu/pirwww>

UniProt (Universal Protein Resource):

<http://www.expasy.uniprot.org/index.shtml>

PDB:

<http://www.rcsb.org/pdb/>

Stanford Microarray Database:

<http://genome-www5.stanford.edu/>

ArrayExpress at EBI:

<http://www.ebi.ac.uk/arrayexpress/>

OPD (Open Proteomic Database):

<http://bioinformatics.icmb.utexas.edu/OPD/>

DIP:

<http://dip.doe-mbi.ucla.edu>

MINT:

<http://mint.bio.uniroma2.it/mint/>

BioCyc:

<http://www.biocyc.org/>

KEGG:

<http://www.genome.ad.jp/kegg/>

TransFac & TransPath:

<http://www.biobase.de/pages/products/transpath.html>

PubMed:

<http://pubmed.gov>

Nucleic Acids Research (NAR) Database List:

<http://www3.oup.co.uk/nar/database/c/>

QUESTIONS FOR DISCUSSION

1. What are the advantages of joint learning from multiple types of data? What are the potential problems of learning jointly from multiple types of data?
2. What biological questions, other than those mentioned in this chapter, can be appropriately addressed using joint learning algorithms?
3. What additional types of data and knowledge can be utilized to improve prediction of the protein function classification? What meaningful kernel representations can be used for these data and knowledge?
4. Which types of genomic or proteomic data can be used for learning regulatory networks? What are the limitations of using microarray data to infer regulatory networks?
5. What are the potential problems of representing known genetic interactions documented in the literature or stored in the genetic interaction databases as binary relations in the form of “gene A activates gene B?” How could this simple representation be enhanced to capture additional information? What computational problems will be introduced by these representational enhancements?