ERAN SEGAL

Research Statement

RESEARCH OBJECTIVES

My long-term research interest is in making scientific discoveries in biology, by designing rich probabilistic model representations that integrate heterogeneous genomic data and developing efficient algorithms to learn these models automatically from data. Recent high-throughput methods have led to an explosion in the availability of such heterogeneous data, allowing us for the first time to understand cellular processes on a genome-wide scale. Transforming these immense amounts of data into *biological information* is a challenging task, and the key to success lies in our ability to combine theoretically-founded algorithms and techniques from computer science, statistics, and machine learning, with deep understanding of the biological domain. Biology is transitioning into an information science and as such, the forthcoming years will be crucial for laying out the foundations and methodologies that will be employed. Design and development of such frameworks for studying key biological questions is the central long-term goal of my research efforts.

CURRENT ACHIEVEMENTS

Probabilistic Modeling Language My thesis research has focused on developing a statistical modeling language (introduced in [5]) that can represent complex interactions in biological domains, and on applying it to study a wide range of biological problems. The models are based on the language of *relational Bayesian networks*, which represents a probability distribution over various entities in a relational domain. In the biological domain, we typically model several classes of objects, such as Gene, Array, Process, Expression, and Interaction. With each class, we also associate a set of observed attributes, such as the *sequence* of a gene or the *condition* of an array, and a set of hidden attributes, such as the *process* a gene participates in, whose value we wish to infer. Such a modeling language has several advantages. First, it allows us to explicitly represent and reason about biological entities in a modular way, as well as model the mechanistic details of the underlying biological system. Second, we can build on the sound foundation of statistical learning methods to develop algorithms that learn the models directly from data. Finally, by designing each model separately, depending on the biological phenomena we wish to study, we can use this general framework to solve a wide range of problems, and obtain biological insights directly from the model.

For example, a living cell coordinates the activation and deactivation of genes by organizing them into regulatory modules and controlling each module through a common regulatory mechanism. To discover this modular organization, including finding the regulatory modules and their actual regulators, we designed a class of models with an explicit *module* entity [6,20]. Genes in the same module are then controlled by the same set of regulators and share the same regulatory mechanism. We used gene expression data to learn these models: genes with similar expression profiles are assigned to the same module; genes whose expression is predictive of the expression of genes in the module, are assigned as the module regulators. Importantly, this design allowed us to read detailed hypotheses about gene regulation directly from the learned models. In collaboration with David Botstein's lab (Genetics, Stanford), we tested three of the novel hypotheses in the wet lab. In all cases, the experimental results supported the computational predictions, suggesting regulatory roles for previously uncharacterized proteins. There is an ongoing debate in the biological community as to the extent to which it is possible to discover regulators from gene expression data alone. Our results provided strong support for our claim that we can, indeed, induce regulation from expression. We published these findings in *Nature Genetics* [1].

Integrating Heterogeneous Data Due to deficiencies in measuring technologies and inherent redundancies in biological systems, we can view each genomic dataset only as a partial and noisy sensor of a biological process. By fusing sensors that originate from different genomic datasets, we can thus obtain much more reliable and robust analyses. Indeed, an important aspect of our modeling language is that it provides a formal framework for performing such sensor fusion. For example, we integrated protein-protein interaction data and gene expression for the task of discovering molecular pathways [4]. We designed the model such that physically interacting proteins, and genes with similar expression profiles, are more likely to participate in the same pathway. This resulted in more functionally coherent pathways compared to standard approaches, and led to potential identification of novel members of pathways. For this work, we received the best student paper award at ISMB, 2003.

In another project, we constructed detailed models of the mechanism by which patterns (motifs) in DNA promoter sequences give rise to observed expression profiles, for the task of finding *transcriptional modules* — sets of genes that are co-regulated through a common combination of motifs [3]. By integrating the sequence and expression into a unified model, we allowed for bidirectional "information flow" when learning the models:

genes with similar expression profiles are more likely to be in the same module, forcing us to find motifs in co-expressed genes; similarly, genes with common motifs affect the module assignment, forcing an organization consistent with regulatory mechanisms. Using only raw sequence and expression data as input, these models recovered many of the known motifs in yeast, several known motif combinations in human [25], and suggested novel hypotheses that are consistent with other data. For this work, we received the best paper award at ISMB, 2003. We also showed how to combine protein-DNA binding data into these models, allowing us to recover a fairly accurate model of the interactions between genes, transcription factors, and motifs in the cell cycle [9].

Multi-Species Models As biological systems are not fully optimized, some of the observed relationships may not be biologically meaningful. However, since the functionally relevant relationships confer a selective advantage, they are more likely to be conserved across evolution. Thus, we can find the key relationships by combining data from different organisms. In collaboration with Stuart Kim's lab (Developmental Biology, Stanford), we developed a method to identify pairs of genes that are co-expressed in multiple organisms, and applied it to 3000 arrays from human, fly, worm, and yeast. This conserved co-expression network provided global insights about evolutionary principles, and predicted the known function of genes significantly better than single species networks. We also presented wet lab experiments supporting some of the novel functional predictions. We published these findings in *Science* [2]. We are now enriching the expressive power of our probabilistic framework to multi-species models in order to address a broad range of biological questions regarding evolution. In particular, we are collaborating with Matthew Scott's lab (Developmental Biology, Stanford) on developing methods for detecting conserved regulatory pathways and studying the degree to which regulatory relationships are conserved.

Learning Learning our models automatically from data poses great computational challenges, stemming from the many inter-dependencies that exist between the various biological entities and from the vast amounts of biological data we include. As part of this learning task, we had to reason in some of the largest and most densely connected graphical models constructed so far, some with over two million hidden variables. As reasoning in such models is intractable, we developed several learning algorithms that exploit problem-specific biological structure, such as the context-specificity of transcription factor binding or the modularity of gene regulation, leading to efficient algorithms. These learning algorithms require us to infer the values of hidden variables. In our models, such inference cannot be done exactly, and we thus scaled up existing approximate algorithms, as well as developed novel approximate inference algorithms [3,7].

Visualization For computational tools to have a broad impact, they must be accompanied by visualization and browsing tools that are easily accessible to biologists. To support this effort, we developed GeneXPress, a software environment for visualization and statistical analysis of genomic data, including gene expression, sequence data, and protein-protein interactions. Currently, 383 scientists from 41 countries are using GeneXPress.

FUTURE DIRECTIONS

Genomic datasets, spanning many organisms and data types, are rapidly being produced, creating new opportunities for understanding the molecular mechanisms underlying human disease, and for studying complex biological processes, such as development and gene regulation, on a global scale. It is clear that computational tools will play a major role in realizing these opportunities, but the challenge will be to develop methods that extract meaningful *information* from the vast amounts of raw data. My long-term research goal is to address these challenges by constructing probabilistic models that integrate heterogeneous data from different organisms and exploit the modularity in biological systems for obtaining efficient representations and learning algorithms. When new types of measurements, such as protein expression levels, sub-cellular localizations, and tissue specific expression levels become available, I plan to design new models for incorporating them in order to obtain a more reliable and complete view of the biological system.

With the growing availability of genomic data, a key challenge is to unravel the genetic blueprint of the cell: identify all the genes, their biochemical functions, physical interactions, and involvement in processes. The next challenge is to understand how these 'parts' assemble into higher order functional units and how these units interact to give rise to fully functional organisms. Thus, I plan to develop higher level representations whose basic building blocks correspond to functional units such as pathways or transcription factor targets. The models will then characterize biological conditions in terms of these units, leading to more informative views of organisms, as we show in a global analysis of human cancer [24]. In constructing these models, we need to address several challenges, such as modeling the uncertainty of gene membership in units, integrating gene level models to support our higher order conclusions, and identifying novel functional units. I believe that this challenging and exciting line of research can lead to qualitative leaps in our understanding of how cells and organisms work.