Mike P. Liang, Olga G. Troyanskaya, Alain Laederach, Douglas L. Brutlag, *and* Russ B. Altman

Computational Functional Genomics

Biology is evolving into an information-rich science due to advances in high-throughput (large scale) experimental techniques that produce hundreds of thousands of data points in a single measurement. The field of computational functional genomics exploits this wealth of information to create models that begin to describe function in a systematic way. A fundamental aspect of computational functional genomics is the problem of functional identification. All genes in an organism have a specific function or functions. This function describes the role that a gene plays in the cell and is generally determined experimentally. The high-throughput experimental techniques mentioned above do not measure function directly, but rather measure features of a gene that are related to its function (e.g., the sequence of the gene). The relationships between the function of a gene and its features are often complex and not well understood, making machine learning algorithms ideally suited for the analysis of this type of biological data. Given a training set of genes with known features and function, a model can be constructed using machine learning to predict the function of all the genes in the organism. Furthermore, the features of an organism can be analyzed in an unsupervised framework to identify genes that have common features and thus possible common or related functions. In this article we present some specific examples of how representing biological data in a machine-learning framework is possible and how these representations contribute to both the prediction and discovery of biological function.

Introduction

High-throughput experimental methodology has transformed biology into an information-rich science. It is now possible to rapidly obtain data on thousands of genes in a single experiment. In the next decades, data from these experiments may provide an integral understanding of biological systems. The depth of understanding of a biological system reflects the accuracy with which it can be simulated and engineered. It is therefore critical that the understanding of biological systems be pushed to the deepest level possible. However high-throughput (large-scale) biological data sets are complex, incomplete, and noisy, and thus development of sophisticated computational methods is necessary.

The first step towards a system-wide understanding of a biological system is functional identification of its components. Currently, approximately one quarter of the yeast genome (http://www.yeastgenome.org/) and greater fractions for other organisms have no known function. Recent breakthroughs in large-scale experimental methods have resulted in the ability to measure specific characteristics of a biological system (e.g., the sequence of a gene or its level of expression) in a high-throughput manner, opening the possibility of automated functional identification. Automated functional identification is based on the principle of functional similarity, such that two genes that share common characteristics will generally share common function. For example, two genes with similar DNA sequences from different organisms will most likely have the same function in their respective organism.

The problem of functional identification of all the genes in a biological system, or functional genomics, is thus one of pattern recognition. Each gene can be characterized by a set of features (e.g., the gene sequence) that can be obtained in a high-throughput manner. Previous experimental work has identified the function of some of these genes. The features of these characterized genes can be used as a training set in a supervised framework for detecting relationships between the features and their respective functions. These patterns can in turn be used to assign (or annotate) putative function to all genes that have been characterized by high-throughput experiments. Moreover, identification of common features in an unsupervised framework may suggest new functional relationships and provide the experimentalists with direction for rational experimental design. In both the supervised and unsupervised frameworks, success is

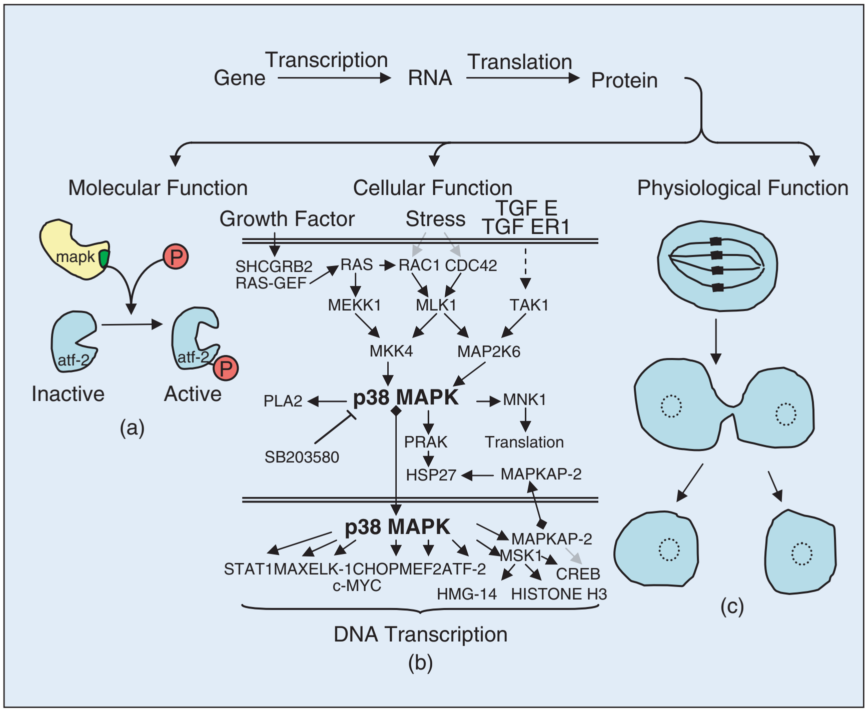
1053-5888/04/$20.00©2004IEEE

contingent upon several criteria. First and foremost, proper feature selection is critical because not all features may be informative. The identification of the appropriate features requires a strong understanding of the underlying biological process. However, the biological process is often poorly understood, making automatic feature selection an important and challenging component of computational functional genomics. Second, proper training set construction and validation is important because biological data sets are inherently biased. Bias is introduced in the data sets because the biologist's decision to study a particular system is influenced by the experimental complexity of the system. For example, organismal development has been extensively focused on the fruitfly because it develops in only a few days. These types of bias must be taken into account to construct general models. Ultimately, the success of any method is measured by its biological accuracy and significance. A successful method produces results that are not only predictive but also lead to a more comprehensive understanding of the underlying biological mechanism.

Functional identification is context specific, such that the function associated with a gene is dependent on the scale at which it is being studied. For example, a gene in the molecular context encodes a protein that has the potential to catalyze a chemical reaction in the cell (e.g., a kinase catalyzes a phosphorylation reaction [Figure 1(a)]). In a cellular context, the phosphorylation reaction can activate a signal transduction cascade [Figure 1(b)]. In the physiological context, the activation of this cascade ultimately leads to cell division [Figure 1(c)]. Although the biological function assigned to the gene is dependent on scale, machine-learning frameworks can nonetheless be used to model and identify function at every scale. The examples of functional genomics research presented below are organized based on the concept of biological scale. The purpose of this article is to review some of the recent work that has been done in computational functional genomics, specifically illustrating how better representations of the data, careful selection of a training set, and better machine learning algorithms can significantly improve functional identification. We present the problem of functional identification in the context of each biological scale. Then we conclude with specific challenges in computational functional genomics where the signal processing community may be able to make significant contributions. This review does not attempt to be comprehensive, and examples have been chosen from our own work and the work of others simply to illustrate the challenges and the types of approaches taken.

Functional Identification

Biological systems are comprised of a multitude of proteins that work in concert to afford the system remarkable efficiency and adaptability. These proteins generally facilitate (or catalyze) specific chemical reactions that allow the system to operate. Genes encode protein sequences, and when one refers to the function of a gene, it is in fact the function of the corresponding protein that is implied. Genes are



▲ *1. The central dogma of molecular biology describes the flow of genetic information as going from the gene (DNA) to RNA (an intermediary molecule) and then to protein. The function of a gene is defined by the function of the protein encoded by the gene. Protein function can be studied in various contexts, illustrated here by an example of the p38 mitogen-activated kinase (MAPK) protein. (a) At the molecular level, MAPK transfers a phosphate group onto another protein to activate it. (b) At the cellular level, MAPK plays a critical role in the MAPK signaling pathway. (c) At the physiological level, activation of the MAPK signaling pathway can serve as a signal leading to cell division.*

comprised of DNA. Cells have developed extremely efficient and accurate molecular mechanisms for replicating DNA. Current high-throughput experimental techniques exploit these mechanisms to rapidly and efficiently measure specific features of genes. Thus, genes are experimentally much easier to characterize than proteins. Genes are also fundamental elements of living systems and thus remain the center of interest for the majority of biological labs. For these reasons, most high throughput biological data is genetic (e.g., gene sequences, gene expression levels).

To treat data in a machine-learning framework, it is imperative to properly represent it. The scale of biological data is broad, ranging from detailed molecular structures to organism wide phenotypes. In this section we review recent advances in the development of novel representations for this data and demonstrate how these representations simplify the application of various machine-learning protocols. The following sections are organized according to biological scale, ranging from molecular to physiological scale. Although the systems change, the fundamental theme of machine learning applied to biological systems remains, and for each example, the different aspects (e.g., features, training set) of the system are clear.

Molecular Function

Molecular function is the biochemical or biophysical activity of a gene. These activities include catalysis of chemical reactions, binding to small molecules or atoms, and interaction with other biological macromolecules. Understanding what molecular function a gene can perform as well as where the activity occurs can provide powerful insights on how the gene works, how the gene may be modified to change its function, and even how to inhibit or activate the function for medical treatment. Two types of data are used for computational studies of molecular function: the primary sequence of the gene and the three dimensional (3-D) structure of the protein encoded by that gene. A key challenge in analyzing molecular function is identifying important functional sites in the protein given its sequence or structure.

Using Protein Sequence Data for Identifying Molecular Function

Protein sequence data is represented as a string with an alphabet size of 20, one for each amino acid type. The universal genetic code relates gene and protein sequences such that if a gene is sequenced, the protein sequence can be easily known. The genome sequencing projects have thus been a major source of protein sequence data. Protein sequence databases like SwissProt [1], UniProt [2], and the Protein Information Resource [3] provide rich repositories for protein sequence data and functional annotation.

A common supervised approach for using protein sequence data to identify molecular function for an unknown protein is to identify known proteins with similar sequence. The process of comparing sequences is called sequence alignment. Sequence alignment places protein sequences such that the optimal number of letters (amino acids) in the sequences is matched (often with gaps and substitutions allowed). Smith-Waterman [4] has addressed optimal local alignment of pairs of sequences using dynamic programming. This optimal alignment method is not always suitable for analysis on large databases of sequences because of their runtime complexity. Heuristic methods like BLAST [5] have been developed to rapidly identify sequences in large databases that are similar to a query sequence. It is important to note that reliable assignment of function requires the query sequence to have sufficiently high sequence similarity to a previously functionally characterized sequence.

Sequence-based methods often focus on specific areas of the protein sequence that are relevant to the protein's function. Mutation or deletion of amino acids in one of these functional areas will negatively affect protein function. As a result, if a set of protein sequences from different organisms are aligned, regions that are well matched (or conserved) among the sequences may correspond to functionally important amino acids. Aligning a set of sequences is known as multiple sequence alignment (MSA). Multiple sequence alignment is a challenging problem and an active area of research. A common computational tool for multiple sequence alignment is ClustalW [7]. Identifying and representing patterns in these alignments is important for predicting functional sites on protein sequences because functional sites often correspond to highly conserved amino acids in the multiple sequence alignment. This pattern of conservation, or profile, is characteristic of the protein's function and can therefore be used to identify function in unknown proteins. Furthermore, the conserved amino acids generally play critical roles in the protein's function (e.g., negatively charged amino acids like glutamate and aspartate often play the role of a base in acid-base catalysis), and identification of these residues can significantly enhance the understanding of the catalytic mechanism of the protein.

An important challenge with multiple sequence alignments is efficient representation of the pattern of conserved amino acids in the alignment for rapid searching of such patterns in sequence databases. PROSITE [8] uses the simplest possible representation of a pattern, a regular expression. Pfam [9], on the other hand, uses a more sophisticated representation, a hidden markov model. Recently, Nevill-Manning et al. have developed a method, eMotif [10], that produces a range of sequence patterns that result in varying degrees of specificity and sensitivity. The highly specific motifs are very useful in assigning function rapidly with few false predictions, while the more sensitive motifs are used to infer weaker relationships that can then be experimentally verified. Finally, Wu et al. developed a more sensitive representation of conserved regions with position specific weight matrices, or profiles, in a method called eMatrix [11]. Profiles specify the probability of occurrence for the amino acids at each position in the conserved region. In many instances the specificity and accuracy of profile-based methods for functional identification is sufficient to correctly annotate the majority of genes in a biological system. This is particularly important because the rate of gene sequencing is increasing exponentially, and annotation of these novel sequences must be carried out in an automated manner.

There are specific biological examples where only a small number of conserved amino acids are necessary for a protein to maintain its function. In these cases, it is very difficult to use sequence similarity to identify function. In fact, recent sequencing projects have revealed that more than half of new genes discovered do not have sufficient sequence similarity to known genes to transfer functional annotation. The information content of sequence data alone may therefore not be sufficient for functional annotation. This has resulted in the development of computational techniques that include other sources of data such as the 3-D structure of proteins.

Using Protein Structure Data for Identifying Molecular Function

The 3-D structure of a protein is becoming an increasingly important source of data for identifying molecular function. Structure and function are intimately related, and this relationship is only revealed when the detailed 3-D atomic structure of a protein is obtained experimentally. The function or activity of a protein can generally be attributed to a specific arrangement of atoms within what is known as the active site of the protein. Structural genomics initiatives [12] are attempting to determine the structure of all proteins on the basis of the direct relationship between structure and function. The Protein Data Bank [13] (PDB: http://www.pdb.org/) provides a central resource for storing publicly available protein structures (mostly from X-ray crystallography and NMR spectroscopy) and is an excellent source of structural data.

The simplest approach for using 3-D structure in a machine learning framework is to use the 3-D atomic coordinates as features. PROCAT [14] identifies conserved atoms in proteins and searches new 3-D structures for occurrences of these structural templates using geometric hashing. A disadvantage of working in Cartesian space is that the atomic positions are not perfectly conserved, resulting in noisy data. Alternatively, Fetrow et al. use distances between atoms as features (FFF [15]), as these have been found to be more highly conserved because molecular function generally depends on the relative distance between atoms and not their absolute coordinates.

In some situations, however, molecular function is still maintained even if one atom is replaced with another, as long as it has similar physicochemical properties. In these situations, using the 3-D coordinates or relative distances is not sufficient; however, conservation of physicochemical properties can be used to identify function. Wei et al. have developed a tool, FEATURE [16], for describing the 3-D structural environment around a functional site using the distribution of physicochemical properties in its microenvironment. Given a set of structures representing a site and a background set of structures that do not have this site, FEATURE automatically identifies statistically significant properties in the environment that discriminate the site from the background. In addition, it provides a statistical 3-D model of the functional site which can be used to predict locations of functional sites on new protein structures. FEATURE uses naïve Bayes to perform discriminant analysis of example sites from nonsites and uses a nonparametric Wilcoxon rank sum test to identify properties significant for function.

For FEATURE to be effective in genomic-scale analysis of protein function, a library of models of functional sites is necessary. Liang et al. have developed SeqFEATURE [17], that integrates sequence analysis information, using sequence motifs, with structural analysis information used by FEATURE to automatically generate a library of models. By integrating the conserved physicochemical properties in the structural environment around sequence motifs, Liang et al. have shown that the resulting structural motif has better performance in predicting functional sites than the sequence motif alone. Because the method is fully automatic, a library of models can be created from a database of sequence motifs. FEATURE and SeqFEATURE provide a way of annotating protein structures with protein function useful for high-throughput annotation of protein function.

Cellular Function

Cellular function is the regulation of

genes and interaction of proteins to perform cellular processes, such as cellular growth, communication, and metabolism. The central dogma of biology states that genes (DNA) are transcribed into ribonucleic acid (RNA), and that RNA is in turn translated into proteins. The regulation of a gene at the transcriptional or translational level thus results in varying activity of the protein product. The level of transcriptional expression of a gene can be measured experimentally using a variety of "GeneChips" or microarrays in a high-throughput manner. It is now possible to measure the relative expression level of over 60,000 genes in a single experiment.

When a gene is transcribed, it is said to be turned on or expressed. Careful regulation of both the timing and the amount of gene expression is required for proper cellular function. Defects in regulation can result in serious consequences for the cell such as uncontrolled cellular growth in cancer metastasis. Proteins interact together in pathways to perform cellular function. There are many types of pathways in a cell, including signaling pathways and metabolic pathways. In signaling pathways, a signal from external stimuli or from a change in the state of a cell is propagated from one part of the cell to another. In metabolic pathways, the cell breaks downs or creates substances to generate energy and to maintain its function. In general, when referring to the cellular function of a gene, the reference is actually to the function of the pathway that the gene participates in. If two genes have the same cellular function, then they participate in the same pathway. Cellular function of a gene can be identified using gene sequence data, microarrays, biological literature, and other data sources. Computational challenges include predicting how genes are regulated, which proteins participate in a pathway, and how proteins interact in a pathway.

Using Gene Sequence Data for Identifying Cellular Function

Special proteins known as transcription factors bind to regions near a gene and regulate gene expression. These specific regions are called transcription factor binding sites and are a type of regulatory element that controls gene expression. An important strategy for identifying gene function is finding regulatory elements in the genes. If a new gene shares a common regulatory element with other genes, the cellular function of the new gene may be inferred from the function of the other genes. Many methods have used this principle to find common regulatory elements from a set of gene sequences.

Using Gene Expression Data and Other Data Sources for Identifying Cellular Function

Genes in the same pathway often have similar expression profiles. A microarray experiment is a powerful way of measuring the level of expression of all genes in a cell in one experiment. Successive microarray experiments can take snapshots of gene expression under different conditions or over a period of time. Finding correlated patterns of gene expression across these experiments can group genes that belong to the same pathway. In addition, by comparing gene expression across different experimental conditions (such as cancer cells versus normal cells), genes that have differential expression can be used as indicators for those conditions. This indirect method of assigning function based on coregulation or interaction is often termed the "guilt by association" method of assigning function. Although genes with the same cellular function often have correlated expressions, the converse is not always true, a limit of the "guilt by association" approach. Genes with correlated expression do not necessarily have the same cellular function. Finding genes that truly have common cellular function is a key challenge in analyzing gene expression data. Stuart et al. address this issue by looking for conserved correlation of expression across *diverse* organisms [18]. If the same genes have correlated expression across various organisms, their cellular function is probably a core function critical for survival. Hence, correlated expression of genes across diverse organisms strongly indicates that the genes have common cellular function.

Protein-protein interactions are also very important for understanding cellular function because proteins that interact are likely to participate in the same pathway. Yeast two hybrid and co-immunoprecipitation experiments are methods for rapidly identifying protein-protein interaction. By performing interaction experiments on all pairs of proteins, an interaction matrix can be generated. These experiments can generate large amounts of data that are often noisy or contain missing values. A key challenge is to find proteins in these interaction matrices that are part of the same pathway as well as to identify their cellular function.

Microarray and yeast two-hybrid experiments are two examples of genome-wide experiments where quality is often sacrificed for scale, resulting in highly noisy data. However, accurate biological conclusions can still be made based on these data if heterogeneous data sources are examined simultaneously. Segal et al. have built a method for predicting pathways by combining gene expression data and transcription factor data in a relational Markov network [19]. Troyanskaya et al. incorporate even more disparate sources with their method MAGIC (multisource analysis by grouping and integration of clusters) [20]. MAGIC combines data from protein interactions, genetic associations, transcription factor binding sites, and gene expression analysis in a Bayesian network. Troyanskaya et al. found that by integrating multiple sources, predicting cellular function of interacting genes was considerably more accurate than using individual data sources alone.

The biomedical literature is also an important source of functional information. This large body of information documents the results of decades of study on cellular function. Public databases like PubMed (http://www.pubmed.org/) store abstracts and links to full text articles of these documents. Article abstracts often outline the key findings in papers. Biomedical literature is thus a rich source of information about cellular function that can be used to inform predictions of function on new genes. A difficulty in working with literature is the unstructured nature of the documents. There have been limited efforts to codify the genes and functions referenced in subsets of the biomedical literature that focus on specific organisms. For instance, articles studying yeast have been manually codified using gene ontology by a team of experts [21]. Automated analysis of biomedical literature still remains a key open challenge.

Raychaudhuri et al. address this problem by combining biomedical literature with gene expression data to determine which genes with correlated expression have the same cellular function. Their neighbor divergence per gene (NDPG) method [22] use coherence of PubMed articles to score how likely genes with correlated expression have the same cellular function. The intuition behind NDPG is that if a group of genes have the same cellular function, then PubMed articles that refer to the group should have semantic neighbors which also refer to the group. These methods demonstrate an approach of dealing with noisy and sparse data through integrating it with other diverse data sources.

Physiological Function

The physiological function of a gene is the effect that a gene has on the entire organism, as is perceived through its phenotype. The phenotype of an organism is its physical appearance or behavior, such as eye color and side effects to drug therapy. Genotype is the genetic makeup of an organism. The phenotype of an organism is affected by its genotype and its environment. A key challenge is to correlate variations in the organism's genotype to variations in its phenotype.

Using Biomedical Literature to Identify Gene and Drug Relationships

Pharmacogenomics studies focus on correlating an organism's genotype with drug efficacy. By understanding how genes and drugs relate, drug therapy can be tailored to an individual's unique genetic makeup. The Pharmacogenomics and Pharmacogenetics Knowledge Base (PharmGKB) [23] provides a central repository for data involving variations in genes and their effect on drug response. Much of the data stored in PharmGKB are manually curated from biomedical literature, which provides an important source of information about genes and drugs. Unfortunately, manual curation of biomedical literature is very slow and inefficient. Because research articles are not stored in a structured format, automating the extraction of gene-drug relationships is an important challenge. Chang et al. use natural language processing techniques to identify gene-drug relationships in the biomedical literature [24]. Chang et al. use co-occurrences of genes and drugs in the literature to first identify whether genes and drugs are related, and then employ a maximum entropy classifier to classify gene-drug relationships. Automated analysis of pharmacogenomics data from biomedical literature is still an active area of research.

Challenges

Understanding gene function at the molecular, cellular, and physiological level each poses numerous challenges. Below we present some key examples of research areas where advances in computational methodology may potentially impact the field. The list below is not exhaustive but represents areas with current high activity and where there is still great potential for computational improvements.

Molecular Function Challenges

Current open questions at the molecular scale include:

▲ 1) *Annotating genomes.* Genomes of many organisms have been sequenced, but identification of the exact sequence and structure of every gene in the genome and their regulation remains a challenge. In the human genome, even the estimates for the exact number of genes range widely from 30,000 to 120,000. In addition, in eukaryotes the same gene can encode for more than one protein through the process called alternative splicing. Recent analyses indicate that 40–60% of human genes are predicted to have alternative splice forms. Identification of gene locations, their regulation, their products, and their function is an open area of research that can benefit from machine learning methods and other engineering principles.

▲ 2) *Identifying functional regions in proteins.* Identifying the exact area of a protein that is responsible for biological function is critical for modeling the fundamental mechanisms of protein activity as well as developing pharmaceutical drugs. Accurate analysis of protein sequence and structure to define specific regions or residues of a protein that can be linked to a specific function is an area where machine learning and signal processing methods can be of great benefit.

▲ 3) *Identifying protein interactions and their mechanisms.* Even when a functional site of a protein is known, understanding what molecules the protein interacts with, as well as the mechanism of the interaction, remains an open question. The dynamics of this interaction is also critically important. For example, in pharmaceutical development it is important to know the strength and transiency of interaction between a drug and protein.

Cellular Function Challenges

Current open questions at the cellular scale include:

▲ 1) *Identifying biological pathways.* To understand how cells operate, it is critical to understand function and regulation of specific pathways in the cell. A key challenge here is to identify which proteins participate in specific pathways (e.g., proteins involved in fatty acid metabolism) and how members of a pathway interact with one another. Furthermore, it is necessary to understand and accurately model pathway components and regulation as well as their fault-tolerant properties. One application of these models would be development of drugs that target nonrobust or nonredundant pathways in pathogens, which is an effective way to fight certain diseases.

▲ 2) *Quantifying the dynamics of biological networks.* Biological pathways interact and interconnect in the cell to form networks. Modeling the structure and dynamics of biological networks is critical for understanding how a cell functions and how it is affected by environmental conditions. A key outcome of this challenge is creating an accurate functional model of the cell.

▲ 3) *Engineering biological networks.* An emerging research area is design of artificial genetics circuits inside living cells. These circuits can be used to study cellular function and regulation in simplified and controlled systems, as well as to create cells with specific properties, such as bacteria engineered for detecting chemicals and for bioremediation [25]. Circuit diagrams can be designed using computational and modeling tools, and then constructed inside cells using experimental techniques.

Physiological Function Challenges

Research on gene function at the organism level has only begun recently and the area poses many open questions. At a broad level, challenges at the organism level include:

▲ 1) *Correlating genotype to organism level effects.* The variety of phenotypes observed among one species is formed through the interaction of each individual's genotype with its environment. These variations in genotype can arise from single nucleotide mutations or from modifications and deletions of whole genes. Computational and experimental researchers need to model effects of changes in the genomic composition of an organism on the organism's fitness and survival. Such models will improve our understanding of fundamental biology and may lead to development of novel treatments for disease that are targeted to each individual's genotype.

▲ 2) *Extracting knowledge from biomedical literature.* Most of human knowledge about biology is summarized in the biomedical literature in the form of volumes of articles reporting biological experiments, clinical studies and their results. Unfortunately, information in these articles is not stored in a structured format and thus is mostly inaccessible to computational analyses. Accurate natural language processing techniques are needed to mine biomedical literature for information and present the resulting data in a format that computers can effectively utilize. Furthermore, development of new controlled vocabularies and ontologies of concepts and relationships is necessary to represent this knowledge in a computationally accessible form.

The biological challenges described above share a common set of computational challenges. In the post-genomic era, large quantities of biological data are being produced, and these data need to be stored, processed, and mined for relevant biological information. These large-scale data are heterogeneous, and thus novel methods are necessary that can deal with integration as well as provide fast and accurate analysis of these diverse data sources. Biological data are often noisy and many problems posed in bioinformatics do not have true gold standard solutions, therefore machine learning methods for biological data should be robust to noise, number, and quality of the examples. These computational challenges are not unique to the biological domain. Expertise from researchers in the signal processing, computer science, and other engineering disciplines can provide critical insight and innovation on applying new methodology to the biological field.

Conclusion

The exponential growth of publicly available data has transformed biology into an information rich science that provides many new and interesting applications for the machine learning community. Biological data is especially challenging to analyze because it is inherently noisy and biased. The research presented in this review illustrates how both effective representation of these data and proper feature selection are critical to the success of a particular computational functional genomics approach. Ultimately, the success of a computational method in genomics is defined by whether it furthers scientists' understanding of the biological system in question.

This review of research in computational functional genomics highlights the types of computational challenges that arise in functional genomics and current methodologies that are employed. Further information about computational biology is available from the International Society for Computa-tional Biology (http://www.iscb. org/) and conferences such as the Intelligent Systems in Molecular Biology conference, Pacific Symposium on Biocomputing, the Research in Computational Molecular Biology conference and the IEEE Computer Society Bioinformatics conference. The signal processing community has much experience in developing robust and rapid techniques for analyzing large data sets that are noisy and biased. We hope that this review provides an introduction to the field of computational functional genomics for members of the IEEE community.

Mike P. Liang, *Alain Laederach*, and *Russ B. Altman* are with the Department of Genetics, Stanford University Medical Center, California. *Olga G. Troyanskaya* is with the Department of Computer Science, Princeton University, New Jersey. *Douglas L. Brutlag* is with the Department of Biochemistry, Stanford University Medical Center, Stanford, California.

References

[1] B. Boeckmann, A. Bairoch, R. Apweiler, *et al.*, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, pp. 365–370, 2003.

[2] R. Apweiler, A. Bairoch, C.H. Wu, *et al.*, "UniProt: The universal protein knowledge base," *Nucleic Acids Res.*, vol. 32 Database issue, pp. D115–9, 2004.

[3] C.H. Wu, L.S. Yeh, H. Huang, *et al.*, "The protein information resource," *Nucleic Acids Res.*, vol. 31, pp. 345–347, 2003.

[4] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.

[5] S.F. Altschul, W. Gish, W. Miller, et al., "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[6] O. Gotoh, "Multiple sequence alignment: algorithms and applications," *Adv. Biophys.*, vol. 36, pp. 159–206, 1999.

[7] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.

[8] P. Bucher and A. Bairoch, "A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation," in *Proc. Int. Conf. Intell Syst. Mol. Biol.*, 1994, vol. 2, pp. 53–61.

[9] E.L. Sonnhammer, S.R. Eddy, and R. Durbin, "Pfam: A comprehensive database of protein domain families based on seed alignments," *Proteins*, vol. 28, pp. 405–420, 1997.

[10] C.G. Nevill-Manning, T.D. Wu, and D.L. Brutlag, "Highly specific protein sequence motifs for genome analysis," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 5865–5871, 1998.

[11] T.D. Wu, C.G. Nevill-Manning, and D.L. Brutlag, "Fast probabilistic analysis of sequence function using scoring matrices," *Bioinformatics*, vol. 16, pp. 233–244, 2000.

[12] S.E. Brenner, "A tour of structural genomics," *Nat Rev Genet*, vol. 2, pp. 801–809, 2001.

[13] H.M. Berman, J. Westbrook, Z. Feng, *et al.*, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–242, 2000.

[14] A.C. Wallace, N. Borkakoti, and J.M. Thornton, "TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites," *Protein Sci.*, vol. 6, pp. 2308–2323, 1997.

[15] J.S. Fetrow, A. Godzik, and J. Skolnick, "Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity," *J. Mol. Biol.*, vol. 282, pp. 703–711, 1998.

[16] L. Wei and R.B. Altman, "Recognizing protein binding sites using statistical descriptions of their 3D environments," in *Proc. Pac. Symp. Biocomput*, 1998, pp. 497–508.

[17] M.P. Liang, D.L. Brutlag, and R.B. Altman, "Automated construction of structural motifs for predicting functional sites on protein structures," in *Proc. Pac. Symp. Biocomput*, 2003, pp. 204–215.

[18] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, pp. 249–255, 2003.

[19] E. Segal, H. Wang, and D. Koller, "Discovering molecular pathways from protein interaction and gene expression data," *Bioinformatics*, vol. 19 Suppl 1, pp. I264–I272, 2003.

[20] O.G. Troyanskaya, K. Dolinski, A.B. Owen, et al., "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proc Natl. Acad. Sci. USA*, vol. 100, pp. 8348–8353, 2003

[21] L. Issel-Tarver, K.R. Christie, K. Dolinski, et al., "Saccharomyces Genome Database," *Methods Enzymol*, vol. 350, pp. 329–346, 2002.

[22] S. Raychaudhuri and R.B. Altman, "A literature-based method for assessing the functional coherence of a gene group," *Bioinformatics*, vol. 19, pp. 396–401, 2003.

[23] M. Hewett, D.E. Oliver, D.L. Rubin, *et al.*, "PharmGKB: The pharmacogenetics knowledge base," *Nucleic Acids Res.*, vol. 30, pp. 163–165, 2002.

[24] J.T. Chang, "Using machine learning to extract drug and gene relationships from text," Stanford Univ., 2003.

[25] M.E. Wall, W.S. Hlavacek, and M.A. Savageau, "Design of gene circuits: Lessons from bacteria," *Nat. Rev. Genet.*, vol. 5, pp. 34–42, 2004.