Integrative Data Mining: The New Direction in Bioinformatics

Machine Learning for Analyzing Genome-Wide Expression Profiles and Proteomics Data Sets

Paul Bertone[1] and Mark Gerstein[2]
[1]Dept. of Molecular, Cellular,
and Developmental Biology
[2]Dept. of Molecular Biophysics and Biochemistry and
Dept. of Computer Science,
Yale University

Biological research is becoming increasingly database driven, motivated, in part, by the advent of large-scale functional genomics and proteomics experiments such as those comprehensively measuring gene expression. These provide a wealth of information on each of the thousands of proteins encoded by a genome. Consequently, a challenge in bioinformatics is integrating databases to connect this disparate information as well as performing large-scale studies to collectively analyze many different data sets. This approach represents a paradigm shift away from traditional single-gene biology, and it often involves statistical analyses focusing on the occurrence of particular features (e.g., folds, functions, interactions, pseudogenes, or localization) in a large population of proteins. Moreover, the explicit application of machine learning techniques can be used to discover trends and patterns in the underlying data. In this article, we give several examples of these techniques in a genomic context: clustering methods to organize microarray expression data, support vector machines to predict protein function, Bayesian networks to predict subcellular localization, and decision trees to optimize target selection for high-throughput proteomics.

Biological Research Is Database Oriented

Databases have defined the information structure of molecular biology for over a decade, archiving thousands of protein and nucleotide sequences and three-dimensional (3-D) structures. As large-scale genomics and proteomics move to the forefront of biological research, the role of databases has become more significant than ever. The current landscape of biological databases includes large public archives, such as GenBank, DDBJ, and EMBL for nucleic acid sequences [1]; PIR and SWISS-PROT for protein sequences [2]; and the Protein Data Bank for 3-D protein structure coordinate sets [3]. Another source of sequence data is dbEST [4], a division of GenBank storing expressed sequence tags (ESTs) from cell lines, which provide information about gene expression in various tissues. Databases such as these have been steadily accumulating gene sequences and protein structures for more than a decade, which are submitted on a per-instance basis from disparate laboratories in the biological sciences community.

In addition to these general repositories of biomolecular data, specialized systems have been developed that extend its interpretation by providing a context for individual sequences and structures. The SCOP, CATH, and FSSP [5] databases classify proteins based on structural similarity, Pfam and ProtoMap [6] identify families of proteins based on sequence homology, while PartsList and GeneCensus [7] give dynamic reports on the occurrence of protein families in various genomes. Databases have also been developed to provide comprehensive access to sequence, expression, and functional data for all the known genes of specific model organisms [8].

Integration of Databases and Large-Scale Surveys

In addition to sequence and structure databases, many diverse experimental data sets have been compiled that focus on various aspects of protein function. However, these need to be related with other data in order to be placed in a useful context, as in-

Integrative database analysis and data mining are essential aspects of modern biology, enabling the synthesis of large-scale portraits of genome function.

dividual bits of information are relatively meaningless outside of this wider scope. For example, an experiment measuring high mRNA copy number for the human gene J00068 carries little intrinsic value until one associates this fact with the gene's functional classification, learning that it codes for the protein actin, an abundant component of skeletal muscle.

The scale of genomics research precludes the traditional single-molecule approach in biology, where all the experimental knowl-edge about a given protein could be obtained through reading the literature. Integrative database analysis and data mining are essential aspects of modern biology, enabling the synthesis of large-scale portraits of genome function.

In practice, data mining often takes the form of statistical surveys, in the sense of demographic censuses, of different populations of genes and proteins. The goal in this analysis is to identify certain outstanding features possessed by a given population. For instance, large-scale surveys have been used to characterize the features of pseudogenes (inactive protein-coding regions) in the worm genome [9], correlate the levels of protein expression with subcellular localization [10], examine the relationship between protein structure with function [11-12], and measure the composition of protein folds in complete genomes [13-14]. Bioinformatics researchers conducting integrative database surveys face the challenge of merging genomics data from many different sources into a common framework. Therefore, the development of standards allowing federated databases to interoperate is essential.

Functional Genomics and Microarray Technology

With the introduction of complete genome sequences, much comprehensive functional analysis has been done on the proteins encoded by an organism's genome. The sequence determination phase of genomics offers only a fraction of the analytical possibilities. Many functional experiments can be performed using that data, and countless opportunities exist to relate those experimental results with other properties via integrative database analysis.

Among the experimental techniques available for genome-wide analysis are gene disruption [15], two-hybrid studies [16], large-scale proteomics [17], silicone elastomer protein chips [18], serial analysis of gene expression (SAGE) [19], and various DNA microarray technologies. Of these, microarrays have become particularly popular due to the highly parallel nature of the experiments and the inherent homogeneity of the data captured. Using array technologies, immobilized oligonucleotide fragments having known sequences are exposed to fluorescence-labeled DNA probes, and the signals corresponding to hybridized fragments are scanned and quantified [Fig. 1(a)]. A number of these systems have been devised, allowing the simultaneous interrogation of thousands of genes in a single experiment. Due to the recent sequencing of complete genomes, laboratories are using these array technologies to generate expression data on a scale that most researchers would have considered nearly impossible just a few years ago.

To construct the array shown in Fig. 1(a), DNA fragments of known sequence are immobilized to each array element. Messenger RNA transcripts from cell extracts are reverse-transcribed and labeled
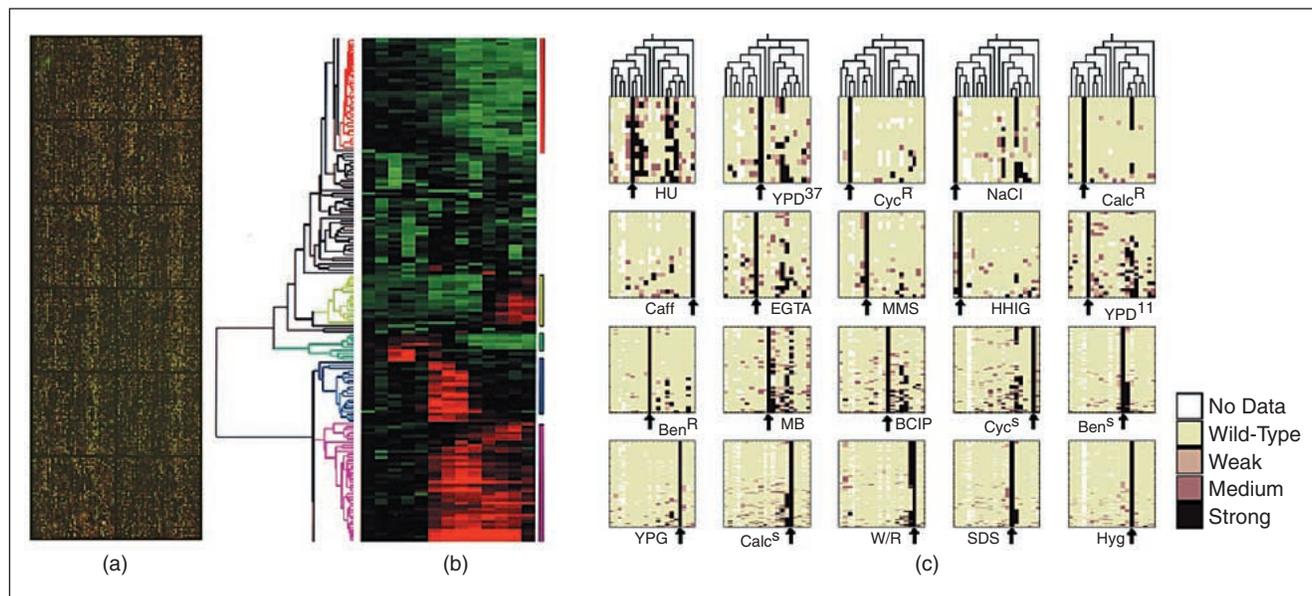
1. (a) Scanned image of a cDNA microarray (see text for details). (b) Hierarchical clustering techniques applied to microarray expression data, adapted from [24] (see text for details). (c) Double-clustering of disruption phenotype data using *k*-means and hierarchical methods [15] (see text for details).

with the fluorophore cyanine-5, serving as hybridization probes. Total genomic DNA is labeled with cyanine-3 and hybridized to all the array elements as a negative control. The two signals are generated separately with a laser using different excitation wavelengths (633 and 543 nm, respectively), and the emissions are scanned and quantified via densitometry. Changes in gene expression are measured as the normalized ratios between mRNA level and background signal for each array element. The image is false-colored after scanning; here, genes that are up-regulated are shown in red, while genes that are down-regulated appear green. In this way, differential expression of thousands of genes can be quantified simultaneously.

Microarray technologies include cDNA arrays [20] and high-density oligonucleotide systems, such as GeneChips [21] and intergenic DNA arrays [22]. Each of these methods is designed to capture a different type of information. GeneChips and cDNA arrays enable genome-wide expression monitoring by measuring mRNA copy number under various cellular conditions, compiling an expression profile for many genes at incremental time points. cDNA microarrays measure gene expression as the difference in signal strength between reverse-transcribed mRNA levels from cells under various conditions. This difference is given relative to a reference value, which yields a ratio for every target gene under investigation. Conversely, GeneChips measure the absolute expression levels of mRNA transcripts, in number of copies per cell.

Intergenic DNA arrays are an emerging technology, involving the construction of high-density oligonucleotide chips that are probed with DNA fragments captured via chromatin immunoprecipitation (sometimes referred to as ChIP chips). DNA-binding proteins, complexed with their cognate recognition sequences, are sequestered from cell extracts with antibody-bound beads. The DNA is then isolated and used to probe the array, hybridizing to elements containing complementary nucleotide sequences. In experiments where transcription factor-bound DNA is used to interrogate a microarray containing intergenic sequences, sites of transcription factor binding can be identified across an entire genome [22]. Evidence of putative gene expression accumulates when both cDNA

and ChIP-derived probes are hybridized to identical arrays and comparisons are made between differential mRNA levels and upstream sites of transcription factor binding. Combined with existing expression analysis methods, this location data can be used to assemble a more comprehensive picture of large-scale genetic function.

Machine Learning Approaches to Genomic Data Analysis
Unsupervised Learning and Clustering

A general problem in data analysis is how to structure information into meaningful taxonomies or categories. This issue is of great importance when trying to infer relationships in diverse biological data sets. Statistical methods for finding trends and patterns in experimental results have played a large role in their interpretation. Principal component analysis (PCA) can be an effective method of identifying the most discriminating features in a data set. This technique usually involves finding two or three linear combinations of the original features that best summarize the types of variation in the data. If much of the variation is captured by these two or three most significant principal components, class membership of many data points can be observed.

Several approaches to this problem employ unsupervised learning to find these categories; that is, no a priori information is required, and generally, no feedback is given to the model to adjust its performance. Unsupervised learning enables pattern discovery by organizing data into clusters, using recursive partitioning methods. The k-means algorithm is a popular instance-based method of cluster analysis. The algorithm partitions data into a predetermined number of categories as instances are examined, according to a distance measure (e.g., Euclidean). Category centroids are fixed at random positions when the model is initialized, which can affect the clustering outcome.

The self-organizing feature map (SOM) [23] consists of a neural network whose nodes move in relation to category membership. As with k-means, a distance measure is computed to determine the closest category centroid. Unlike k-means, this category is represented by a node with an associated weight vector. The weight vector of the matching node, along with those of neighboring nodes, are updated to more

closely match the input vector. As data points are clustered and category centroids are updated, the positions of neighboring nodes move in relation to them. The number of network nodes that constitute this neighborhood typically decrease over time.

While both algorithms require the number of clusters to be empirically chosen, the SOM algorithm overcomes some limitations of k-means by imposing global relationships between clusters, thereby improving interpretability. Like PCA, the SOM is capable of reducing high-dimensional data into a 1- or 2- dimensional representation. The algorithm produces a topology-preserving map, conserving the relationships among data points. Thus, although either method may be used to effectively partition the input space into clusters of similar data points, the SOM can also indicate relationships between clusters.

Supervised Learning and Classification

Analysis of large data sets that contain diverse information often involves the explicit application of supervised learning. This generally involves dividing the data set features into two categories: *predictors*, or features in a data set that are relevant for learning, and the *response variable*, or property to be classified. Given the heterogeneous information presented by integrated bioinformatics databases, a number of potential correlations between predictors and response variables may be discovered, depending on the subset of features used and the classifications sought after.

Machine learning can be applied to a wide variety of biological information,

Interpretability of results is a key factor in selecting a particular machine learning method.

both to partition the data into categories and to classify previously unseen examples. Supervised learning is conducted in two phases: training and testing the classifier model. Using this strategy, the data set is divided into two mutually exclusive sets. The first set is used to train the model, where correct classifications/responses of the input examples are known a priori. This information is used to improve the performance of the model and reduce the classification error rate, a process that incrementally adjusts an *n*-dimensional hyperplane that serves to partition the data set into categories. Afterward, unseen instances in the test set are classified according to the partitioning established during the training phase.

Classification Performance versus Ease of Interpretation

For biological research applications, interpretability of results is a key factor in selecting a particular machine learning method. By deciphering the mapping between points in feature space and learned categories, one can apply classification results in a practical context. The ability to decipher this mapping depends largely upon the internal representation used by a particular algorithm. For example, decision trees are quite useful in this respect, because rules can be extracted from the tree that discriminate amongst classes. Backpropagation neural networks, while

useful classifiers, employ a numerical weight matrix to establish nonlinear mappings that can be difficult to interpret.

Biological Data Mining Applications

Organizing Microarray Data

Clustering algorithms are being applied to microarray data sets with increasing regularity; they are often incorporated into microarray image analysis software and are therefore frequently used to visualize local and global relationships among hybridization signals captured by the array. Currently, hierarchical clustering is the most popular technique employed for microarray data analysis [24]. Hierarchical methods are based on building a distance matrix summarizing all the pairwise similarities between expression profiles and then generating cluster trees (also called dendrograms) from this matrix. Genes that appear to be co-expressed at various time points are positioned close to one another in the tree, whose branch lengths represent the degree of similarity between expression profiles.

Agglomerative methods such as *k*-means and self-organizing maps have also been used to cluster microarray data. The SOM algorithm has been applied to identify cell differentiation patterns and distinguish between cancer cell types via array expression profiling [25]. Using both hierarchical and neighborhood analysis techniques, genes exhibiting similar mRNA levels tend to cluster together, revealing patterns in gene expression data [Figure 1(b)]. In this software display, microarray intensity values are depicted as colored blocks, where rows represent individual genes, and columns correspond to incremental time points when successive experiments were performed. Thus, each row constitutes an expression profile for a given gene. Red blocks indicate high mRNA levels, green blocks indicate low levels. The data was organized via hierarchical clustering to group genes with related expression profiles.

While useful visualization tools, the analytical utility of either type of clustering approach is limited in terms of interpreting microarray results. These techniques serve to group data points based on changes in mRNA levels under various cellular conditions. Although cluster membership of related data points can indicate coexpression or co-repression of the genes they represent, the existence of common regulatory pathways based on partitioning expression profiles remains speculative in

the absence of external information. Additionally, there is no reason to believe that co-expressed genes are evolutionarily related as hierarchical ordering might imply.

Partitioning Phenotype Data with Multiple Clustering Techniques

Where appropriate, clustering methods can be combined to establish a multitiered partitioning of a data set. An example of this approach is the tandem use of *k*-means clustering with hierarchical clustering as applied to genome-wide phenotypic analysis [15]. Yeast cells containing transposon insertions were sorted with the *k*-means algorithm based on growth conditions, to produce clusters of genes with similar phenotypes. The same data was then ordered hierarchically, producing trees of experimental assays, and clustering those which yield transformants with shared phenotypes [Fig. 1(c)]. Using this method, assays can be identified that are capable of selecting functionally related genes.

In Fig. 1(c), both examples of hierarchical clustering use the Pearson coefficient to measure similarity between gene expression profiles. Given two normalized expression ratio profiles, the Pearson correlation coefficient is given by the dot product

$$R_{ij} = \frac{1}{N-1} X_i \cdot X_j$$

where N is the number of elements represented by the profiles X_i and X_j. The measurement expression ratio profile x can be used to compute a *Z*-score from the normalized profile X with the following equation

$$X(k) = \frac{x(k) - x_{\text{avg}}}{\sigma_x}$$

where $X(k)$ and $x(k)$ are the kth components of the profiles X and x, respectively, x_{avg} is the average, and σ_x is the standard deviation of values in x. The correlation coefficient matrix R can then be computed for a group of genes, where each matrix element R_{ij} corresponds to the Pearson correlation coefficient between genes i and j. The average correlation coefficient R_{avg} is found by averaging the elements of R, including the diagonal. This statistic indicates the overall similarity of the expression profiles in a group of genes.

Classifying Gene Function with Support Vector Machines

An example of a supervised learning method applied to functional genomics

data is the use of support vector machines (SVMs) to classify yeast ORFs into several functional categories by correlating mRNA expression with prior knowledge of gene function [26]. The SVM maps an n-dimensional input space onto a higher-dimensional feature space, simultaneously transforming a nonlinear class boundary into a simple hyperplane. The new feature space is embodied as a set of nonlinear combinations of the original features. If we consider a dataset whose points are linearly separable in the new feature space, their corresponding convex hulls (the tightest enclosing convex polygons) do not overlap. The *maximum margin hyperplane* is defined as the hyperplane that is maximally distant from both convex hulls; this will bisect the shortest line connecting them. The data points closest to the maximum margin hyperplane are called *support vectors* and uniquely define the maximum margin hyperplane (the minimum number of support vectors from each class is one). Finding the support vectors and the parameters that determine the maximum margin hyperplane can be accomplished with a standard constrained quadratic optimization scheme.

To apply the SVM for gene classification, a set of examples was assembled containing genes of known function, along with their corresponding microarray expression profiles. The SVM was

then used to predict the functions of uncharacterized yeast ORFs based on the expression-to-function mapping established during training. Supervised learning techniques appear to be ideal for this type of functional classification of microarray targets, where sets of positive and negative examples can be compiled from genomic sequence annotations.

Predicting Subcellular Localization with Bayes' Rules

Bioinformatics data mining often involves the aggregation of proteomic data from multiple sources, followed by the application of machine learning techniques to identify or predict various macromolecular properties from these features. One such application of this strategy was the assignment of thousands of proteins in the yeast genome to the subcellular compartments to which they are most likely transported [27]. In this study, gene expression levels from a number of SAGE, GeneChip, and cDNA microarray experiments were accrued and cross referenced with protein localization information from public databases. Each protein was first assigned a default probability of being localized to one of several compartments. These were then updated using a Bayesian system that incorporates a number of external features to arrive at probabilities for the localization of each

protein [Fig. 2(a)]. Training the model on a set of examples for which this information is known made possible the prediction of the localization of unknown proteins. This analysis also enabled the elucidation of the relationships between expression and compartmental localization, with cytosolic proteins being highly expressed, and nuclear and membrane-related proteins exhibiting lower expression levels [Fig. 2(b)].

In Fig. 2(a), the charts depict the state vector for a protein m, representing the distribution of probabilities of protein m localizing to various subcellular compartments. Bar graphs illustrate the vectors for each feature considered. The prior state for a given protein is sequentially updated with Bayes' rule using feature vectors that represent different attributes of the protein (e.g., NLS = nuclear localization signal sequence, GLY = glycosylation site); e.g.,

$$Pm(C|\text{feature}) = Pm(C)$$
$$\times P(\text{feature}|C) / Z$$

where Z is a normalization factor equal to the product of the fraction of proteins localized in each compartment (C) having a particular feature, and the prior probability of protein m being transported to that compartment, summed over all compartments:

2. (a) Bayesian system for predicting subcellular localization [27]. Charts depict the state vector for a protein m, representing the distribution of probabilities of protein m localizing to various subcellular compartments. Bar graphs illustrate the vectors for each feature considered. (b) Total predicted compartmental populations for the yeast genome. Estimates are calculated according to the distribution of a population vector, obtained by summing the probability state vectors of all the proteins considered.

$$Z = \sum_{C} P\big(\text{feature}|C\big) \times Pm\big(C\big).$$

The predicted compartment for each protein is indicated by the probability distribution of the final state vector. In Fig. 2(b), estimates are calculated according to the distribution of a population vector, obtained by summing the probability state vectors of all the proteins considered.

Using Decision Trees to Optimize High-Throughput Proteomics

The future of biological databases will likely include systems that are specifically designed to capture data in a format that facilitates retrospective analysis. This requires careful standardization of experimental parameters, thereby rendering the data amenable to subsequent computation. As discussed above, computing techniques are readily applied to microarray results, partially due to the inherent homogeneity of these data sets. Similarly, a goal in the fusion of databases with data mining applications is to standardize highly divergent data sets as much as possible, making them accessible to machine learning algorithms.

Recently, a combined approach was developed to facilitate distributed collaboration among many laboratories using a shared database system and to subsequently predict macromolecular properties from structural proteomics data via decision tree analysis [28]. Experimental parameters for protein cloning, purification, biophysical characterization, NMR, and X-ray crystallization were recorded in a database. These values were standardized across many laboratories and institutions, providing a common data format for retrospective analysis. Using the database to generate training sets for supervised learning, decision trees [29] were used to classify proteins as either soluble or insoluble, based on features of their amino acid sequences. Useful rules relating these features with protein solubility were then determined by tracing the paths through the decision trees (Fig. 3). Protein solubility strongly influences whether a given protein is a feasible target for structure determination, so the ability to predict this property can be a valuable asset in the optimization of high-throughput projects.

In Fig. 3, experimental data for protein cloning, purification, biophysical characterization, and structure determination are uploaded from many distributed workstations to a central database server via the Internet. These are stored as standardized values and used as features for data mining analyses. A number of response variables can be predicted in this manner. Decision trees were built with the C4.5 algorithm [29] to sort proteins based on solubility, using training sets compiled from the database. The model was trained using features derived from the protein sequences, such as the composition of various amino acid groups (e.g., aliphatic = C(AIGLV), aromatic = C(FWY)), secondary structure features, presence of hydrophobic regions, entropic measures of sequence complexity, etc. Ellipses represent the decision tree nodes, filled to indicate the number of soluble versus insoluble proteins at each vertex. These values appear to the right and left of each node, respectively. Rules that discriminate between classes were then extracted from decision trees by examining the paths from root to leaf nodes that lead to correct classifications. In this example, it was found that proteins having a combined composition of acidic residues (C(DE)) greater than 18% were likely to be soluble, while insoluble proteins generally had less than 18% acidic residues, possessed a stretch of amino acids with average hydrophobicity < -0.78 kcal/mol (Hphobe), and contained fewer than 16% acidic amino acids and their amides (C(DENQ)).

Conclusion

As with biological sciences in general, the emphasis of computational biology has changed in recent years from the characterization of individual molecules to the analysis of genome-wide expression profiles and proteomics data sets. In contrast to traditional single-gene experimenta-
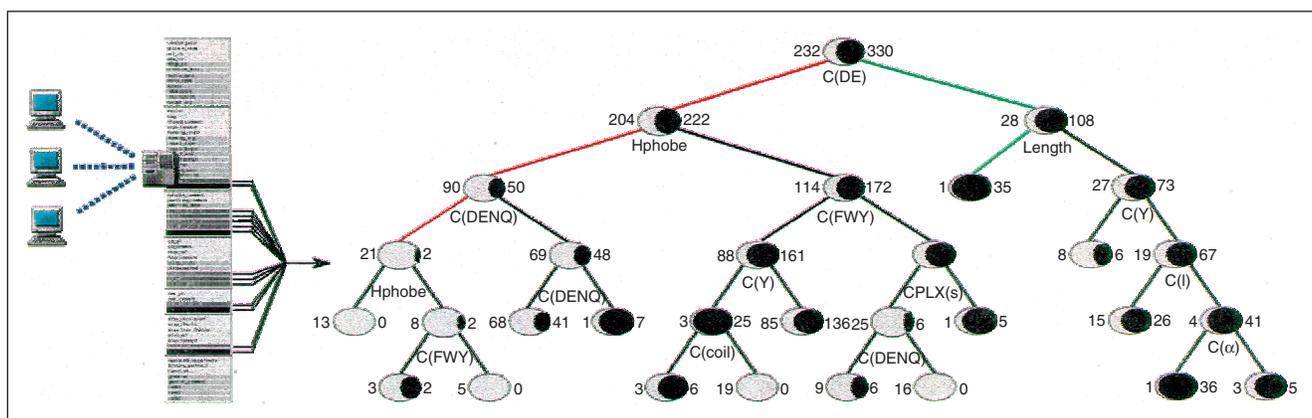
3. Integrated database and machine learning system for optimizing high-throughput proteomics, adapted from [28] (see text for details).

tion, comprehensive functional data sets are being produced, the scale of which has increased with the tandem sequencing of entire genomes and the advent of microarray technologies. An essential part of this work is the integration of a wide variety of experimental data to assemble a larger picture of biological function and frame pertinent information within a meaningful context. These composite data sets are conducive to extensive computational analysis and present new opportunities for data mining.

Machine learning is a useful tool for interpreting genomic information. Both supervised and unsupervised approaches can often be used to analyze the same kinds of data, depending on the desired result and the range of features available. While supervised learning can be used to classify examples according to a given response variable, unsupervised learning can aid in revealing previously unknown relationships without a priori information. However, meaningful clustering relationships are often difficult to discern. Although supervised models require an explicit training and testing regime, they can be used to predict the class of unknown examples based on previous learning by approximating a target function or discovering classification rules. This capability can allow investigators to interpret data classifications and potentially elucidate functional properties.

Bioinformatics calls for a mosaic of computing techniques to facilitate the effective organization and interpretation of data generated by functional genomics projects. Large-scale experiments, such as those performed with microarrays, yield large homogenous data sets that are well suited for computational analysis. Consequently, new types of databases have been created to handle this information [30]. These information systems are designed to provide management and web-based retrieval of microarray data. Currently under debate is whether data sets should be normalized as they are entered into a particular system, and to what degree the details of experimental conditions should be recorded. While the microarray intensity values are well suited to relational table storage, experimental parameters vary widely across individual projects and currently defy uniform standardization.

The data storage and analysis requirements of genomics research will likely promote the development of hybrid approaches that merge database design with computational methods. To facilitate this, it will be necessary to establish consistent formats for database interoperation, which will add greatly to the utility of experimental results in terms of potential data mining applications.

Paul Bertone received the B.M. degree in 1993 and the M.S. in computer science in 1996 from the University of Massachusetts and the M.A. in cognitive and neural systems in 1998 from Boston University. He is currently a Ph.D. candidate in the Department of Molecular, Cellular, and Developmental Biology at Yale University. His research interests include intergenic and maximal-coverage DNA microarray design and the application of machine learning strategies to biological data analysis, particularly in relation to human genomics.

Mark Gerstein is an assistant professor in the Department of Molecular Biophysics and Biochemistry at Yale University. He is also a joint appointee in the Department of Computer Science. He received is A.B. in physics from Harvard College in 1989 and his Ph.D. from Cambridge in biophysics in 1993. He did postdoctoral work at Stanford and took up his post at Yale in early 1997. Since then he has received a number of young investigator awards (e.g., from the Navy and Keck foundations) and has published appreciably in biological science journals (80 in total). His research is focused on bioinformatics and he is particularly interested in large-scale integrative surveys, biological database design, macromolecular genomtry, molecular simulation, genome annotation, gene expression analysis, and Bayesian systems for data mining.

Address for Correspondence: Mark Gerstein, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520. Tel: +1 203 432 6105. Fax: +1 360 838 7861. E-mail: Mark.Gerstein@yale.edu.

References

[1] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, and D.L. Wheeler, *Genbank, Nucleic Acids Res.*, vol. 28, pp. 15-18, 2000; Y. Tateno, S. Miyazaki, M. Ota, H. Sugawara, and T. Gojobori, "DNA bank of Japan (DDBJ) in collaboration with mass sequencing teams," *Nucleic Acids Res.*, vol. 28, pp. 24-26, 2000; W. Baker, A. van der Broek, E. Camon, P. Hingamp, P. Sterk, G. Stoesser, and M.A. Tuli, "The EMBL nucleotide sequence database," *Nucleic Acids Res.*, vol. 28, pp. 19-23, 2000.

[2] W.C. Barker, J.S. Garavelli, H. Huang, P.B. McGarvey, B. Orcutt, G.Y. Srinivasarao, C. Xiao, L.S. Yeh, R.S. Ledley, J.F. Janda, F. Pfeiffer, H.W. Mewes, A. Tsugita, and C. Wu, "The Protein Information Resource (PIR)," *Nucleic Acids Res.,* vol. 28, pp. 41-44, 2000; A. Bairoch, and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Res.*, vol. 28, pp. 45-48, 2000.

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235-242, 2000.

[4] M.S. Boguski, T.M. Lowe, and C.M. Tolstoshev, "dbEST–database for 'expressed sequence tags,'" *Nat. Genet.*, vol. 4, pp. 332-333, 1993.

[5] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, pp. 536-540, 1995; C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton, "CATH- A Hierarchic Classification of Protein Domain Structures," *Structure*, vol. 8, pp. 1093-1108, 1997; L. Holm and C. Sander, "Touring protein fold space with Dali/FSSP," *Nucleic Acids Res.*, vol. 26, pp. 316-319, 1998.

[6] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 28, pp. 263-266, 2000; G. Yona, N. Linial, and M. Linial, "ProtoMap: Automatic classification of protein sequences and hierarchy of protein families," *Nucleic Acids Res.*, vol. 28, pp. 49-55, 2000.

[7] J. Qian, B. Stenger, C.A. Wilson, J. Lin, R. Jansen, S.A. Teichmann, J. Park, W.G. Krebs, H. Yu, V. Alexandrov, N. Echols, and M. Gerstein, "PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information," *Nucleic Acids Res.*, vol. 29, pp. 1750-1764, 2001.

[8] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein, "SGD: Saccharomyces Genome Database," *Nucleic Acids Res.*, vol. 26, pp. 73-80, 1998; H.W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil, "MIPS: A database for genomes and protein sequences," *Nucleic Acids Res.*, vol. 28, pp. 37-40, 2000; L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth, "WormBase: Network access to the genome and biology of Caenorhabditis elegans," *Nucleic Acids Res.*, vol. 29, pp. 82-86, 2001; W.M. Gelbart, M. Crosby, B. Matthews, W.P. Rindone, J. Chillemi, S. Russo Twombly, D. Emmert, M. Ashburner, R.A. Drys-

dale, E. Whitfield, G.H. Millburn, A. de Grey, T. Kaufman, K. Matthews, D. Gilbert, V. Strelets, and C. Tolstoshev, "FlyBase: a Drosophila database. The FlyBase consortium," *Nucleic Acids Res.*, vol. 25, pp. 63-66, 1997.

[9] P. Harrison, N. Echols, and M.B. Gerstein, "Digging for dead genes: An analysis of the characteristics of the pseudogene population in the C. elegans genome," *Nucleic Acids Res.*, vol. 29, pp. 818-830, 2001.

[10] A. Drawid, R. Jansen, and M. Gerstein, "Genome-wide analysis relating expression level with protein subcellular localization," *Trends Genet.*, vol. 16, pp. 426-430, 2000.

[11] H. Hegyi and M. Gerstein, "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome," *J. Mol. Biol.*, vol. 288, pp. 147-164, 1999.

[12] J.M. Thornton, C.A. Orengo, A.E. Todd, and F.M. Pearl, "Protein folds, functions, and evolution," *J. Mol. Biol.*, vol. 293, pp. 333-342, 1999.

[13] M. Gerstein, "Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census," *Proteins*, vol. 33, pp. 518-534, 1998.

[14] Y.I. Wolf, S.E. Brenner, P.A. Bash, and E.V. Koonin, "Distribution of protein folds in the three superkingdoms of life," *Genome Res.*, vol. 9, pp. 17-26, 1999.

[15] P. Ross-Macdonald, P.S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K.H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heidtman, F.K. Nelson, H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G.S. Roeder, and M. Snyder, "Large-scale analysis of the yeast genome by transposon tagging and gene disruption," *Nature*, vol. 402, pp. 413-418, 1999.

[16] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, pp. 623-627, 2000.

[17] D. Christendat, A. Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J.R. Cort, V. Booth, C.D. Mackereth, V. Saridakis, I. Ekiel, G. Kozlov, K.L. Maxwell, N. Wu, L.P. McIntosh, K. Gehring, M.A. Kennedy, A.R. Davidson, E.F. Pai, M. Gerstein, A.M. Edwards, and C.H. Arrowsmith, "Structural proteomics of an archaeon," *Nat. Struct. Biol.*, vol. 7, pp. 903-909, 2000.

[18] H. Zhu, J.F. Klemic, S. Chang, P. Bertone, A. Casamayor, K.G. Klemic, D. Smith, M. Gerstein, M.A. Reed, and M. Snyder, "Analysis of yeast protein kinases using protein chips," *Nat. Genet.*, vol. 26, pp. 283-289, 2000.

[19] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett Jr., P. Heiter, B. Vogelstein, and K.W. Kinzler, "Characterization of the yeast transcriptome," *Cell*, vol. 88, pp. 243-251, 1997.

[20] D. Shalon, S.J. Smith, and P.O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Res.*, vol. 6, pp. 639-645, 1996.

[21] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, et al., "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nat. Biotechnol.*, vol. 14, pp. 1675-1680, 1996.

[22] V.R. Iyer, C.E. Horak, C.S. Scafek, D. Botstein, M. Snyder, and P.O. Brown, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, pp. 533-538, 2001.

[23] T. Kohonen, *Self-organization and Associative Memory.* Berlin, Germany: Springer-Verlag, 1984.

[24] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expresssion patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 14863-14868, 1998.

[25] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 2907-2912, 1999; T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999; D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, and P.O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat. Genet.*, vol. 24, pp. 227-235, 2000; C.M. Perou, S.S. Jeffrey, M. Van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, J.C.F. Lee, D. Lashkari, D. Shalon, P.O. Brown, and D. Botstein, "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 9212-9217, 1999; A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[26] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat. Acad. Sci. USA*, vol. 97, pp. 262-267, 2000.

[27] A. Drawid and M. Gerstein, "A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," *J. Mol. Biol.*, vol. 301, pp. 1059-1075, 2000.

[28] P. Bertone, Y. Kluger, N. Lan, D. Zheng, D. Christendat. A. Yee, A.M. Edwards, C.H. Arrowsmith, G.T. Montelione, and M.B. Gerstein, "SPINE: An integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics," *Nucleic Acids Res.*, vol. 29, no. 13, pp. 2884-2898, 2001.

[29] J.R. Quinlan, *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kauffman, 1993.

[30] National Center for Biotechnology Information (ncbi.nlm.nih.gov/geo); O. Ermolaeva, M. Rastogi, K.M. Pruitt, G.D. Schuler, M.L. Bittner, Y. Chen, R. Simon, P. Meltzer, J.M. Trent, and M.S. Boguski, "Data management and analysis for gene expression arrays," *Nat. Genet.*, vol. 20, pp. 19-23, 1998; J. Aach, W. Rindone, and G.M. Church, "Systematic management and analysis of yeast gene expression data," *Genome Res.*, vol. 10, pp. 431-445, 2000.