Efficient Data Mining for Enabling Genome-wide Computing

Wei Wang
Department of Computer Science
University of North Carolina at Chapel Hill
weiwang@cs.unc.edu

Abstract

The volume and diversity of the data acquired in biomedical applications offer unique challenges to data mining researchers. The solutions that effectively turn data into information and knowledge will advance not only the field of data mining but also the understanding of the underlying science. This paper will present new computational challenges faced by designing methods for enabling genome-wide computing and potential approaches to tackles these challenges. A mouse reference population called Collaborative Cross is used as an example to illustrate computational difficulties in population-level genome-wide analysis.

1. Background

With the realization that a new model population was needed to understand human diseases with complex etiologies, a genetically diverse reference population of mice called Collaborative Cross (CC) was proposed. The CC is a large, novel panel of recombinant inbred (RI) lines that combines the genomes of genetically diverse founder strains to capture almost 90% of the known variation present in laboratory mice and that is designed specifically for complex trait analysis. It provides a translational tool to integrate gene functional studies into genetic networks using realistic population structures, which will be essential to understand the intricacies of biological processes such as disease susceptibility. In turn, CC becomes the focal point for cumulative and integrated data collection, giving rise to the detection of networks of functionally important relationships among diverse sets of biological and physiological phenotypes and a new view of the mammalian organism as a whole and interconnected system. It has the potential to support studies by the larger scientific community incorporating multiple genetic, environmental, and developmental variables into comprehensive statistical-supported models describing differential disease susceptibility and progression. Equally important, the CC is an ideal test bed for predictive, or more accurately, probabilistic biology, which will be essential for the deployment of personalized medicine.

2. Research Issues in Data Mining

The volume and diversity of the data collected in CC offers unique challenges, whose solutions will advance both our understanding of the underlying biology and the tools for computational analysis. For instance, it is commonplace to use clustering algorithms to search for patterns among homogeneous datasets, such as mRNA expression data. However, it is unclear how best to cluster, and thus find patterns within, multimodal datasets, such as those combining genomic, mRNA expression assays, and phenotypic measurements. Yet in their totality these measurements surely tell hidden stories about cause and effect, identify biomarkers, or predict clinical course and outcome.

2.1. Analysis of High Dimensional Data

It is widely recognized that noise and diversity are challenges when analyzing microarrays to discover those genes whose expression is modified by experimental conditions or when examining the genotypes of individuals raised in diverse environments to find the loci underlying complex traits. In both cases, exploratory, non-hypothesis driven, data analysis is a crucial first step.

Clustering has been the most popular approach of analyzing gene expression data and has proven successful in many applications, such as discovering gene pathway, gene classification, and function prediction. There is a very large body of literature on clustering in general and on applying clustering techniques to gene expression data in particular. Several representative algorithmic techniques have been developed and experimented in clustering gene expression data, which include but are not limited to hierarchical clustering, self-organizing maps (Torkkola et al, 2001), and graphic theoretic approaches (e.g., CLICK (Sharan and Shamir, 2000)). These clustering methods seek full space clusters across all genes and all experiments. When applied to microarrays, they often produce a partition of genes that both precludes the assignment of genes to different clusters and fails to exclude irrelevant experiments. They are incapable of discovering gene expression patterns visible in only a subset of experimental conditions. In fact, it is common that a subset of genes are co-regulated and co-expressed

under a subset of conditions, but tend to behave independently under other conditions.

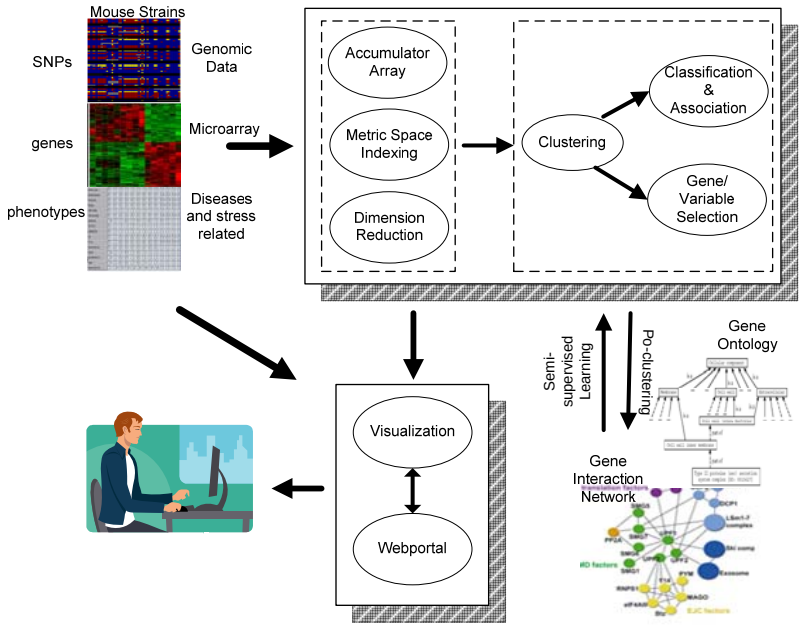
Recently, *biclustering* has been developed to uncover the local structures inside the gene expression matrix. Cheng and Church (2000) are among the pioneers in introducing this concept. Their biclusters are based on uniformity criteria, and they presented a greedy algorithm to discover them. Plaid (Lazzaroni, 2000) presented another model to capture the approximate uniformity in a submatrix in gene expression data and look for patterns where genes differ in their expression levels by a constant factor. Ben-Dor et al. (2001) discussed approaches to identify patterns in expression data that distinguish two subclasses of tissues on the basis of a supporting set of genes that results in high classification accuracy. Segal et al. (2001) described rich probabilistic models for relations between expressions, regulatory motifs and gene annotations. Its outcome can be interpreted as a collection of disjoint biclusters generated in a supervised manner. Tanay et al. (2002) defined a bicluster as a subset of genes that jointly respond across a subset of conditions, where a gene is termed responding under some condition if its expression level changes significantly under that condition with respect to its normal level. Ben-Dor et al. (2002) introduced the model of OPSM (order preserving submatrix) to discover a subset of genes identically ordered among a subset of conditions. It focuses on the coherence of the relative order of the conditions rather than the coherence of actual expression levels. For example, in the gene expression data of patients with the same disease, the genes interfering with the progression of this disease should behave similarly in terms of relative expression levels on this set of patients. These types of pattern can be observed in data from nominally identical exposure to environmental effects, data from drug treatment, and data representing some temporal progression, etc. One major drawback of this pioneering work is the strict order of the conditions enforced by the OPSM model.

Biclustering is also referred to as *subspace clustering* or *co-clustering* in the field of computer science, which has two main branches. One branch of subspace clustering algorithms divides both the set of objects and the set of attributes into disjoint partitions, where the partitions maximize global objective functions (Dhillon et al., 2003; Chakrabarti et al., 2004). Even though a globally optimal partition may be reached, the local properties of each cluster are hard to characterize. The other branch of subspace clustering algorithms eliminates the restriction of disjoint partitions by looking for clusters satisfying desired properties. These clustering algorithms are also called pattern-based algorithms. Unlike partition-based algorithms that search for the best global partitions, pattern-based algorithms allow one object to be in multiple clusters in different subspaces. Several pattern-based algorithms have been developed for different cluster properties. A commonly adopted property is that the set of points in a cluster are spatially close to each other in some subspace of the original high dimensional space (Agrawal et al., 1998; Aggarwal et al., 1999; Aggarwal and Yu, 2000). Note that subspace clustering based on spatial distance is limited in its ability to find clusters with high correlations. In biological applications, genes with different expression levels may still exhibit consistent up and down regulation patterns (called co-regulation patterns) under a subset of conditions. Recently, algorithms such as residue-based biclustering (Cheng and Church, 2000), Order preserving biclustering (Ben-Dor et al., 2002; Liu and Wang, 2003; Liu et al, 2004b) and the search of shifting and scaling patterns (Wang et al., 2002; Yang et al, 2005; Xu et al, 2006) were developed to look for specific co-regulation patterns. Our team at UNC is one of the leading groups conducting research in subspace clustering and has made several key contributions (Wang et al., 2002; Liu and Wang, 2003; Liu et al, 2004b ; Yang et al, 2005; Xu et al, 2006) in pattern-based subspace clustering algorithms.

2.2. Machine Learning Techniques in Prediction of Phenotypes

There has been extensive research in training classifiers for predicting phenotype values from gene expression data. Several classification techniques, including K-nearest neighbor classifier, decision tree, Support Vector Machine (SVM), and logistic regression, have been widely used. Among them, the margin-based classifiers including the support vector machine (SVM) and penalized logistic regression (PLR) have scored many successes. For example, SVM and PLR have been implemented with high accuracy using microarray gene expression data for cancer study (Brown et al., 1999; Mukherjee et al., 1999; Furey et al., 2000; and Zhu and Hastie 2004). In binary classification or supervised learning, margin-based techniques usually deliver high performance by implementing the idea of large margins. Specifically, given a training data set of n samples $\{(x_i, y_i), i=1,\ldots,n\}$ obtained from unknown probability distribution $P(x,y)$, where $y_i \in \{+1, -1\}$ is the outcome (of a given phenotype) of individual i with input x_i, then the goal is to build a classifier to predict class y for a new subject with given x. For such a problem, machine learning is performed by constructing a function f, mapping from x to y, such that $sign(f(x))$ is the classification rule. An important concept, so called margin $yf(x)$, is critical for the success of margin-based classifiers. For each sample pair (x_i, y_i), the margin $y_i f(x_i)$ indicates the correctness and strength of classification of x_i by f.

Figure 1. Data mining pipeline



A desirable classifier is one with good generalization ability, which is measured by the Generalization Error (GE). The GE, defined as the probability of misclassification, can be written as $Err(f)=P(yf(x)<0)=0.5E(1-sign(yf(x))$. A margin-based classifier with a loss function $V(u)$ tries to minimize the GE by using the loss function V which mimics the role of 1-*sign* or commonly called 0-1 loss, i.e., $0.5(1-sign)$. For example, SVM uses the hinge loss function with $V(u)=[1-u]_+$, see Lin (2000); penalized logistic regression (PLR) adopts the logistic loss $V(u)=log(1+exp(-u))$, see Wahba (1998); AdaBoost employs the exponential loss function $V(u)=exp(-u)$, see Friedman, Hastie and Tibshirani (2000); and the Ψ loss satisfies $U> \Psi(u)>0$ if $u \in (0, \tau]$ and $V(u)=1-sign(u)$ otherwise (Shen et al., 2003).

Despite the success in predicting discrete phenotype classes, margin-based classifiers cannot be applied to estimate the probability distribution of a phenotype directly. In our study, we are interested in both. For instance, it is more important to estimate the susceptibility to cancer of a given mouse line than to only give a boolean classification. Therefore, significant additional development is required in order to use large margin classifiers.

2.3. Challenges

A number of computational challenges lie ahead. These include, but are not limited to, dealing with data heterogeneity, handling high complexity, and the presence of noise. This project aims to address these issues and provide data mining and statistical analysis tools to enable learning from multiple types of data. This 1) will provide a framework from which mathematical models of the underlying biochemistry, genetics, and physiology can be developed; 2) will lead to the identification of biomarkers for the disease and 3) will describe new methods for prediction of disease progression and classification of mouse lines. Conceptually, the data can be thought of as organized into a large matrix. Each mouse line corresponds to a column and the rows represent SNPs, gene expressions, and phenotypic measurements captured, as illustrated in Figure 1.

Several key characteristics of this large data matrix complicate its analysis:

- The dimensionality is high since the data matrix contains massive amounts of information on (relatively) few subjects and there exist both complex correlations and causal relationships between variables.
- The data matrix is comprised of disparate measurements including both continuous and discrete variables, which may not be directly comparable to each other.
- The data matrix is not static, but growing, both in terms of adding new CC lines and measurements.

The data will eventually contain high-density SNPs (Single Nucleotide Polymorphism), or even whole genome sequences, for at least hundreds of CC lines and millions of phenotypic measurements (molecular and physiological) and other derived variables.

- Individual items may be contaminated, noisy or simply missing, which makes detectable relationships hard to "see", and thus hard to interpret.
- The number of unknowns far exceeds the number of knowns since relatively little is known about associations between polymorphisms to gene expression pathways to phenotypic observations. Moreover, it is likely that there is more than one pathway related to a given phenotypic observation, possibly characterized by different gene expression patterns.

Consequently, the number of potential hypotheses is extremely large, making it intractable to generate and test every possibility. New data structures and data mining methods are needed to address these challenges. We need to develop novel and scalable data management and mining techniques to enable high throughput genetic network analysis, real-time genome-wide exploratory analysis, and interactive visualization. This requires new methods to support instant access and computation for any user-specified regions and enable fast and accurate correlation calculation and retrieval of loci with high linkage disequilibrium.

3. Summary

In this presentation, I will outline the research scope depicted in Figure 1, which includes the following three directions.

- Developing efficient data structures and access methods to support efficient analysis of high-density data and interactive visualization
- Designing efficient methods for genetic network analysis using subspace pattern discovery techniques
- Developing efficient classification approaches to build prediction models based on the subspace patterns discovered above

In particular, I will discuss the successes and open problems from our experience in developing and integrating novel techniques in bioinformatics, data management, data mining, statistics, and knowledge management, to discover, summarize, and use subspace patterns for efficient analysis and visualization.

4. References

Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., and Park, J. S. (1999). Fast algorithms for projected clustering. In SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data, pages 61–72, New York, NY, USA. ACM Press.

Aggarwal, C. C. and Yu, P. S. (2000). Finding generalized projected clusters in high dimensional spaces. In SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pages 70–81, New York, NY, USA. ACM Press.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In SIGMOD Conference, pages 94–105.

Ben-Dor, A., Friedman, N., and Yakhini. Z. (2001) Class discovery in gene expression data. In RECOMB 2001.

Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2002) Discovering Local Structure in Gene Expression Data: The Order-Preserving Submatrix Problem. In RECOMB 2002.

Brown, M. W., Grundy, D. W., Liu, N., Christianini, C., Sugnet, M., Ares Jr., and Haussler, D. (1999). Support vector machine classification of microarray gene expression data. UCSC-CRL 99-09, Department of Computer Science, University of California Santa Cruz.

Chakrabarti, D., Papadimitriou, S., Modha, D. S., and Faloutsos, C. (2004). Fully automatic cross-associations. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 79–88, New York, NY, USA. ACM Press.

Cheng, Y. and Church, G. (2000) Biclustering of expression data. In ISMB, 2000.

Dhillon, I. S. and Guan, Y. (2003) Information theoretic clustering of sparse co-occurrence data. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03), pages 517–521, 2003.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. Annals of Statistics. 28:337-407.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M. and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906-914.

Lazzeroni, L. and Owen, A. (2000) Plaid models for gene expression data. http://www.stanford.edu/ owen/plaid/

Lin, Y. (2000). Some asymptotic properties of the support vector machine. Technical report 1029, Department of Statistics, University of Wisconsin-Madison.

Liu, J. and Wang, W. (2003) Subspace clustering by tendency in high dimensional space. In ICDM 2003.

Liu, J., Yang, J., and Wang, W. (2004a) Gene ontology friendly biclustering of expression profiles. Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB), 2004.

Liu, J., Yang, J., and Wang, W. (2004b) Biclustering of gene expression data by tendency. Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB), 2004.

Liu, J., Wang, W. and Yang, J. (2004c) A framework for ontology-driven subspace clustering, Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 623-628, 2004.

Liu, J., Strohmaier, K., and Wang, W. (2004d) Revealing true subspace clusters in high dimensions, Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), pp. 463-466, 2004.

Liu, J., Paulsen, S., Xu, X., Wang, W., Nobel, A., and Prins, J. (2006) Mining Approximate frequent itemset in the presence of noise: algorithm and analysis, Proceedings of the 6th SIAM Conference on Data Mining (SDM), 2006.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J. and Poggio, T. (1999). Support vector machine classification of microarray data. Technical Report AI Memo1677, MIT.

Segal, E., B. Taskar, A. Gasch, N. Friedman and D. Koller. (2001) Rich probabilistic models for gene expression. Bioinformatics, 17, S243-S252.

Sharan, R. and Shamir, R. (2000) Click: A clustering algorithm with applications to gene expression analysis. In ISMB, pages 307-216.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003). On Ψ-learning. Journal of the American Statistical Association, 98, 724-734.

Tanay, A., Sharan, R. and Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics, Vol.18, pages S136-S144.

Torkkola, K., Gardner, R. M., Kaysser-Kranich, T., and Ma, C. (2001) Self-organizing maps in mining gene expression data, Information Sciences Volume 139, Issues 1-2 , November 2001, Pages 79-96.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: B. Scholkopf, C. J. C. Burges and A. J. Smola (eds), Advances in Kernel Methods: Support Vector Learning, MIT Press, 125-143.

Wang, H., Wang, W., Yang, J., and Yu, P. S. (2002). Clustering by pattern similarity in large data sets. In SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pages 394–405, New York, NY, USA. ACM Press.

Xu, X., Tung, A. K. H., Lu, Y., and Wang, W. (2006) Mining shifting-and-scaling co-regulation patterns on gene expression profiles, Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), 2006.

Yang, J., Wang, H. Wang, W., and Yu, P. (2005) An improved biclustering method for analyzing gene expression profiles. International Journal on Artificial Intelligence Tools (IJAIT), vol. 14, no. 5, pp. 771-789.

Yang, T., Liu, J., McMillan, L., and Wang, W. (2006) A fast approximation to multidimensional scaling, by Proceedings of the ECCV Workshop on Computation Intensive Methods for Computer Vision (CIMCV), 2006.

Zhu, J. and Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. Biostatistics, 5(3), 427-444.