Biomarker Selection from High-Dimensionality Data

Data management has become an overwhelming task in the biomarker discovery field as technologies have become more efficient and high-throughput. Classical approaches fall short of providing the necessary power to analyze the high-dimensionality data provided by modern genomics and proteomics studies. Currently, efforts are underway to reduce the dimensionality of the overall data set to a small, meaningful subset that can comprehensively explain the hypothesis being tested, while maintaining the integrity of the biological data itself.

Shawn E. Levy, Alexander Statnikov, and Constantin Aliferis

As technology development in the biological sciences continues to rapidly evolve, more and more researchers are presented with the challenge of analyzing and interpreting large amounts of data as investigative technologies become more efficient and high-throughput. The genome sequencing projects, coupled with technology advancements in genomics and proteomics, have provided scientists with the ability to cast a genome-wide net and capture an extraordinary amount of data about their system of interest. Unfortunately, few are prepared to fully appreciate data on this scale, and furthermore, classical statistical approaches are unable to achieve power when applied to such data sets since the number of potential factors (e.g., observed genes, proteins, or genetic variants) exceeds the number of samples analyzed. Although the challenges associated with the analysis of high-dimensionality data have been well recognized in various disciplines, biologists are just beginning to embrace the massive amount of data that is available from studies utilizing modern genomic or proteomic techniques. As shown in nearly all manuscripts published utilizing these techniques, a fractional percentage of data obtained from genomic and proteomic technologies is reported as important to the question being analyzed. Methodologies to reduce the dimensionality of the overall data set (ranging typically from tens of thousands to hundreds of thousands of predictor variables) down to a small, meaningful, reliable, and comprehensive subset that explains the hypothesis or condition being tested is a very active area of

Shawn E. Levy, Alexander Statnikov, and Constantin Aliferis work in the Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN. He can be reached via e-mail at: shawn.levy@vanderbilt.edu.

research. These efforts have brought together typically diverse disciplines of mathematics, biology, computer science, physics, and statistics with the underlying goal of developing statistical and informatic tools and methodologies to reduce the dimensionality of the data to a level meaningful to the researcher while maintaining the integrity of the data in its biological context.

Effective identification of biomarkers for any phenotype of interest, be it a disease, treatment response, or genetic alteration, relies on effective experimental design and the application of appropriate data-processing techniques to yield a meaningful data set. While a detailed examination of these methodologies and techniques is outside the scope of this review, it is important to state that the most effective experiments begin with a hypothesis in a biological context followed by designing a profiling experiment around defined questions that can be answered by the experiment. For example, the design of an experiment to identify a group of genes able to differentiate adenocarcinoma from squamous cell carcinoma in the lung (class discovery/class prediction) (1) would be very different than an experiment designed to identify genes that are differentially expressed in a single tissue of a mouse model developed using transgenic technologies. The hypothesis of the experiment and its appropriate design, as well as sample consideration, are important at the design stage when one considers the resolution of the experiment. Using the adenocarcinoma versus squamous cell carcinoma example above, the number of samples that would be required to identify a classification and prediction model is very likely to be significantly less than the number of samples required to not only classify the sample but also to predict which patients within those samples will have a good versus poor prognosis or a particular response to therapy. Many such factors must be considered at the outset of the profiling experiment, often informed by preliminary experiments, to maximize the likelihood that a list of appropriate biomarkers predictive or prognostic of the phenotype of interest can be developed. Overall, the importance of appropriate experimental design and the impact of normalization, signal, and background filtering and overall statistical design cannot be overemphasized. Two recent supplements and books (2-5) offer excellent reviews of microarray technologies applied to designing and analyzing DNA microarray experiments with themes that can be extended to other profiling techniques.

The identification of biomarkers that provide fundamental or key insight into the understanding of complex disease is often very challenging, even in the context of genome-scale profiling experiments. There is significant evidence that the vast majority of complex diseases, including cancer (6), are genetically linked (7). Many of the diseases most commonly reported in the popular press, including heart disease, diabetes, and obesity, are believed to have a complex polygenic basis. Given the widespread nature of polygenic diseases compared to single gene diseases, such as cystic fibrosis, the potential impact on public health is enormous. The fundamental problem that many of the researchers leading these efforts face is the inability of classical statistical approaches to achieve power in the analysis of genome-scale data sets. Invariably, the number of variables being assayed vastly out numbers the number of samples in the study. As referenced above, this problem is not unique to biological investigations. Other disciplines that collect large amounts of data on comparatively few subjects are faced with the same challenges (e.g., text categorization, automated detection of epidemics, computational biology, and analysis of complex and time-evolving signals). Efforts in these fields as well as in the field of bioinformatics are beginning to provide methods to effectively and efficiently meet these challenges. The remainder of this brief review will discuss recent work from our group and others on methods that can be applied to complex biological data sets to identify key biomarkers that are informative for the phenotype of interest.

Genome scale profiling experiments all have thousands of predictor variables (or discriminatory features) and relatively few individual samples. Developing analysis models for such data structures can be complicated and the computational methods required are often intensive. Close on the heels of the first manuscripts describing the application of microarrays to biological analysis came several seminal papers describing discrimination and cluster analysis methods for RNA expression analysis (8-12). Most of this work made use of various clustering algorithms (agglomerative hierarchical clustering, divisive k-means clustering, or self-organizing maps) to characterize individual samples, identify patterns in the data, or discriminate between samples. These clustering methods, especially hierarchical, have become almost synonymous with the microarray literature and a red-green, heat-map representation of differential gene expression has become a widely accepted method to illustrate patterns in large amounts of data. However, none of these early analysis techniques include statistical methods for modeling gene expression patterns. Significant work done around the same time as the early clustering work was aimed at assessing parametric (13-15) and non-parametric (16) models to address questions regarding formal assessment of differences between gene patterns or the fit of a specific model to the data. The application of Bayesian approaches was quickly recognized as a powerful tool in the analysis of high-dimensionality data. The value of the Bayesian approach is further emphasized given the high degree of noise and variability often seen in microarray data. In general terms, Bayesian statistics can be described as a method that predicts the probability of a model based on the data. This is in contrast to frequentist or classical statistics where probability is based on tests of significance by supposing that a hypothesis is true (the null hypothesis) and then calculating the probability of observing a statistic (i.e., a function of the data) at least as extreme as the one actually observed during hypothetical repeated experiments (this is the P-value). In other words, classical statistics gives the probability of crucial aspects of the data taking specific value ranges and assuming a model, thus enabling rejection of models - or model parameter values - that have a small probability. Bayesian approaches predict the probability of a model (or of model parameters) given the data. While a more detailed discussion of the advantages of Bayesian approaches in relation to biological data is outside the scope of this review, we direct you to the following recent references (15, 17-20).

Although genes are often named based on a predicted role or function, most genes are involved in multiple pathways and this involvement may be dynamic depending on tissue, developmental time, or disease (21). Bayesian decomposition has shown great promise in elucidating these relationships using a model based on prior information (that can be continually added and modified as more data is available) as genes are not restricted to a single group and can be in several groups simultaneously. This application of Bayesian approaches using a biological context to frame the data analysis has shown great promise in recent work and further strengthens the potential of Bayesian approaches for identifying biomarkers based on statistical models in a biological context by utilizing networks of biological pathways (22). The complexity of human genetic networks makes this technique challenging, but also emphasizes the strength of applying Bayesian decomposition approaches to analyze data. The current efforts to codify the mouse (23) and human phenome (24) (a comprehensive cataloging of physiological, anatomical, and behavioral data) will provide a rich data source for building biological models and using Bayesian methods to test the validity of these models.

The data methods discussed so far are by no means comprehensive. Many other pattern recognition techniques have been described for or applied to microarray data sets over the past several years. These include Genetic Algorithms, Support Vector Machines, several types of Neural Networks, Markov models, and others. A review article by Valafar describes these and other methods in great detail (25). Although these pattern recognition methods are excellent tools for the analysis of genome-scale data sets, most are complex to a novel user and can be very computationally intensive. In an effort to help overcome the steep learning curve in applying some of these methods, the authors and their collaborators have performed a comprehensive evaluation of the major multicategory classification algorithms (26) and based on this evaluation, have developed a software system that supports the effective application of Support Vector Machine approaches for novice users (27). Additional work by our group and collaborators has described the application of a Markov Blanket algorithm in an efficient, stable, and novel manner for variable selection from complex data sets that can be applied to classification, regression, and prediction studies (28). Finally, another family of multivariate analysis methods is based on Principle Component Analysis (PCA). Such methods have been applied to microarray data with success (29). PCA, as a tool to reduce the dimensionality of the overall dataset to a meaningful subset and then apply gene annotations to that subset using gene ontology terms, pathway information, or information from the previously mentioned phenome projects, may be a promising means to identify biomarkers from a biological perspective.

The purpose of this review was to introduce and describe biomarker selection methods that extend beyond the application of a t-test or basic ANOVA. While these statistical methods are efficient and effective in most contexts, it is important to consider other potentially more robust data analysis procedures

Table I:

Name	Version	Developer	Supervised classification
ArrayMiner ClassMarker	5.2	Optimal Design, Belgium	• K-Nearest Neighbors • Voting
Avadis Prophetic	3.3	Strand Genomics, USA	• Decision Trees • Neural Networks • Support Vector Machines
BRB ArrayTools	3.2 Beta	National Cancer Institute, USA	• Compound Covariate Predictor • Diagonal Linear Discriminant Analysis • Nearest Centroid • K-Nearest Neighbors • Support Vector Machines
caGEDA	(accessed 10/2004)	University of Pittsburgh & University of Pittsburgh Medical Center, USA	• Nearest Neighbors methods • Naïve Bayes Classifier
Cleaver	1.0 (accessed 10/2004)	Stanford University, USA	• Linear Discriminant Analysis
GeneCluster2	2.1.7	Broad Institute, Massachusetts Institute of Technology, USA	• Weighted Voting • K-Nearest Neighbors
GeneLinker Platinum	4.5	Predictive Patterns Software, Canada	• Neural Networks • Support Vector Machines • Linear Discriminant Analysis • Quadratic Discriminant Analysis • Uniform/Gaussian Discriminant Analysis
GeneMaths XT	1.02	Applied Maths, Belgium	• Neural Networks • K-Nearest Neighbors • Support Vector Machines
GenePattern	1.2.1	Broad Institute, Massachusetts Institute of Technology, USA	• Weighted Voting • K-Nearest Neighbors • Support Vector Machines
Genesis	1.5.0	Graz University of Technology, Austria	• Support Vector Machines
GeneSpring	7	Silicon Genetics, USA	• K-Nearest Neighbors • Support Vector Machines
GEPAS	1.1 (accessed 10/2004)	National Center for Cancer Research (CNIO), Spain	• K-Nearest Neighbors • Support Vector Machines • Diagonal Linear Discriminant Analysis
MultiExperiment Viewer	3.0.3	The Institute for Genomic Research, USA	• K-Nearest Neighbors • Support Vector Machines
PAM	1.21a	Stanford University, USA	• Nearest Shrunken Centroids
Partek Predict	6.0	Partek, USA • Discriminant	• K-Nearest Neighbors • Nearest Centroid Classifier
Weka Explorer	3.4.3	University of Waikato, New Zealand	• K-Nearest Neighbors • Decision Trees • Rule Sets • Bayesian Classifiers • Support Vector Machines • Multi-Layer Perception • Linear Regression • Logistic Regression • Meta-Learning Techniques (Boosting, Bagging)

Cross-validation for performance estimation	Automatic model selection for classifier & gene selection methods	URL
Yes	No	http://www.optimaldesign.com/ArrayMiner
Yes	No	http://avadis.strandgenomics.com/
Yes	No	http://linus.nci.nih.gov/BRB-ArrayTools.html
Yes	No	http://bioinformatics.upmc.edu/GE2/GEDA.html
Yes	No	http://classify.stanford.edu
Yes	No	http://www.broad.mit.edu/cancer/software
Yes	No	http://www.predictivepatterns.com/
Yes	No	http://www.applied-maths.com/genemaths/genemaths.htm
Yes	No	http://www.broad.mit.edu/cancer/software
No	No	http://genome.tugraz.at/Software/Genesis/Genesis.html
Yes	No	http://www.silicongenetics.com
Yes	Limited (for number of genes)	http://gepas.bioinfo.cnio.es/tools.html
Yes	No	http://www.tigr.org/software/tm4/mev.html
Yes	Limited (for a single parameter of the classifier)	http://www-stat.stanford.edu/~tibs/PAM/
Yes nalysis of gene selection algorithms)	Limited (does not allow optimization of the choice	http://www.partek.com/
Yes	No	http://www.cs.waikato.ac.nz/ml/weka/

Table 1 provides a summary of several multicategory selection tools, their capabilities, and a URL for access to the tool. Many of these software tools are designed to assist the novice user in applying the classification method to their data set. It is important to note that the applicability of the tool from an experimental design perspective must be determined by the user. Table 1 adapted from (27).

that offer additional insight into the data set being analyzed. These techniques are often able to handle error, noise, or variability in the data more efficiently. For example, consider a two-sample t-test applied to a small data set (30) that then ranks genes based on a test statistic or corresponding P-value. This method assumes that the gene expression values are normally distributed and no correlation exists among the genes. Testing corrections can help with some of these issues but the underlying result is that the genes identified as most informative are those that show the largest differential expression. When searching for informative biomarkers for disease prediction, classification, prognosis, or drug target action or response, this assumption must be carefully considered. In nearly all cases, it is worth investigating the application of one or more of the techniques reviewed here. Table 1 provides a summary of several software tools that implement many of the methods described here.

Particular experimental design criteria and hypotheses will define when the techniques are appropriate and they nearly universally outperform more classical methods. As profiling technologies continue to evolve and increase in complexity, comprehensiveness, and sensitivity, more robust data analysis methods will become more vital to the success of these types of experiments. How the challenges of this new level of data complexity will be met will be an interesting and fertile research area. As the complexity of genomic data continues to increase, the analysis methods must match the pace of the technology in order to become a valuable tool in understanding the complexities of the human genome.

Citations

1. N. Yamagata, Y. Shyr, K. Yanagisawa, et al., Clin Cancer Res 9, 4695-4704 (2003).
2. Various authors, Nature Genet 32, 461-552 (2002).
3. Various authors, Nature Genet 37, 1-45 (2005).
4. S. Knudsen, in Guide to Analysis of DNA Microarray Data. 2nd ed, (John Wiley and Sons, New Jersey, 2004).
5. G. Parmigiani, E.S. Garrett, R.A. Irizarry and S.L. Zeger, in The Analysis of Gene Expression Data: Methods and Software. Statistics for Biology and Health, K. Dietz, Ed. (Springer, New York, 2003).
6. R.S. Houlston and J. Peto, Oncogene 23, 6471-6476 (2004).
7. B. Rannala, Am J Pharmacogenomics 1, 203-221 (2001).
8. P. Tamayo, D. Slonim, J. Mesirov, et al., Proc Natl Acad Sci USA 96, 2907-2912 (1999).
9. P.T. Spellman, G. Sherlock, M.Q. Zhang, et al., Mol Biol Cell 9, 3273-3297 (1998).
10. M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Proc Natl Acad Sci USA 95, 14863-14868 (1998).
11. T.R. Golub, D.K. Slonim, P. Tamayo, et al., Science 286, 531-537 (1999).
12. T. Hastie, R. Tibshirani, M.B. Eisen, et al., Genome Biol 1, 3.1-3.21 (2000).
13. M.A. Newton, C.M. Kendziorski, C.S. Richmond, et al., Comput Biol 8, 37-52 (2001).
14. M. West, C. Blanchette, H. Dressman, et al., Proc Natl Acad Sci USA 98, 11462-11467 (2001).
15. J.G. Ibrahim, M.H. Chen and R.J.Gray, Journal of the American Statistical Association 97, 88-99 (2002).
16. B. Efron and R. Tibshirani, Genet Epidemiol 23, 70-86 (2002).
17. P. Baldi and A.D. Long, Bioinformatics 17, 509-519 (2001).
18. A.D. Long, H.J. Mangalam, B.Y. Chan, et al., J Biol Chem 276, 19937-19944 (2001).
19. D. Husmeier, Biochem Soc Trans 31, 1516-1518 (2003).
20. D. Husmeier, Bioinformatics 19, 2271-2282 (2003).
21. K.H. Wolfe and W.H. Li, Nat Genet 33 Supplement, 255-65 (2003).
22. M.F. Ochs, T.D. Moloshok, G. Bidaut and G. Toby, Ann N Y Acad Sci 1020, 212-226 (2004).
23. M.A. Bogue and S.C. Grubb, Genetica 122, 71-74 (2004).
24. R.G. Steen, (2005), http://hpd.mcw.edu/
25. F. Valafar, Ann N Y Acad Sci 980, 41-64 (2002).
26. A. Statnikov, C.F. Aliferis, I. Tsamardinos, et al., Bioinformatics 21, 631-543 (2005).
27. A. Statnikov, I. Tsamardinos, Y. Dosbayev and C.F. Aliferis, Int J Med Inform 74, 491-503 (2005).
28. C.F. Aliferis, I. Tsamardinos and A. Statnikov, AMIA Annu Symp Proc 21-25 (2003).
29. A. Wang and E.A. Gehan, Stat Med 24, 2069-2087 (2005).
30. J.M. Satagopan and K.S. Panageas, Stat Med 22, 481-499 (2003).