



New Directions for Postgenomic Bioinformatics

1405 Bernerd Place Rockville, MD 20851

www.Biomind.com

Machine Learning Algorithms for Clinical and Research Microarray Data Analysis

Mining Microarray Data to Discover:

Disease Biomarkers & Complex Genetic Relationships

Biomind LLC WHITE PAPER

January 2006

Objective:

Demonstrate the ability of Biomind's machine learning algorithms to more accurately identify disease biomarkers in microarray data than conventional analysis methods

Commercial Capability:

Data analysis collaboration in which the most accurate disease biomarkers are identified from clinical microarray data

Molecular biomarkers associated with disease and disease predisposition may be used for diagnostic purposes in the early detection and characterization of various disorders. Microarray and SNP data have been used extensively based upon their respectively high resolution of gene expression and polymorphism. And, while diagnostic, pharmacogenomic, and research uses for such biomarkers have proliferated, methods for their identification have standardized. Biomind has developed software which sifts through large, complex microarray datasets to accurately identify biomarkers implicit in clinical disease data. The software uses machine learning algorithms which integrate the Gene Ontology (GO) and Protein Information Resource (PIR).

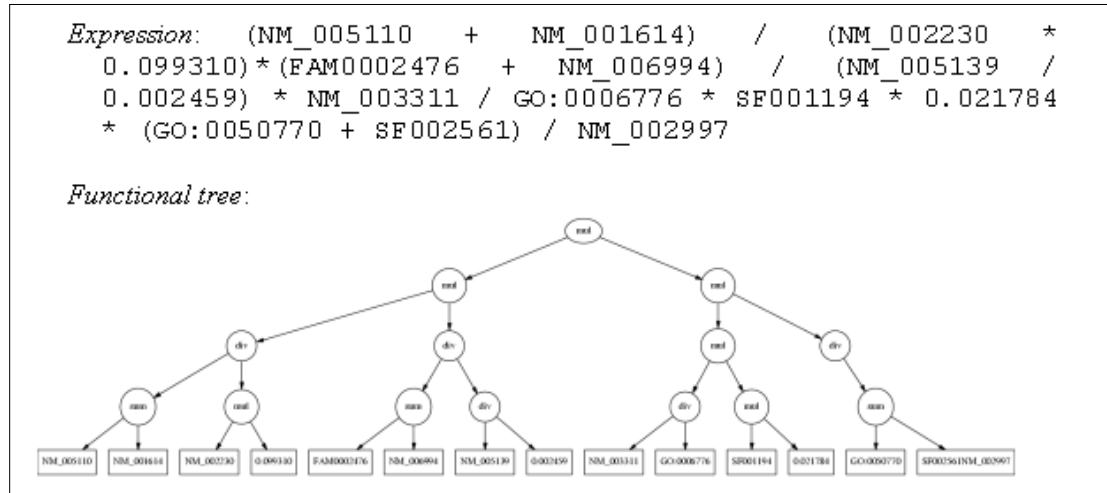
Traditional methods for identifying biomarkers in microarray data rely upon differential expression followed by clustering techniques. These methods give insight into the effects of individual genes, considered in isolation. However, many experimentally important genes are not significantly (over or under) expressed in the relevant samples, and are therefore ignored by differentiation analysis. Machine learning algorithms which search for nonlinear patterns and integrate relevant knowledge resources can identify a more complete set of experimentally important gene and gene features. These nonlinear patterns better identify the biomarkers which explain clinical outcomes and endpoints. The process highlights

relevant genes, gene combinations and gene interactions implicit in the microarray data.

Biomind's machine learning algorithms generate a set of mathematical rules or *classification models* which best explain how microarray data is distributed among predetermined categories (e.g. case vs. control, time series etc.). Mathematicians refer to this sort of machine learning process as *supervised categorization*. Decision trees, neural networks, logistic regression, support vector machines, and genetic programming are all examples of supervised categorization algorithms. Biomind utilizes *support vector machines* and *genetic programming* in its commercial analysis software, *ArrayGenius*TM.

Biomind's classification models combine genes, gene combinations, gene ontologies, and protein families, collectively referred to as *features*, in complex algebraic equations. Features are derived from the experimental microarray data and its links to the Gene Ontology and Protein Information Resource. The 100 most valuable mathematical rules found for dividing the data between categories are listed. The members of this "model ensemble" essentially serve as diagnostic rules. For example, a rule serving as a classification model is illustrated in Fig. 1.

Figure 1: BioMind ArrayGenius™ Classification Model Example



ArrayGenius™ then extracts the most commonly seen features from these models in a more simple list format. The features themselves are then clustered in order to see which occur together—again, feature referring to gene, gene ontology, or protein family. This is a paradigm-shift: the identification of features which appear together in multiple machine learning models in a model ensemble, instead of viewing genes grouped (clustered) by differential expression.

Biomind and its collaborators have written several papers awaiting publication which describe these analyses (see endnote)

The genes and gene features (biomarkers) which ArrayGenius™ generates should be examined for their power in predicting clinical outcomes. For evaluation purposes, this is most easily done by comparing the software's ability to diagnose a verifiable clinical condition versus the most accurate genomic predictors currently reported in the scientific literature.

Figure 2 demonstrates that Biomind’s analysis predicts the clinical condition with comparable or greater accuracy in several circumstances (SVM=Support Vector Machines, GP = Genetic Programming).

Figure 2: ArrayGenius’ diagnostic accuracy

Dataset	Method	Accuracy with Enhancement	Accuracy without Enhancement	Accuracy in Literature
Lung Cancer	SVM	100.0%	100.0%	93.3%
	GP	97.0%	91.3%	
Prostate Tumor	SVM	100.0%	94.1%	73.5%
	GP	97.0%	100.0%	
ALL/AML	SVM	67.6%	94.1%	100.0%
	GP	79.4%	73.5%	
Aging Brain	SVM	100.0%	95.0%	---
	GPC	95.0%	70.0%	
DLBCL	SVM	94.8%	97.4%	97.5%
	GPC	81.8%	77.9%	

These results, among others, validate ArrayGenius’ ability to identify important biomarkers. A careful analysis of the classification models underlying these statistics also suggests that Biomind’s general feature identification results are more important in explaining experimental conditions than differential expression analysis. Extensive wet biology experimentation may bear this out; and such work is currently underway in the context of Biomind’s existing research collaborations.

Biomind's algorithms are more powerful than differential expression in biomarker identification because:

1. Nonlinear interactions among genes and functional groups are identified

2. Multiple, diverse biological data sources are considered in the analysis
3. Clustering based upon machine learning features identifies **combinations** of features with even greater predictive capability
4. Diagnostic rules are automatically generated

Biomind has also successfully applied these algorithms to SNP and DNA mutation associations in clinical data. Biomind and the University of Virginia have discovered heteroplasmic mitochondrial DNA mutations associated to Parkinson's disease. And, SNP combinations have been used by Biomind and CDC researchers to identify and predict Chronic Fatigue Syndrome patient samples.

ImmPort Web Portal for the NIH NIAID

Biomind, in conjunction with Northrop Grumman, has participated in the construction of ***ImmPort***, a web portal for the analysis of immunology and infectious disease, functional-genomic data. Via this portal site, NIH NIAID and its funded researchers will use Biomind's array analysis and SNP algorithms to correlate biomarkers to disease states. The portal securely accepts microarray data from disparate research sites and analysis it on a Linux cluster running Biomind ArrayGenius.

In Immport as in other Biomind applications, Biomind's advanced data analysis algorithms enable more accurate characterization of disease biomarkers. Biomarkers that were not otherwise apparent are identified. Existing biomarkers of great complexity are simplified to several genes (features)—driving down assay costs. Biomind's machine learning software platforms are enabling translational medicine.

Working Papers:

Smigrodzki, R., Goertzel B., Pennachin C., Coelho L., Prosdocimi F., and Parker W. *Genetic Algorithm for Analysis of Mutations in Parkinson's Disease*

Goertzel, B., Pennachin C., Coelho L., Vernon S., Whistler T. *Identifying Complex Biological Interactions based on Static Gene Expression Data via Classification-Model-Usage-Based Clustering.*

Pennachin, C., Coelho, L., Goertzel, I., Queiroz, M., Prosdocimi, F., Lobo, F., and Goertzel, B. *Knowledge-Guided Analysis of Gene Expression Data Using Genetic Programming, Support Vector Machines, and the Gene Ontology and PIR Databases.*

Queiroz, M., Prosdocimi, F., Goertzel, I., Lobo, F, Pennachin, C., and Goertzel, B. *Inferring Gene Ontology Category Membership via Cross-Experiment Gene Expression Data Analysis.*

Goertzel, B, Pennachin C, Coelho L, Vernon S, Whistler T. *Analysis of Microarray Data via Symbolic Regression and Hierarchical Evolutionary Computing.*