Biomind ArrayGenius and GeneGenius:
Web Services Offering Microarray and SNP Data Analysis via Novel Machine Learning Methods

Ben Goertzel, Cassio Pennachin, Lucio Coelho, Leonardo Shikida, Murilo Queiroz

Biomind LLC
1405 Bernerd Place
Rockville MD 20851
ben@goertzel.org, cassio@biomind.com, lucio@biomind.com, kenji@biomind.com, murilo@biomind.com

Abstract

Analysis of postgenomic biological data (such as microarray and SNP data) is a subtle art and science, and the statistical methods most commonly utilized sometimes prove inadequate. Machine learning techniques can provide superior understanding in many cases, but are rarely used due to their relative complexity and obscurity. A challenge, then, is to make machine learning approaches to data analysis available to the average biologist in a user-friendly way. This challenge is addressed by the Biomind ArrayGenius product, an easy-to-use Web-based system providing microarray analysis based on genetic programming, kernel methods, and incorporation of knowledge from biological ontologies; and GeneGenius, its sister product for SNP data. This paper focuses on the obstacles faced and lessons learned in the course of creating, deploying, maintaining and selling ArrayGenius and GeneGenius – many of which are generic to any effort involving the creation of complex AI-based products addressing complex domain problems.

Introduction

Biology these days is characterized by large and complex datasets, whose analysis and interpretation poses a major challenge. This challenge is most commonly met via statistical methods, but the shortcomings of conventional statistical techniques are becoming more widely recognized: surprisingly often, they miss highly significant patterns in the data. In cases where quantitative comparison is possible, it has often been shown that machine learning techniques can provide better performance than standard statistical methods. And, in cases where qualitative rather than quantitative results assessment is appropriate, it has often been observed that machine learning methods give rise to insights beyond those offered by standard statistics.

Examples of these general points occur, for example, in the analysis of gene expression microarray data and SNP data, which are the types of data analyzed by the Biomind ArrayGenius and GeneGenius products that we will discuss here. Gene expression datasets contain information, typically for a number of tissues drawn from a number of individuals, on the measured expression levels of various genes (or ESTs) in those individuals. SNP datasets contain information, typically for a number of individuals, on the single-nucleotide polymorphisms found in each individual's DNA. Since humans have around 25000 genes, these datasets may become quite voluminous. Furthermore, microarray data is typically very noisy, so individual gene expression values must be taken with a grain of salt, whereas patterns spanning multiple gene expression values may be taken more seriously.

Suppose for example one has gene expression data regarding tissue samples from 50 individuals with lung cancer, versus 50 controls. A typical, simplistic statistical approach would be to look for genes whose expression levels are highly differentiated between the case and control categories. But this approach obviously does not provide maximum performance in terms of either quantitative categorial discrimination or qualitative biological understanding. What a machine learning approach provides is the capability to find complex combinations of genes that distinguish case from control. This "supervised categorization" approach generally provides a higher classification accuracy, and also does a better job of pinpointing which genes and gene-combinations are most important for the categorial distinction. A similar conclusion holds for SNP data, as well as other types of biological data. These facts are well documented in the scientific literature, using case studies regarding a variety of different diseases; see e.g. (Kothapalli, 2002; Lyons-Weiler, 2003; Rockett, 2003; Simon, 2003).

So, why are machine learning methods not more widely used for bioinformatic data analysis? The main reason seems to be their unfamiliarity to the biological community. Biologists are typically trained in statistics, but not in AI or machine learning.

The motivation underlying the creation of the Biomind ArrayGenius and GeneGenius products was a desire to make advanced machine learning methods available to the biology community, in a way that does not require deep AI

expertise on the part of the biologist end-users of the products. The products provide an easy-to-use graphical interface allowing machine-learning analysis of gene expression and SNP datasets, using a selection of standard
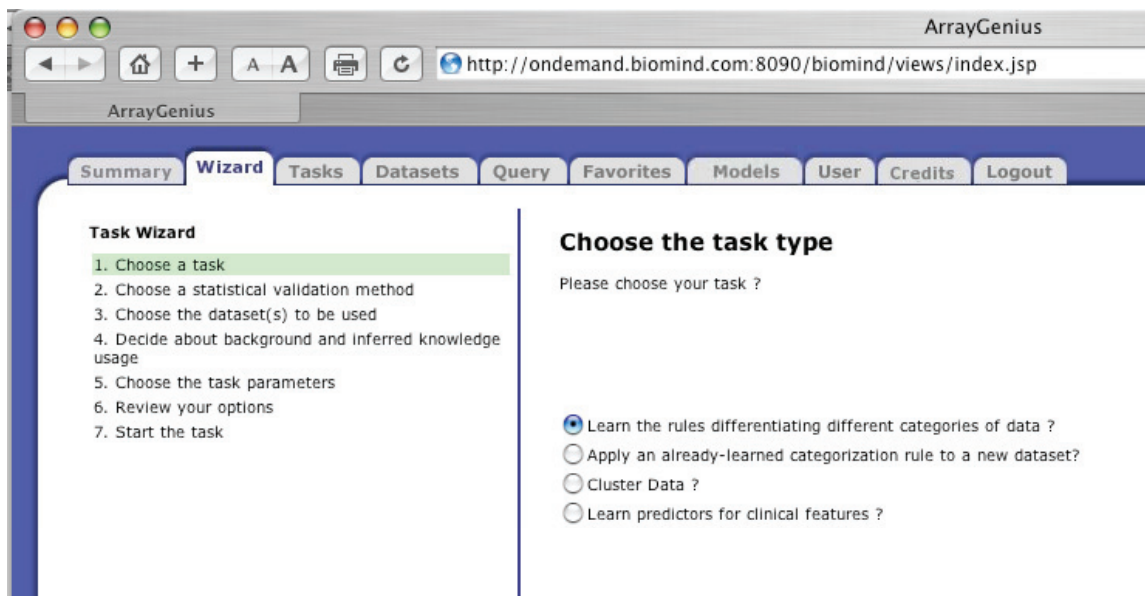

Figure 1. The Biomind ArrayGenius Task Wizard: Step 1. The Task Wizard leads the user through a series of steps, gathering information about the type of data analysis they want to do. Behind the scenes, the software then launches an ensemble of machine learning tasks intended to fulfill the user's high-level specifications.

supervised categorization and clustering algorithms, and also using novel machine learning techniques developed by Biomind LLC's research staff. These products have now been sold to a number of customers, and utilized to generate research results published in biology journals.

The algorithms underlying the ArrayGenius and GeneGenius products have been discussed in prior publications, as have some of the scientific results obtained using the products. This paper takes a somewhat different focus; it is concerned mainly with the challenges faced and the lessons learned in the context of creating, deploying, maintaining and selling the products. Many of the issues faced are generic ones that are relevant to anyone involved with creating a complex AI-based software product addressing a complex application domain.

Biomind ArrayGenius

ArrayGenius, Biomind LLC's flagship product, is an enterprise software system for microarray data analysis, which delivers advanced analytical functionality to end users via an easy-to-use Web UI. ArrayGenius combines advanced machine learning algorithms with massive volumes of background information, including biological ontologies, to deliver powerful data analysis functionality.

Practical applications include:

- **Biomarker Discovery**, including cases where biomarkers involve complex multiple gene interactions
- **Biological Processes Interpretation**, including dynamics and pathways underlying diseases and other phenotypic characteristics
- **Personalized Medicine**, including identification of individuals likely to suffer toxic reactions to particular drugs
- **Fundamental Research**, including gene function and metabolic pathway research

Hierarchical and partitioning based clustering methods are offered, but the core of the product AI-wise consists of

- supervised categorization technology, including Genetic Programming (Koza, 1992) plus Support Vector Machines (Cristianini and Shaw-Taylor, 2000.)
- methods for using biological ontologies to create "enhanced feature vectors" for input to supervised categorization algorithms.
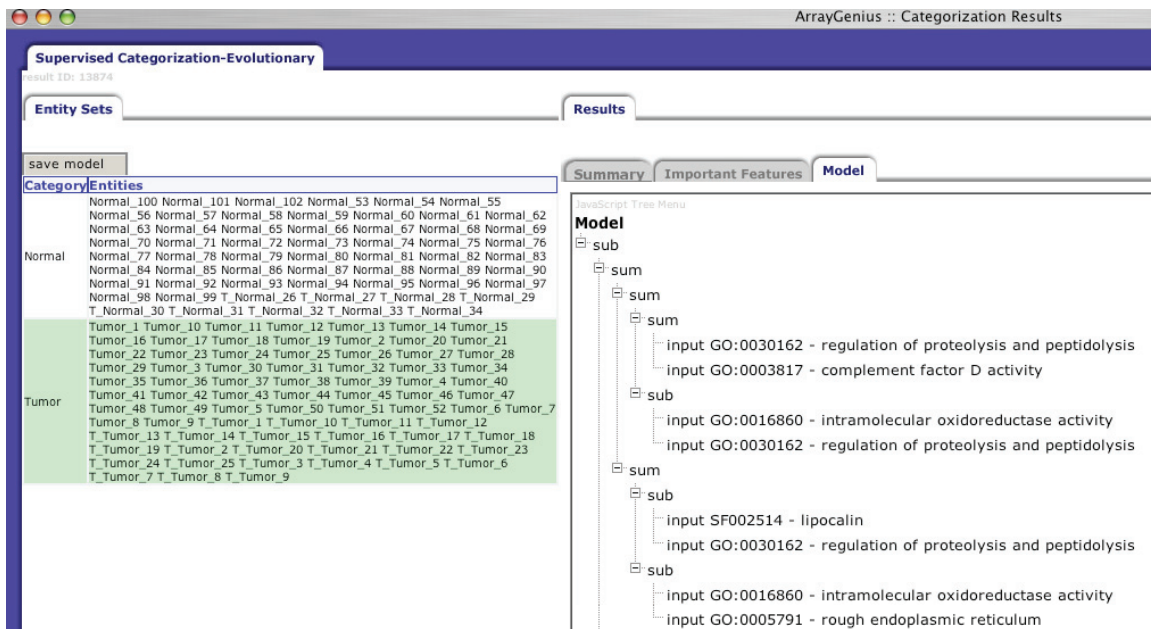
Figure 2. A Genetic Programming categorization model visually displayed within the Biomind ArrayGenius. The image only shows part of the categorization model, for space reasons. The terminals of the program tree refer to Gene Ontology (GO) and Protein Information Resource (PIR) categories, which are evaluated on an individual's microarray profile via averaging the gene expression values of the genes within the category, as described in (Goertzel et al, 2006a).

- a set of unique tools for statistically analyzing ensembles of supervised classification models, so as to provide added qualitative biological insigh into the genes and gene-combinations critical to a dataset

The high classification accuracy provided by the supervised categorization algorithms is one selling-point, but a lesson learned early on in the sales process was that accuracy alone is not of interest to most biological researchers. Microarray data analysis is often considered exploratory rather than definitive due to the highly noisy nature of the data, and so biologists are largely interested in using the data to find out which genes, gene combinations and biological processes are most relevant to the dataset. As well as delivering high classification accuracies, ArrayGenius includes novel methods for listing important genes and gene-combinations. The *most important features* reporting function, which highlights individual features common among hundreds or thousands of supervised categorization models, is unique to ArrayGenius, and provides deeper insight when compared with rudimentary analytical methods such as tabulation or clustering of most highly expressed genes.

When microarray data is loaded into ArrayGenius, the first thing it does is compare the data to its internal ontologies -- built based on the Gene Ontology (Gene Ontology Consortium, 2000) and proteomic databases (Tan and Gilbert, 2006) -- and create an enhanced feature vector for each individual whose microarray profile is represented in the dataset. The enhanced feature vector contains the original microarray data for the individual, plus extra entries indicating the degree to which various biological processes and protein families are expressed in the individual. ArrayGenius uses the enhanced dataset along with the original one, in all its subsequent analyses.

The user can make cluster dendrograms -- both the familiar kind where one sees which genes cluster together in the dataset, and a new kind, where one observes which biological processes cluster together in the dataset.

And when the dataset involves microarray data samples that can be divided into two or more categories (which may be Case vs. Control, they may represent different time points, etc.), then the user can ask ArrayGenius to learn classification models -- mathematical rules that predict whether a sample belongs to one category or another, using the gene expression values in the sample and also the inferred expression values of biological processes and

protein families in the sample. In many cases the product's categorization algorithms find extremely accurate rules -- our studies show that, after a bit of experimentation with parameter values, it generally beats the best algorithms from the academic literature. Sometimes the rules are surprisingly simple, other times they're more complex.

Because the categorization algorithms used are complex, their behavior can depend on various configurable parameters. The user has the option to set these parameters, but through the Task Wizard (to be described below) the user also has the option to bypass this process and allow the software to automatically search parameter space. The data is divided into in-sample and out-of-sample portions, and a variety of parameter values are tried on the in-sample portion; the parameter-set that led to the best in-sample results is then tested out-of-sample. This approach is heavy in its use of processor time, but undemanding on the user. For instance, for Support Vector machines the automated parameter search can choose the kernel function and its parameters. For Genetic Programming it can choose the operator set and other evolution parameters. The automated parameter search can also explore different methods of feature selection.

Generally ArrayGenius will learn a lot of classification models for a dataset, not just one or two -- and it can then study which genes, processes and families occur most often across all its models. This is a novel way of detecting which genes, processes or families are most important to the biological phenomenon being studied in the dataset. The most important features, in this sense, will often not be the ones most differentiated in expression between the categories of interest. That's because ArrayGenius is figuring out which genes, processes and families are most important, not in terms of their solitary activity, but in terms of their interactions with other genes, processes and families.

And once the user has obtained their results, they can interpret them via following hyperlinks into the Gene Ontology database, into PubMed, into various other online resources -- and into the BiomindDB, the product's own integrative data resource that provides useful functions like finding the research articles that focus on particular combinations of genes and processes.

In sum, ArrayGenius's novel approach provides a lot of information that more traditional microarray analysis doesn't:

- Ontological data integrated into the analytical process, so that classification rules and clusters involve biological processes and structures and families, not just raw gene expression values.
- A sophisticated understanding of which biological processes are important to the phenomena under analysis
- Extremely accurate classification rules, useful for diagnostics and other purposes
- New kinds of knowledge, indicating biological relationships and research directions that would

otherwise go unnoticed. A smarter way to analyze microarray data.

Specific examples of the utilization of ArrayGenius to analyze various microarray datasets are provided in (Goertzel et al, 2006a; Pennachin et al, 2006), along with comparison to results from the published literature on some of the same datasets, and in-depth commentary on the biological usefulness of the results.

The initial version of the product proved difficult for biologist users to master, due to the complexity of configuring algorithm parameters and interpreting results. However, these deficits were remedied in an upgraded version, which incorporates:

- A Task Wizard, which leads the user through a simple series of choices (phrased in nontechnical language) and then automatically launches machine learning tasks based on their decisions
- Intensively hyperlinked results screens, which allow the user to explore the data analysis results in conjunction with various online databases as well as the internal BiomindDB

Biomind GeneGenius

The general approach exemplified in the ArrayGenius product is not intrinsically restricted to microarray data; in principle it may be profitably applied to many different types of biological data. Following our work with microarrays we have extended the approach to SNP data, in the form of the GeneGenius product. The collaborative use of GeneGenius by CDC biologists and Biomind LLC's technical staff has led to some interesting discoveries, including the first concrete evidence in favor of a genetic basis for Chronic Fatigue Syndrome, as reported in (Goertzel et al, 2006). The Chronic Fatigue Syndrome application also provided an interesting opportunity to study microarray and SNP data derived from the same subjects, revealing for example that genes related to glucocorticoid metabolism emerges as "important features" from the classification model ensembles learned for both types of data.

Delivery Mechanisms and Business Models

It has not proved viable to create a "one size fits all" delivery mechanism for ArrayGenius or GeneGenius. The business and technical arrangements have been different in different cases. This reflects the social and technical complexities of the bio-IT market.

Three different delivery mechanisms have been used for the ArrayGenius and GeneGenius software:

- Biomind OnDemand, in which users upload their data to the ondemand.biomind.com website, and run analysis over the Web
- Managed hosting, which works similarly to Biomind OnDemand, except that a particular customer gets a dedicated server hosted by Biomind, thus providing additional data security and guaranteed processor time
- On-site installation of the product, which is ideal for customers who have restrictive data security policies, or who wish to integrate the Biomind software tightly with their other software systems

In the managed hosting and on-site installation approaches, customers pay an annual subscription fee followed by an annual renewal fee. In the OnDemand scenario, on the other hand, customers buy a certain number of processing units, each of which buys them a certain amount of processing time.

So far there have been two major customers for the ArrayGenius/GeneGenius product line: the Centers for Disease Control and Prevention (through a direct purchase order, coupled with a series of consultation contracts), and the National Institutes of Health (indirectly, via a contract between the NIH and Northrop-Grumman AI, on which Biomind LLC is a subcontractor). There have also been a handful of minor customers, all universities. Different customers have made different choices regarding delivery mechanism and business arrangement.

The CDC purchased ArrayGenius under a "managed hosting" plan, in which Biomind LLC maintained a dedicated server for CDC use, and created accounts for CDC research staff. Simultaneously they retained Biomind LLC as consultants, to provide ongoing assistance in using the software, and also to run separate analyses in parallel with the analyses run by the CDC staff. Biomind staff visited the CDC on-site to provide tutoring in product usage.

On the other hand, the university customers, having fewer datasets to process each year, have preferred to purchase units of time on the Biomind OnDemand server.

Finally, the NIH has pursued a different arrangement, involving a combination of a product purchase (for an on-site installation of ArrayGenius and GeneGenius) with a consulting contract with Biomind LLC. Northrop-Grumman IT is leading a multi-company-and-university team in the NIH-funded creation of a Web portal called ImmPort, a data repository and data analysis and visualization destination for NIH-funded immunologists. ArrayGenius and GeneGenius form part of ImmPort's analytical functionality.

The original arrangement was that the Biomind software systems would be installed at Northrop-Grumman along with the ImmPort servers, and would serve as part of the ImmPort back-end. Thus, although from a Biomind point of view this was an on-site installation, from an end-user point of view the Biomind software was still to be hosted remotely (rather than on-site at the universities and labs of the biologist end-users). In the course of the evolution of the ImmPort project, however, this original plan was modified, and it was found architecturally inconvenient to utilize ArrayGenius and GeneGenius as separate software processes from the rest of the ImmPort server. So, what ultimately occurred was that the ArrayGenius and GeneGenius products were broken down into software components and integrated with the ImmPort server piece-by-piece.

The Development, Deployment and Maintenance Process

Developing, deploying and maintaining complex software is never easy, but the ArrayGenius and GeneGenius products presented particular challenges because of the presence of two difficult aspects beyond those present in the generic large-scale, server-side software project: artificial intelligence and bioinformatics. Each stage of the product lifecycle formed a multidimensional learning experience for all concerned.

Design and Development Challenges

The largest challenge faced during the development phase was the fact that the needed expertise to create the core aspects of the products did not exist in any one member of the development team. Finding individuals expert in any two of the three areas {large-scale software engineering, machine learning, postgenomic bioinformatics} is difficult enough, and finding individuals expert in all three areas is more so. Fortunately the individual managing the development process did have expertise in all three areas to a considerable degree, but he was dividing his time between this and several other projects. The diversity of expertise required meant that a higher degree of cooperation between team members was required than in most software engineering projects.

Dealing with the various biological databases required within the product also proved a particular challenge, as these databases are often of highly variable quality internally, and interpreting them well enough to iron out their various irregularities requires considerable expertise both in databases and in biology.

Another set of challenges, on the product-design side, was incurred by the use of relatively complex AI algorithms rather than the simpler statistical algorithms more typical in bioinformatics. The complexity of these AI algorithms means that there are many possible ways to configure any given analytical task, and many possible ways to interpret the results. So there are tradeoffs between simplicity and flexibility. The AI-oriented members of the development team were accustomed to working with data-analysis tools on the command line, a modality which affords a great variety of options for specifying algorithm parameters, doing data preprocessing and visualizing and analyzing the results. On the other hand, the biologically-oriented members of the team were

more interested in maximizing the simplicity of the process of launching and interpreting analysis tasks.

On the task-launching side, this conflict was ultimately resolved by providing two different ways to carry out analysis within the product: the Task Wizard which launches "meta-tasks" that require very little user configuration, and then the capability to launch individual analysis tasks via specifying their detailed parameters. On the results-presentation side, a small set of visualization tools were provided within the product, along with the capability to export data in spreadsheet form for external analysis and visualization. Creating robust data visualization tools within the products was found to require more software engineering time than could be afforded, in part because doing complex visualization in a Web context can be difficult.

Deployment and Maintenance Challenges

Deployment of the product has proved relatively unproblematic. Maintenance on the other hand has proved subtler due to the dependence of the product on various biological databases. It was initially hoped that the updating of the product's internal databases, based on changes in relevant external databases, could be fully or largely automated. However, this proved overoptimistic, and as it stands periodic updates to the products internal databases must be done by hand, as external databases evolve. Currently this is done on a quarterly basis.

Combining Product Sales with Consulting Services

While the technical and product-design challenges involved in creating ArrayGenius and GeneGenius were considerable, probably the subtlest issues confronted during the product life-cycle have been related to the "business model" for the products – specifically related to the balancing of product sales and consulting in the bioinformatics arena.

At one point, after a difficult sales meeting, a Biomind LLC insider made the observation that "Bioinformatics customers fall into two groups: those who don't understand what machine learning is or why it would have any value, and those who understand it so well they want to code everything themselves." This is an overexaggeration but has a germ of truth to it. In reality, ArrayGenius and GeneGenius customers have been drawn from the categories of

- Customers who do understand the value of machine learning approaches, in general, but lack the expertise to implement them
- Customers who have the know-how to implement machine learning algorithms and have done so, but recognize that the ArrayGenius and GeneGenius products embody particularly

sophisticated approaches that would be very time-consuming to successfully emulate.

What has been interesting to see (and was not originally anticipated) is that in both of these customer categories, the ArrayGenius and GeneGenius products have generaly been assessed in terms of their synergy with the expertise of the Biomind LLC scientific/technical team. That is, customers have not viewed themselves as merely buying a product, but rather as buying a product together with the active collaboration of the individuals who developed the product and possess expertise in using it. Both of the major product customers, so far, have also been major consulting-services customers; and the majority of minor product customers have also been minor consulting customers.

This interweaving of product sales and consultation appears to be a natural consequence of the complexity of the AI algorithms underlying the product, combined with the relative lack of experience of the end users (biologists) with such algorithms. While it is easy for inexperienced end users to load data into the products and run analyses, and relatively simple for them to do basic results interpretation (for instance, delving into the literature to explore the importance of genes tagged as "important"), it is inarguable that the Biomind team can carry out results interpretation in a more sophisticated and insightful way than nearly any end users.

In time, we expect that AI methods will come to pervade the bioinformatics industry, so that "bio-AI experts" will become as pervasive as biostatisticians are today. At that point, the coupling of product sales with consulting services will be largely obsolete, as sophisticated interpretation of results will be carried out by in-house bio-AI experts on the customer side. But now this pervasion is still in its early stages, and so we expect that the habitual combination of product sales and consulting will exist for some time to come.

Conclusion

We have described the Biomind ArrayGenius and GeneGenius products, which allow biologists to use the Web to carry out machine learning based analyses of microarray and SNP data. We have also reviewed many of the challenges that arose during the course of designing, developing and maintaining the products; and the combination of product sales and consulting services that emerged over time as the de facto means of product delivery. As noted, these products have presented particular complexities from a variety of standpoints, due to their combination of large-scale, server-side software engineering, advanced and in some respects novel AI technology, and use of bioinformatics concepts and biological databases. However, they have been successfully delivered to customers, and they have been utilized to arrive at a number of interesting research conclusions, some published in the scientific literature, and

some which would not have been arrived at had the products not existed.

Many of the lessons learned in the course of creating, deploying, maintaining and selling the ArrayGenius and GeneGenius products are applicable more generally. Certainly there would be many similar aspects involved in the creation of compute-power-intensive, AI-based data analysis products for any highly specialized domain area.

At a very high level, the one lesson that emerges most clearly from our experience with these products is the power of, and the necessity of, intensive collaboration between individuals with different backgrounds. Within the development team, close collaboration of individuals with software engineering, AI and biology backgrounds was key. And, after delivery of the first product version to the first customer (the CDC), close collaboration between developers and customers became key. Many of the current product features came into existence due to requests from customers; and, the internal architecture of the next version of the product will be completely different from previous versions, due in part to the impetus the NIH, a current customer, has given us to partition the software into smaller, more autonomous components. In the end, although the ArrayGenius and GeneGenius products began with an original and somewhat iconoclastic vision, what they have evolved into is the result of a collective cognitive process spanning many individuals with different backgrounds, affiliated with different institutions, and bearing different relationships to the products.

References

Cristianini, N. and J. Shaw-Taylor. 2000. Support Vector Machines, Cambridge University Press.

Gene Ontology Consortium. 2000. Nat. Genet. 25, 25-29.

Goertzel, Ben, Cassio Pennachin, Lucio Coelho, Brian Gurbaxani, Elizabeth B. Maloney, James F. Jones. 2006. Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes are predictive of chronic fatigue syndrome, Pharmacogenomics

Goertzel, Ben, Cassio Pennachin, Lucio de Souza Coelho, Elizabeth B. Maloney, James F. Jones, Brian Gurbaxani. 2006. Allostatic Load is Associated with Symptoms in CFS Patients, Pharmacogenomics

Goertzel, Ben, Lucio Coelho, Cassio Pennachin and Mauricio Mudada. 2006. Identifying Complex Biological Interactions based on Categorical Gene Expression Data. Proceedings of Conference on Evolutionary Computing 2006, Vancouver CA

Kothapalli, R., S. J. Yoder, S.Mane and T.P. Loughran, 2002. *BMC Bioinformatics* 3(1), 22

Koza, John. 1992. Genetic Programming. MIT Press.

Lyons-Weiler, J. 2003. *Applied Bioinformatics* 2(4), 193-195

Pennachin, Cassio, Ben Goertzel Lucio Coelho, Izabela Freire Goertzel, Murilo Queiroz, Francisco Prosdocimi, Francisco Lobo. 2006. Learning Comprehensible Classification Rules from Gene Expression Data Using Genetic Programming and Biological Ontologies, Proceedings of CIBB 2006, Genova, Italy

Rockett, J. C. 2003. *Drug Discovery Today* 8(8), 343.

Simon, R., M.D. Radmacher, K. Dobbin, L.M. McShane, 2003. *J. Nat. Cancer. Inst.* 95(1), 14-18.

Tan, A. and D. Gilbert. 2006. *Appl. Bioinformatics* 2, S75-S83 (2003).