Incorporating microarray analysis into a visual programming language

Andrew C. Whittam

The microarray has become the standard way to obtain a holistic view of gene expression in a cell. The tools needed to interpret the data produced by a microarray are scattered throughout the web. Along with the problems of finding programs, issues such as compatibility and usability bar the way. An online program that combines multiple microarray tests and genomic databases into a user friendly but flexible environment with capabilities for the novice to advanced user would reduce the work of the interested scientist. The incorporating of analysis tools in BioBike has allowed for a single program for scientists to use to analyze and interpret their microarray data.

Introduction

With the increase in genomic sequencing and popularity of microarrays the biological sciences are traveling ever further into an age of massive datasets. The complete genomes of more than 100 organisms had been sequenced in 2001. 6 years later 200 genomes were sequenced in 2007 alone. In parallel with the increase in mapping of genes has been the rise of the microarray, an experiment that allows the measuring of gene expression. The microarray has become the standard way to obtain a holistic view of gene expression in a cell. Scientists who utilize this resource face the problem of too much data and too little usable information. Through the advancement in technology has the ability to collect and interpret the massive amounts of data produced by microarrays become possible. Many tools and procedures have been developed to address this issue of extracting information from large amounts of data. For example in “Significant Analysis of Microarrays” Tusher, Tibshirani and Chu describe a technique for determining significant differences in individual gene expression when faced with few repetitions commonly found in microarray experiments. With some of these theories and algorithms there are computer programs to download and install but along with these comes all the baggage of a standalone offline application. There are such issues as finding the program (if it exist), compatibility, documentation, formatting, and simple knowledge of the program. After publishing a paper describing a technique to cluster or group genes with similar expression, the Elsien lab created a computer program for download that performed the clustering. Through the limiting user input and privileges some of these applications are simple to use but lack the ability to be tailored to address the specific scientist’s needs. On the other hand other applications give the user a multitude of choices and variables to tailor their analysis but lose usability and simplicity in the process. Still after results are computed by these programs additional analysis may have to be done to truly understand the results and their context within the genome. The combination of these tools into a single program would prove useful to scientist. The framework for such a program already exists in BioBike, a visual programming language that links complete genomes with customized functions and tests. Recently the BioBike program advanced forward with the addition of microarray data to the tools. Incorporating microarray analysis into this visual programming language would consolidate some of the steps for microarray interpretations.

Materials and Methods

We decided to use Significant Analysis of microarrays (SAM) and Cluster as the two microarray functions we would incorporate into BioBike. Both are well established procedures that have papers published describing their methods.

SAM

SAM is used to distinguish between significant variations of a gene expression due to a tested environment with that of
random change. SAM was first described in the 2001 paper “Significance analysis of microarrays applied to the ionizing radiation response” by Tusher, Tibshirani and Chu. SAM uses a modified t-test to rank individual gene expression across repetitions after which it compares that list to gene scores calculated on permutations of the microarray data. The idea is that if genes are significantly changing they should deviate from the randomly generated scores. The SAM function produces a list of genes, both up and downregulated that have been ranked significant.

Cluster
Clustering of microarray data was described in “Cluster analysis and display of genome-wide expression patterns” by Eisen, Spellman, Brown, and Botstein in 1998. Clustering is the use of algorithms to organize genes according to similarity in pattern of gene expression. Using the data as vectors a distance formula is used to determine similarity. Cluster outputs a chart mapping related genes into nodes. From this researchers can link genes that potentially may have similar regulation factors.

LISP
Since BioBike is programmed in the computer language LISP and Biolisp (a sublanguage of LISP) we chose to write the microarray analysis programs in Lisp as well. Tailoring a program specifically to BioBike allowed for faster results along with less software complications. After writing the programs in LISP the code was incorporated into the BioBike framework to produce graphical representations for the user. SAM would return a list of genes that were deemed significantly up or down regulated. Cluster will group genes and return the results in a graphical cluster chart.

Discussion
The solution of an online program that combines multiple microarray tests and genomic databases into a user friendly but flexible environment with capabilities for the novice to advanced user allows reduces the steps of microarray analysis. The implementing of these functions into the BioBike has allowed the investigator to focus less on the collecting, understanding and execution of the tools and more on the scientific investigation. Allowing the user the amount of freedom he or she desires with the option to edited function keeps accessibility and simplicity in the experiment. The environment also promotes the combination of experiments and functions. The freedom of the scientist to explore different microarray tests and genomic functions in concert promotes an atmosphere of creativity and innovation. The conalidation of steps of microarray analysis into one program reduces the work of assembling and using the tools needed.

References


Results

Fig. 1
Screenshot of SAM analysis in BioBike language
Fig. 2
Screenshot of Cluster analysis in BioBike language

Fig. 3
Screenshot of SAM analysis results after execution in BioBike language
Fig 4
Screenshot of Cluster analysis results after execution in BioBike language