Incorporating microarray analysis into a visual programming language

Andrew C. Whittam

The mircoarray 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 analysis 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[4]. In parallel with the increase in mapping of genes has been the rise of the mircoarray, 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 micorarray analysis into this visual programming language would consolidate some of the steps for microarray interpretations.

Materials and Methods

We decided to use Significant Anaylsis 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

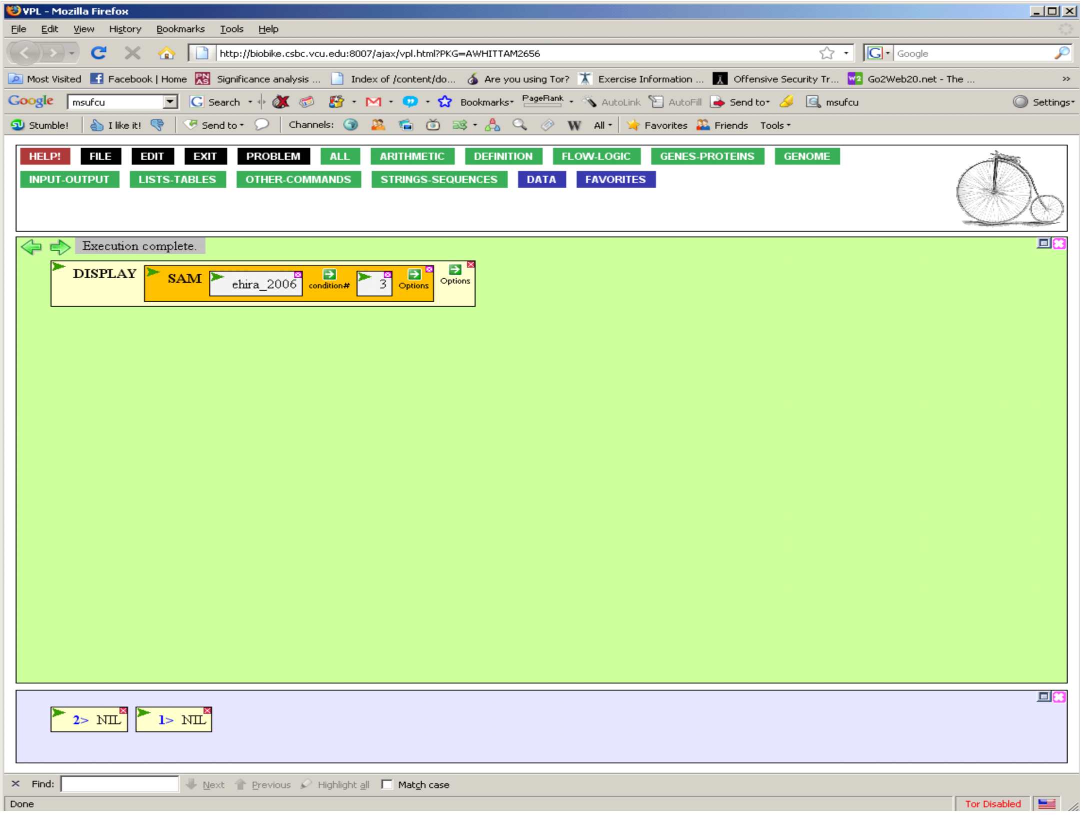
[1] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein
Cluster analysis and display of genome-wide expression patterns
PNAS, Dec 1998; 95: 14863 - 14868.

[2] Significance analysis of microarrays applied to the ionizing radiation response
Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu
Proceedings of the National Academy of Sciences, U.S.A. 98:5116-5121.

[3] http://ramsites.net/~biobike/ - BioBike Website

[4] Liolios K, Mavrommatis K, Tavernarakis N, Kyrpides, NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata.   NAR 36, D475-D479

Results



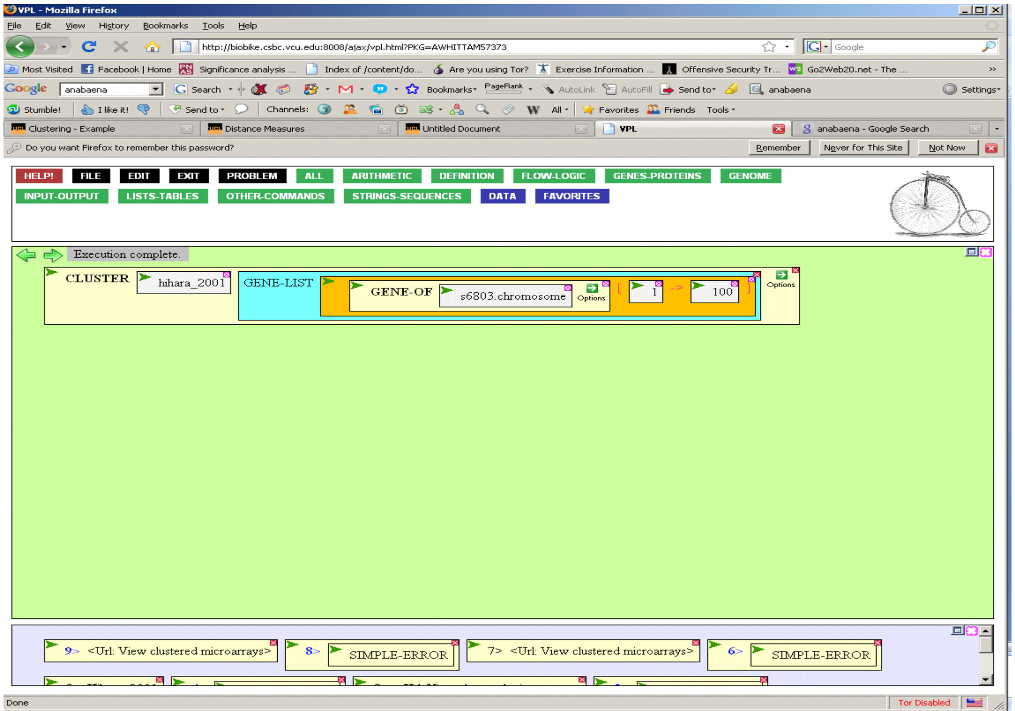
VPL - Mozilla Firefox
http://biobike.csbc.vcu.edu:8007/ajax/vpl.html?PKG=AWHITTAM2656
HELP! FILE EDIT EXIT PROBLEM ALL ARITHMETIC DEFINITION FLOW-LOGIC GENES-PROTEINS GENOME
INPUT-OUTPUT LISTS-TABLES OTHER-COMMANDS STRINGS-SEQUENCES DATA FAVORITES
Done
Tor Disabled

Execution complete.

```
DISPLAY  SAM  ehira_2006  condition#  3  Options  Options
```

```
2> NIL    1> NIL
```

Fig. 1
Screenshot of SAM analysis in BioBike language



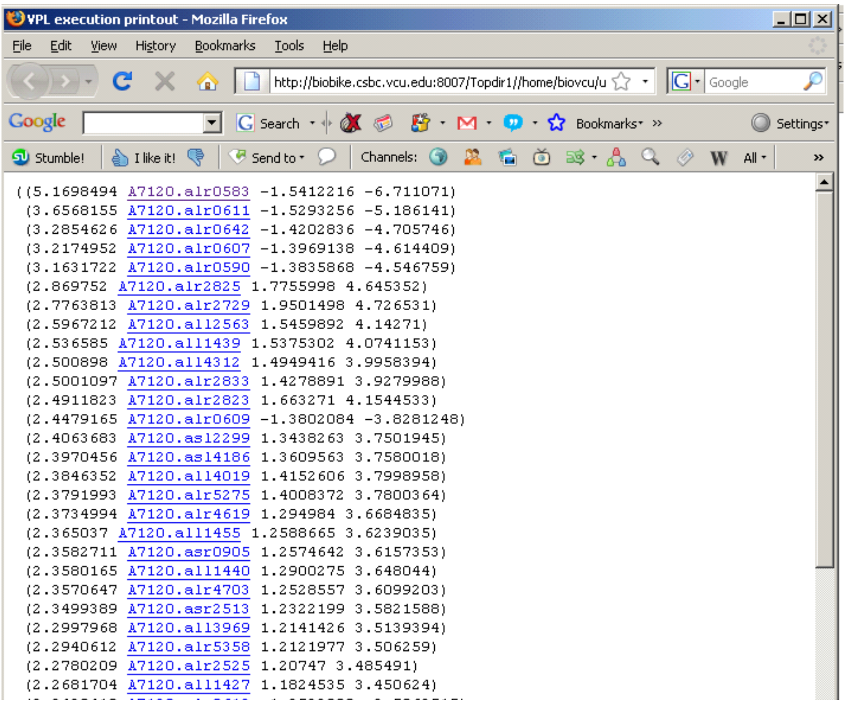
VPL - Mozilla Firefox
http://biobike.csbc.vcu.edu:8008/ajax/vpl.html?PKG=AWHITTAM57373
Clustering - Example
Distance Measures
Untitled Document
VPL
anabaena - Google Search
Do you want Firefox to remember this password?
Remember
Never for This Site
Not Now
HELP! FILE EDIT EXIT PROBLEM ALL ARITHMETIC DEFINITION FLOW-LOGIC GENES-PROTEINS GENOME
INPUT-OUTPUT LISTS-TABLES OTHER-COMMANDS STRINGS-SEQUENCES DATA FAVORITES
Done
Tor Disabled

Execution complete.

```
CLUSTER  hihara_2001  GENE-LIST  GENE-OF  s6803.chromosome  Options  [ 1 -> 100 ]  Options
```

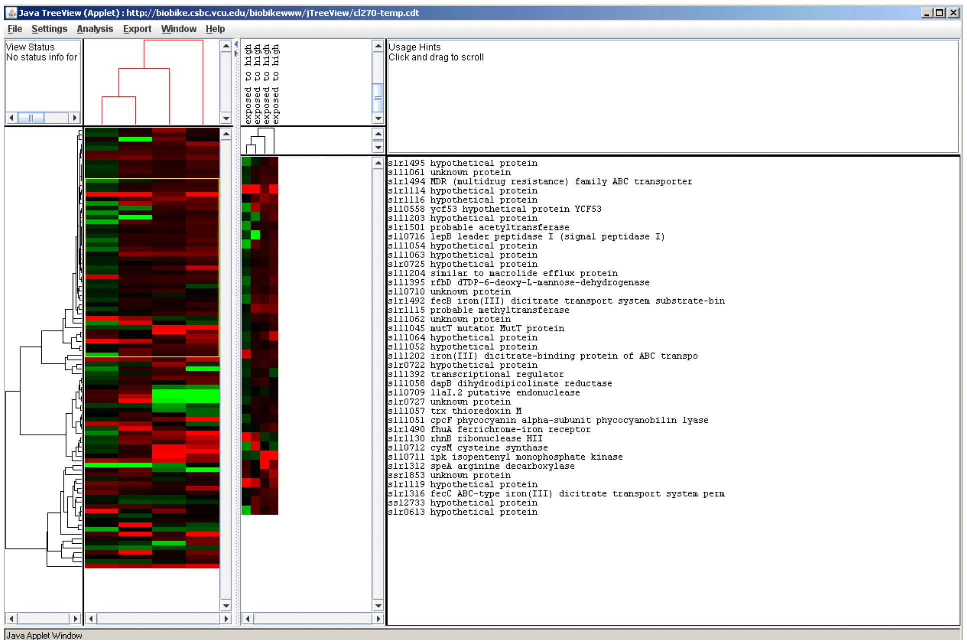
```
9> <Url: View clustered microarrays>   8> SIMPLE-ERROR   7> <Url: View clustered microarrays>   6> SIMPLE-ERROR
```

Fig. 2
Screenshot of Cluster analysis in BioBike language


VPL execution printout - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://biobike.csbc.vcu.edu:8007/Topdir1//home/biovcu/u
Google
Search
Bookmarks
Settings
Stumble!
I like it!
Send to
Channels:
All

```
((5.1698494 A7120.alr0583 -1.5412216 -6.711071)
 (3.6568155 A7120.alr0611 -1.5293256 -5.186141)
 (3.2854626 A7120.alr0642 -1.4202836 -4.705746)
 (3.2174952 A7120.alr0607 -1.3969138 -4.614409)
 (3.1631722 A7120.alr0590 -1.3835868 -4.546759)
 (2.869752 A7120.alr2825 1.7755998 4.645352)
 (2.7763813 A7120.alr2729 1.9501498 4.726531)
 (2.5967212 A7120.all2563 1.5459892 4.14271)
 (2.536585 A7120.all1439 1.5375302 4.0741153)
 (2.500898 A7120.all4312 1.4949416 3.9958394)
 (2.5001097 A7120.alr2833 1.4278891 3.9279988)
 (2.4911823 A7120.alr2823 1.663271 4.1544533)
 (2.4479165 A7120.alr0609 -1.3802084 -3.8281248)
 (2.4063683 A7120.asl2299 1.3438263 3.7501945)
 (2.3970456 A7120.asl4186 1.3609563 3.7580018)
 (2.3846352 A7120.all4019 1.4152606 3.7998958)
 (2.3791993 A7120.alr5275 1.4008372 3.7800364)
 (2.3734994 A7120.alr4619 1.294984 3.6684835)
 (2.365037 A7120.all1455 1.2588665 3.6239035)
 (2.3582711 A7120.asr0905 1.2574642 3.6157353)
 (2.3580165 A7120.all1440 1.2900275 3.648044)
 (2.3570647 A7120.alr4703 1.2528557 3.6099203)
 (2.3499389 A7120.asr2513 1.2322199 3.5821588)
 (2.2997968 A7120.all3969 1.2141426 3.5139394)
 (2.2940612 A7120.alr5358 1.2121977 3.506259)
 (2.2780209 A7120.alr2525 1.20747 3.485491)
 (2.2681704 A7120.all1427 1.1824535 3.450624)
```

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



View Status
No status info for

Usage Hints
Click and drag to scroll

```
slr1495 hypothetical protein
sll1061 unknown protein
slr1494 MDR (multidrug resistance) family ABC transporter
slr1114 hypothetical protein
slr1116 hypothetical protein
sll0558 ycf53 hypothetical protein YCF53
sll1203 hypothetical protein
slr1501 probable acetyltransferase
sll0716 lepB leader peptidase I (signal peptidase I)
sll1054 hypothetical protein
sll1063 hypothetical protein
slr0725 hypothetical protein
sll1204 similar to macrolide efflux protein
sll1395 rfbD dTDP-6-deoxy-L-mannose-dehydrogenase
sll0710 unknown protein
slr1492 fecB iron(III) dicitrate transport system substrate-bin
slr1115 probable methyltransferase
sll1062 unknown protein
sll1045 mutT mutator MutT protein
sll1064 hypothetical protein
sll1052 hypothetical protein
sll1202 iron(III) dicitrate-binding protein of ABC transpo
slr0722 hypothetical protein
sll1392 transcriptional regulator
sll1058 dapB dihydrodipicolinate reductase
sll0709 llaI.2 putative endonuclease
slr0727 unknown protein
sll1057 trx thioredoxin M
sll1051 cpcF phycocyanin alpha-subunit phycocyanobilin lyase
slr1490 fhuA ferrichrome-iron receptor
slr1130 rhnB ribonuclease HII
sll0712 cysM cysteine synthase
sll0711 ipk isopentenyl monophosphate kinase
slr1312 speA arginine decarboxylase
ssr1853 unknown protein
slr1119 hypothetical protein
slr1316 fecC ABC-type iron(III) dicitrate transport system perm
ssl2733 hypothetical protein
slr0613 hypothetical protein
```

Fig 4

Screenshot of Cluster analysis results after execution in BioBike language