An Overview of Projects in Computing for Genomic Analysis
Biocore Technical Report 1

Bioconductor Core

November 30, 2002

Accompanying the proliferation of genomic data have been many different software tools and initiatives. We will consider a representative sample of these initiatives and demonstrate the likely conflicts and interactions with these different groups. We have no intention of reimplementing good procedures; rather, we will use the flexibility of R's design and interoperability to access them directly (if possible) using only a relatively small amount of interface code.

1 Open-Bio Initiatives

The Open Bioinformatics Foundation supports a set of related open source projects for various programming languages. These include *bioperl*, *biopython*, *biocorba*. These projects are loosely related developers that contribute source code for routines of interest to them to a common repository. There are good tools for carrying out sequence analysis but few if any for analyzing more quantitative data that arise from DNA microarray or protein array technologies.

R differs from these languages in that it is interactive and provides a rich repository of well-designed and tested numerical and statistical algorithms. By providing well-defined interfaces (via class definitions) we make it much easier for the developers in these different projects to access and utilize these capabilities in R.

The R environment has long relied on interoperability to access software written in other languages. A substantial effort has been made by Dr. Temple Lang (through the Omegahat initiative, `www.omegahat.org`) to provide interoperability between the S language (of which R is one dialect) and languages such as Perl, Python, and Java. It should be noted that Dr. Temple Lang's initiatives are generally bidirectional. They are directed at both providing services to programmers normally resident in R and to programmers normally resident in the other languages, thereby making many of our developments available to other groups such as *bioperl*.

2 Commercial Systems

There are a variety of commercial systems available. GeneSpring, Genesight, and Affymetrix GeneChip Software Solutions appear to be the most popular. GeneSpring is available for Microsoft Windows, Mac OS, and most UNIX systems. Genesight is only available for Microsoft Windows systems. Affymetrix GeneChip Software Solutions combines the Microarray Suite (MAS) software, the Laboratory Information Management System (LIMS), the MicroDB Software, and the Data Mining Tool (DMT). They are all designed to work with the Affymetrix GeneChip technology (1) and are only available for Windows.

One of these commercial systems, Iobion's GeneTrafic `www.iobian.com`, utilizes Bioconductor tools. The Chief Scientific Officer of Iobion Informatics recently presented a talk in Washington, D.C. from September 10-13, 2002 during the CHI Second Annual Microarray Data Analysis and Data Visualization Conference one Advancements in Affymetrix Probe-Level Data Analysis. See `www.iobion.com/presentations/presentations.html`. In this talk one of the Bioconductor tools was featured.

3 Academic Offerings

Another general class of systems and tools are those that are free only for academic or non-commercial purposes. These tools are generally developed at non-profit academic or governmental institutions. They are not open source. Examples include GeneCluster, a stand alone Java system from the Whitehead Institute Center for Genome Research (5). It provides the ability to filter and pre-process data in a variety of ways, cluster expression profiles using the SOM algorithm, and view the results. Installation support is available for Microsoft Windows and Macintosh.

TIGR's microarray tools `www.tigr.org/software/` are an example. TIGR Microarray Data Analysis System (MIDAS) is a microarray data quality filtering and normalization tool that allows raw experimental data to be processed through various data normalizations, filters, and transformations via a user-designed analysis pipeline. Currently implemented normalization and data analysis algorithms include total-intensity normalization, Lowess (Locfit) normalization, flip-dye consistency checking, replicates analysis, intensity-dependent z-score filtering (slice analysis), etc. MIDAS is implemented by Java language and thus a platform-independent application. It requires JDK v1.3 or higher. Refer to the included manual for details. Users may license the program or the entire source code.

[MADAM] Microarray experiments produce large amounts of data for even the simplest of experiments. In order to analyze data from many experiments that data must be stored in an accessible form, such as in a database. MADAM (MicroArray DAta Manager) is a java-based application designed to load and retrieve microarray data to and from a database (also supplied with the software). MADAM provides data entry forms, data report forms and additional applications necessary to maintain microarray data for further analysis. Madam requires JRE 1.3.1. Users may license the program or the entire source code. [MultiExperimentViewer] TIGR MultiExperiment Viewer (MEV) is a Java application designed to allow the analysis of microarray data to identify patterns of gene expression and differentially expressed genes. Numerous normalization, clustering and distance algorithms have been implemented, along with a variety of graphical displays to best present the results. MEV was written to be flexible and expandable, and supports a variety of input and output formats. MEV requires version 1.2 or higher of Sun's JRE and J3D package. Users may license the program or the entire source code. [TIGR ArrayViewer] A software tool designed to facilitate the presentation and analysis of microarray expression data, leading to the identification of genes that are differentially expressed. ArrayViewer is written in Java for cross-platform compatibility and reads and writes data using flat files or a database through stored procedures, See the ArrayViewer Overview as a Adobe Acrobat PDF File. Machines that lack the requirements for the MultiExperiment Viewer may use ArrayViewer for single experiment analysis. [TIGR SpotFinder] TIGR Spotfinder is a software tool designed for Microarray image processing using the TIFF image files generated by most microarray scanners. TIGR Spotfinder was written in C/C++ for PCs running Windows.

Gary Churchill's Statistical Genetics Group Software is based on R. We are working to incorporate it into Bioconductor. This group has software for QTL mapping (Marker regression programs, Pseudomarker programs, R/qtl programs) and Gene expression (ANOVA programs for microarray data (Matlab version, R/maanova 0.8-1)

DNA-Chip Analyzer, or dChip, `www.dchip.org`, (4) is software for analyzing Affymetrix data. It provides an alternative for computing expression level data plus various clustering and interactive tools for exploring expression level data. It is only available for Windows.

A tool for the analysis of cDNA microarrays is provided at `biosun1.harvard.edu/~ctseng/download.html`. It implements the methods described in (6) and is only available for Windows.

The Significance Analysis of Microarrays (SAM) package `www-stat.stanford.edu/~tibs/SAM`, (7), implements a multiple testing procedure which can be used for both cDNA and oligo microarray data, protein expression data, and SNP chip data. It provides analysis tools based on statistical inference.

Cluster, TreeView, and GMEP, developed by Michael B. Eisen and collaborators, are an integrated set of programs for analyzing and visualizing the results of complex microarray experiments (`rana.lbl.gov`, (3), see the Appendix for a copy of Dr. Eisen's paper). Cluster performs a variety of types of cluster analysis and processing on large microarray datasets, including: hierarchical clustering, self-organizing maps (SOMs), k-means clustering, and principal component analysis. TreeView permits graphically browsing the output from Cluster. Both tree-based and

image based browsing of hierarchical trees can be performed. The programs are only available for Windows.

Combined Expression Data and Sequence Analysis (GMEP) (2) is a software package for computing genome-mean expression profiles from expression and sequence data.

These products are very easy to use and provide a valuable resource for end-users. However, the associated development cost is usually high. The software that we are proposing to create will drastically reduce the development time for prototype modules (and deliverable code in the R system). Once an idea has been tested and deemed useful, stand-alone versions with simplified user interfaces can be created and deployed. This project will provide researchers with better tools for testing and creating novel analyses.

Finally, we consider the GeneX and CyberT projects (`www.ncgr.org/genex`). These projects aim to provide an Internet-available repository of gene expression data with an integrated tool set that will enable researchers to analyze their data and compare their results with other such data. We have already contacted these groups to identify potential points of conflict. There were in fact none, and they are very interested in adopting methodology coming out of this project. By requiring developers to adhere to a certain standard we allow service providers such as GeneX and CyberT to adopt tools much more easily.

If there are standard classes and methods for handling specific types of genomic data the service providers need only develop tools to transform data from their storage format (usually some form of relational database is used) into the class structure needed. Thus one set of extraction routines can suffice for multiple analysis packages. Standardization provides enormous benefits to all concerned.

In addition to software development there are a number of initiatives for standardization of data formats and other things. These include, MIAME and MAGEML. We are closely tracking these developments and are in a position to adopt (and provide interface routines for) these standards as they are more widely adopted. Many of them are rapidly developing and it seems prudent not to expend too many resources on formats that are likely to undergo substantial revision.

References

[1] Affymetrix. *Affymetrix Microarray Suite User Guide*. Affymetrix, Santa Clara, CA, version 5 edition, 2001. 1

[2] D. Y. Chiang, P. O. Brown, and M. Eisen. Visualizing associations between genome sequence and gene expression data using genome-mean expression profiles. *Bioinformatics*, 17:S49–S55, 2001. 3

[3] M. B. Eisen, P. T. Spellman, P. O. Brownand, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA*, 95(25):14863–14868, 1998. 2

[4] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science USA*, 98:31–36, 2001. 2

[5] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, , and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps. *Proceedings of the National Academy of Science USA*, 96:2907–2912, 1999. 2

[6] G. C. Tseng, M.-K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong. Issues in cdna microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects,. *Nucleic Acids Res*, 29:2549, 2001. 2

[7] V. Tusher, R. Tibshirani, and C Chu. Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences USA*, 98:5116–5121, 2001. 2