2002 Annual Report for the Bioconductor Project, DFCI

R. Gentleman

May 29, 2008

1 Introduction

The Bioconductor project (`http://www.bioconductor.org`) was initiated in 2001 at the Dana Farber Cancer Institute. Institutional funding from the High Tech Industry Multidisciplinary Research Fund was used to provide programming support for the project.

1.1 Broad Goals

The broad goals of the Bioconductor project are to enable good data-analytic and inferential practice in computational biology and to provide a platform that allows scientists (both biologists and statisticians) to develop and rapidly deploy new innovative computational methods. The mechanisms used to carry out these goals include:

- providing statisticians with tools in a programming environment that they are comfortable with,
- providing biologists with simplified access to the necessary tools (both existing and those yet to be developed) needed for computation,
- a dedication to the development of high–quality object–oriented software with commitments to encapsulation, extensibility and documentation.

Computational biology is a wide and diverse field. There is currently no consensus view on which tools will be important or on how to develop systems that take advantage of all relevant data. However, it is clear that there will be a need for more non–numerical computing and for computing/algorithms on non–standard data structures (such as graphs). Thus, it is important to build computational infrastructure that supports the construction of tools rather than designing or implementing specific tools. We feel that our emphasis should be more on the development of tools to build applications rather than on the applications themselves. Particular analyses may come and go, but the need for new applications based on new data or amalgamations of data will always exist.

1.2 A short history

The project began in 2001. In September the first programmer joined the team and in January of 2002 the second programmer joined the team. The project has been underway for approximately 11 months at the time of the writing of this document. I am extremely surprised and pleased with the progress made to date. It is important to realize that none of the principals had much experience in these areas prior to starting this program and hence much of the first year has been a learning experience. That said, the Bioconductor project is now widely known and widely used and is seen as a leader in the development of technology and methodology for analysing genomic data.

We have assembled a wide and diverse team of statisticians, programmers and biologists. That team has produced a large amount of software (approximately 5 times the quantity that would be expected according to most commercial software development estimates). These components are now being used widely (they have been ported by commercial companies) and are sufficiently mature to form the building blocks of our next steps forward. Additionally the main participants are starting to produce papers describing the tools that we have developed. There is always a trade off in this field between the need to develop software and the need to write papers describing that software. In the early stages of the project I felt that is was more important to write software and to sacrifice publication in favor of tool development. This emphasis has already started to shift to a more sustainable situation.

2 Specific Achievements

There have been 15 packages developed and included in Release 1.0, May 1, 2002. These packages and many that are still being developed are available under the GPL (or LGPL) and can be freely downloaded from the Bioconductor web site. The focus of these initial offerings is the analyses of microarray experiments. However, it should be noted that the knowledge gained is more broadly applicable. We anticipate a regular 6 month release schedule.

We have developed and distributed technology that uses sound design principles to be applied to issues of assembling biological annotation data in a format suitable for analyses.

One of our basic principles is to get these tools into the hands of biologists. This can be achieved through a number of different mechanisms.

- Education in the use of these software via short courses and other training sessions: There have been 2 courses presented and 2 more planned for 2002. Additional special sessions are planned for the fall at the DFCI.

- Distribution of specific and detailed *HowTo* documents that contain viable code and explain how to perform specific tasks: We have pioneered the use of these tools developed as part of the R project and now have approximately 25 such documents. We cannot hope to write all necessary documents of this form and hence have developed tools that will allow others to contribute.

- Development of widgets (simplified graphical user interfaces) that allow for *point–and–click* access to many different analyses: This is progressing well but is a very large programming job (usually 1.5 to 2 times as large as developing the underlying algorithms).
- Development of browser–based methods: While this area has the most promise it is the one area where we have had the least success to date. We have adopted a new technology (web–services) that we believe will be successful and plan to have some prototype tools developed by the end of 2002.

2.1 Talks on Bioconductor

Presentations by Rafael Irizarry are at `http://www.biostat.jhsph.edu/~ririzarr/talks.html`
Presentations by Robert Gentleman are at `http://biosun1.harvard.edu/~rgentlem/HTML/talks.html`
Presentations by Sandrine Dudoit are at `http://stat-www.berkeley.edu/users/sandrine/presentations.html`
Presentations by Vincent Carey are at `http://www.biostat.harvard.edu/~carey/bioinf.html`

2.2 Newsletters

Bioconductor projects and their authors have been major contributors to the R News Letter, http://www.cran.r-project.org/doc/Rnews/. The March 2002 issue is primarily about Bioinformatics. In the June 2002 issue there are substantial Bioconductor contributions as well.

2.3 Papers

All papers published by members of the Bioconductor project, using Bioconductor developed resources are listed on a separate page. Some (but not all) papers submitted for publication are also listed.

2.4 Ph. D. Students

One student graduated this year from the department of Biostatistics, M. Tadesse, where a portion of her Ph.D. dissertation was based on technology and ideas generated by the Bioconductor project. There are currently two other Ph.D. students (both under the supervision of R. Gentleman), D. Scholtens and B. Ding whose Ph.D. thesis will be based on technology and ideas generated by the Bioconductor project.

Ph. D. students at both U.C. Berkeley and Johns Hopkins have been involved in the development of packages. The senior members of the project have helped them to develop programs and programming styles that will be beneficial to future developments.

2.5 Grants

One large grant was submitted in March to the NIH under the BISTI program. Amount requested in direct costs is approximately 500,000.00. Other research grants will be applied for in connection with this project during the coming year. In particular support for theoretical and computational models for dealing with data on graphs is an intended focus of research.

2.6 Website activity

The Bioconductor website was developed and initiated in early 2002. We have monitored activity on that site. While it is not possible to estimate numbers of user etc. we have the following information available. Since May 1, 2002 we have had 5689 unique IP blocks hit bioconductor.org (these are defined as unique values of the first 9 digits in the IP address). Under this last condition the Biostatistics department at the DFCI would count as one hit. There were 12,842 *visits* to downloadable pages (defined as a unique IP accessing pages w/ no more then a 30 min pause between hits) and 130,505 files downloaded (defined as downloadable files & wiki) the total amount downloaded was 20,749,469 KB.

3 Future Plans and Developments

Through a very generous donation organized by Dr. J. D. Iglehart we have obtained funding to provide the DFCI with an extremely valuable computational resource. We purchased 16 dual processor pentium computers and sufficient disk storage space to handle many (but not all) different computational problems that arise in computational biology. This system should be up and running very shortly and we will begin using it to perform analyses and computational experiments. We will make the facility broadly available to other researchers within the DFCI.

In the next year our plans are as follows:

- Attack the problem of supporting biostatistical analysis on large data resources contained in data warehouses. Primary development requirement: effective and secure methods for allowing S language statistical analysis functions to operate with external references to data objects, to avoid copying very large quantities of data prior to analysis.

- Continue with widget development to provide encapsulated simple user interfaces to packaged analysis sequences (file browser to identify input data, function browser to identify analysis desired, parameter selection buttons, result browser and serializer).

- Development of tools to enable the creation of transparent web based applications (see following section for an example).

- Development of tools for creation and visualization of network models. A newly established collaboration with researchers at AT&T and Lucent Technologies,

the creators of GraphViz `http://www.research.att.com/sw/tools/graphviz`, will greatly aid in the development of these tools.

- Creation of tools for improving transparency and efficiency of various analytical tasks in molecular biology and genomics, including improved image analysis for blot and gel data (*gelTools* package sketch at Vincent Carey's web site), improving R resources for sequence analysis (incorporation of BLAST family of algorithms, methods for SNP annotation), and development of disease-specific resources for genomic applications in e.g., HIV and obesity research.

4 Specific Example

Recently inspired by questions by Dr. K. Polyack (DFCI) we began exploring the necessary tools to provide a web–based tool for finding nearest neighbors in SAGE data. The following structure seems quite reasonable and easy to implement given all the work and tool development that has been done in the bioconductor project.

- construct a table in any database (Postgres will be easiest for us) that contains all the SAGE data. Tags form the rows, samples the columns.
- use database connections from within R to extract subsets of the data (chunking) and compute the distances to the target gene.
- return the top K genes
- use HTML, JavaScript and R to allow us to develop an interactive web page that activates R to carry out the requested analyses and return the top genes in HTML for the researcher.

It seems that such applications may be of more general interest. Similar analyses can be carried out on Affymetrix data, on cDNA data or an combined data. Thereby enabling reuse and reanalysis of existing data. We plan to develop a tool kit that would allow a moderately capable programmer to easily construct such an application in a reasonable time frame (a few days). Such applications, if designed to interact can for the basis of powerful, leading edge, computational biology tools.