2003 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.

During 2003 the group of developers grew to 22 persons from various institutions around the world. The mailing list has become very active with one or two serious discussions per day occurring. There are currently 500 subscribers to the list.

2 Specific Achievements

Release 1.2, our third release occurred in May 2003. There are 30 packages in Release 1.2 and 39 in the development arm. This will eventually become release 1.3 (scheduled for mid October 2003). Of these 5 packages are contributed from other research groups.

Most packages are released under an Open Source license while others have a free for non-commercial use license. Our focus has remained on the analyses of microarray experiments. Although much of the recent effort has been on the development of tools and methodologies for dealing with network data (such as protein-protein interactions, metabolic networks and a variety of other examples).

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.

2.1 Bioconductor in the News

The Bioconductor Project has been mentioned in news articles in Nature and the R project was mentioned in a news article in Science. We view these as positive accomplishments. Among the goals of the project were to facilitate use and to become leaders in the field and these articles indicate that we are achieving those goals.

- Nature 424, 119 (10 Jul 2003) News
- Nature 424, 610 - 612 (07 Aug 2003) News Feature

2.2 Ph. D. Students

There are currently three 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. E. Whalen has begun her Ph.D. studies and will be working on methods for network visualization.

Insightful Corporation has provided partial funding for a PhD student to HSPH Biostatistics for work in enhancing the collaboration between their work in this are and 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.

2.3 Commercialization

Many different commercial entities are developing links to R and Bioconductor. These include the Insightful Corporation `www.insightful.com` and SpotFire `www.spotfire.com`. Again, we feel that these actions indicate the increasing relevance and consumer driven demand for the sorts of innovation that the Bioconductor project is providing.

2.4 Grants

One large grant was submitted in March to the NIH under the BISTI program. Amount requested in direct costs is approximately 600,000.00. We have been advised that the BISTI grant will be funded but have not yet received formal notification of funding.

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.

2.5 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 August 1, 2002 we have had 35,545 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 137,972 *visits* to downloadable pages (defined as a unique IP accessing pages w/ no more then a 30 min pause between hits) and 1, 926,262 files downloaded (defined as downloadable files & wiki). These are approximately five fold (adjusted for time) increases over the previous year and all indications are that the rate of increase is itself increasing.

Since many packages rely on one central package we also track downloads of this package to get some sense of overall use. From August 2002 to August 2003 there were 26,203 downloads. From the inception of the project to August 2003 there were a total of 30,095.

We also produce a variety of meta-data packages for annotating genes and gene products. Approximately 9,000 of these have been downloaded during the past year.

3 Future Plans and Developments

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. This goal was identified in 2002 but we did not have the resources to address it. This year we have a post-doc who just arrived and will be working directly on these issues (funded from a P20 grant at HSPH).

- 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). Much work was done in this area during 2002 and we plan further developments during 2003/4.

- 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. This project has done very well during 2002 and will continue to be a major focus of the project.

4 Collaborations

We continue to work directly with scientists at the DFCI and other locations. These include (but are not limited to) the examples given below

- With Dr. K. Polyak we are developing tools for analysing SAGE data.

- With Dr. A. Miron we are developing tools for analysing SELDI-TOF data.

- Creating software packages to automate the process of assembling gene homology data for Dr. Vinay Akumar of NIH.

- Annotating/analyzing human and mouse gene chip data for Dr. Kenneth Christopher of Brigham and Women's Hospital.