2004 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. In 2003 we were awarded a BISTI grant R33HG002708.

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.

We intend to develop both software infrastructure tools and end user packages both by developers closely associated with the project (so-called *core developers*) and from the broader scientific community. It is not our intention that all such efforts be released through the Bioconductor web-site, but rather that we foster good programming and statistical practice within the bioinformatic, computational biology communities.

2 Specific Achievements

Release 1.3, our fourth release occurred in October 2003. There are 49 different software packages in Release 1.3 and approximately 80 packages in the development arm. Release 1.4 is scheduled for May of 2004.

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. However, tools for other technologies, such as SAGE, arrayCGH, protein mass spectrometry are also available through the project. A number of different machine learning tools have been developed and are being distributed.

A second major emphasis is the development of tools for dealing with graphs and networks.

Bioconductor was awarded a prize from Insightful Corporation for Innovative Software design.

2.1 Ph. D. Students

There are currently two Ph.D. students (both under the supervision of R. Gentleman), E. Whalen and K. Rader are working on network visualization and machine learning projects.

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.

Dr. B. Wittner is a post-doc funded from the BISTI grant.

2.2 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.

2.3 Website activity

We have had a few missteps in updating our web page. The current plan is for the new website to appear concurrently with Release 1.4 (or shortly thereafter).

Since October 30, 2003 we have had 35,653 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 117,172 *visits* to downloadable pages (defined as a unique IP accessing pages w/ no more then a 30 min pause between hits). and 953,583 files downloaded (defined as a file with the extension *.tar.gz*, *.zip*, or *.pdf*).

Since many packages rely on one central package we also track downloads of this package to get some sense of overall use. From October 2003 to April 2004 there were 15,214 downloads. From the inception of the project to October 2003 there were a total of 30,185.

We also produce a variety of meta-data packages for annotating genes and gene products. Approximately 31,689 of these have been downloaded during the past year. There were another 74,078 downloads of other data packages, including CDF data, probe sets, and experimental data.

2.4 Mailing List

The Bioconductor mailing list has been steadily growing in size and activity. There are currently 775 members on this mailing list and there have been 1,674 messages between November 1, 2003 and April 29, 2004. In the month of April 2004 there were 109 individual posters and accounts for 271 messages while March 2004 had 393 messages from 129 individuals.

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.