Bioconductor Annual Report 2005

Seth Falcon
Fred Hutchinson Cancer Research Center

October 2005

Contents

1 Summary of Core Tasks and Challenges

1.1 Automated package building and testing

The Bioconductor project provides access to its packages through package repositories hosted on `bioconductor.org`. One of the services provided to the Bioconductor community is the automated building and testing of all packages.

Maintaining the automated build and test suite and keeping the published package repositories updated requires a significant amount of time on the part of the Seattle Bioconductor team.

Challenges include:

- Dealing with cross-platform issues and especially with difficulties automating package building on Windows.

- The build requires substantial compute power and we must maintain many different platforms for testing.

1.2 Annotation data package building

The Bioconductor project synthesizes genomic and proteomic information available in public data repositories in order to annotate the probes of standard microarray chips. These annotation data packages are made available to the community and allow Bioconductor users to easily access meta data relating to their experimental platform.

In order to sythesize data from the various public repositories, we must maintain automated tools that can parse the available information. Due to quickly changing data standards, the maintaince of the code used to produce the annotation packages requires constant attention.

1.3 Other Tasks

- Providing user and developer support on project mail lists.

- Developing new functionality and improving architecture of key packages.

- Orchestrating the Bioconductor releases that occur every six months.

2 Size of Project

The Bioconductor project is comprised of R packages contributed by a worldwide bioinformatics community. There are currently 93 active developers and 149 contributed packages in Bioconductor's development repository. The project also maintains 241 annotation data packages that aid in the analysis of data from microarray experiments. Table **??** tracks the growth of the project over the semi-annual releases.

Release	1.1	1.2	1.3	1.4	1.5	1.6
Package Count	20	30	49	81	100	123

Table 1: Number of contributed packages included in each of the Bioconductor releases. Releases occur twice per year.

3 Bioconductor Electronic Mail Lists

The project maintains two email lists, `bioconductor`[1] and `bioc-devel`[2]. The `bioconductor` list is a forum for user questions, project announcements, and general discussion of interest to the Bioconductor community. As of September, 2005 the list has 1334 subscribers (individuals who receive mail from the list).

The `bioc-devel` list is a forum for package contributors' questions and discussion relating to the development of Bioconductor packages. As of September, 2005 this list has 202 subscribers.

Both lists provide a means of disseminating project news and a space for members of the community to share their knowledge about use of Bioconductor packages and best practices for data analysis.

Table 2 lists the number of posts and number of unique authors as a monthly average over the past four years.

4 The Bioconductor Website

The Bioconductor website, `http://bioconductor.org`, serves over 6000 unique visitors and over 200GB of content every month. In September, 2005, the site served 242GB of content of which 212GB (87%) corresponded

[1] `http://www.stat.math.ethz.ch/mailman/listinfo/bioconductor`

[2] `http://www.stat.math.ethz.ch/mailman/listinfo/bioc-devel`

Year	Posts/month	Authors/month
2002	60	26
2003	232	95
2004	320	135
2005	362	139

Table 2: Monthly average number of posts and number of unique authors for the bioconductor mail list for the years 2002–2005.

to package downloads. We host the website on a dual-Xeon 3.0GHz server with 2GB of RAM from Dell. Table 3 shows the number of unique visitors and outgoing bandwidth over the past four months.

Month	Unique Visitors	Bandwidth (GB)
Jun 2005	9009	277
Jul 2005	6341	232
Aug 2005	6101	195
Sep 2005	6133	242

Table 3: Webserver statistics for bioconductor.org for June–September, 2005

5 Package Building and Testing

The Bioconductor project is committed to providing packages for all computing platforms common in the bioinformatics community. We currently provide packages for Linux, Solaris, and most UNIX-like variants as well as Windows, and OS X.

To ensure that packages are consistently documented, easy to install, and functioning properly, we run a nightly build during which we test all packages in the development repository. Table 4 provides details on the systems we currently have available for the nightly build.

Platform	CPU	RAM	Build Time (Hours)
Linux 64-bit	dual-Xeon 3.40 GHz	4 GB	5
Linux 32-bit	dual-Xeon 2.80 GHz	4 GB	6
Windows 32-bit	dual-Xeon 2.80 GHz	2 GB	6
Solaris 64-bit	4x - Sparc 1.3 GHz	16 GB	12

Table 4: Servers used to build and test Bioconductor packages along with the number of hours required for a build/test cycle of all packages.

6 Accomplishments

6.1 Published Bioconductor Monograph

The Bioconductor "monograph", *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* was released on September 1, 2005.

6.2 Papers Citing Bioconductor

Based on a PubMed search for "bioconductor", there have been 18 publications citing Bioconductor since January 2005. They are listed in the bibliography of this report.

6.3 Bioconductor Courses

The following Bioconductor training courses were held in 2005:

- Bioconductor Training, Boston, Massachusetts, USA, July 21st-23rd 2005
- Bioconductor Short Course, Seattle, Washington, USA, June 1st-3rd 2005
- Computational and Statistical Aspects of Microarray Analysis, Bressanone-Brixen, June 19th-25th 2005

6.4 BioC2005 Conference

The Gentleman Lab organized a conference to highlight current Bioconductor developments and to provide a forum for discussing the use and design of software for analyzing data arising in biology with a focus on Bioconductor and genomic data.

The *BioC2005: Where Software and Biology Connect* conference was held in Seattle at the Fred Hutchinson Cancer Research Center on August 16–17, 2005. Over 87 scientists attended. The conference consisted of 11 talks from leading researchers in computational biology and 12 hands-on lab sessions presented by Bioconductor package developers.

Of the 12 attendees who responded to our conference evaluation survey, all 12 reported that they would consider coming again. Below are a few quotes from the survey responses:

Did the talks address topics of interest?

> Yes. It was very informative to hear from people doing leading edge analysis. I got a lot of new ideas to pursue.

What could have been improved?

> Perhaps more worked examples that we could have as references for after the class. I love the BioC vignettes and tutorials. So I see the labs as more of that, which is wonderful

> Seems like one of the main problems facing bioconductor is the quantity of microarray data. I'd like to have a nitty-gritty tutorial for people who want to incorporate c/c++ code into R.

> Perhaps I just missed this info ... but with greater advance notice, I could have gotten the software installed on my computer. The fact that this issue is so tricky is, ultimately, a comment on one of the real tough issue around Bioconductor itself. It demands quite alot from the users in terms of skill and determination.

> A lot of time was spent loading software and getting it to work - more copies of the relevant pieces available or ask people to load them before coming?

What was your main reason for coming to BioC2005?

> The labs. Get up to date on the latest BioC packages. Learn some new packages.

What did you like best about the conference?

> I really liked the format of the morning talks and the afternoon labs. I think the interactive labs after lunch was much preferred over possibly nodding off in the early afternoon. I really liked having a chance to interact with other core lab folks like Jim MacDonald as well as people critical to bioc like Seth, just to name a few.
>
> Good speakers and good crowd. A number of bioconductor developers were there and interested in interacting with others. Talks were also good.

7 Project Participants and Key Personnel

7.1 Gentleman Lab Members

These individuals, all working in the Gentleman Lab at the Fred Hutchinson Cancer Research Center in Seattle, Washington, played a central role in executing project objectives during 2005.

Seth Falcon Scientific programmer, project and release manager.

Brian Hodges System administrator, build manager.

Chenwei Lin January 2005 – June 2005. Developer in charge of annotation package construction and data management.

Ting-Yuan Liu August 2005 – Present. Developer in charge of annotation package construction and data management.

Alan Katz January 2005 – August 2005. R programmer, Windows build manager.

7.2 Bioconductor Core Developers

The Bioconductor Core Developers guide the project through direct project development and participation in design discussions.

- Douglas Bates, University of Wisconsin, USA.
- Ben Bolstad, Division of Biostatistics, UC Berkeley, USA.
- Vince Carey, Harvard Medical School, USA.

- Tony Chiang, Fred Hutchinson Cancer Research Center, USA.
- Marcel Dettling, Federal Inst. Technology, Switzerland.
- Sandrine Dudoit, Division of Biostatistics, UC Berkeley, USA.
- Byron Ellis, Harvard Department of Statistics, USA.
- Seth Falcon, Fred Hutchinson Cancer Research Center, USA
- Laurent Gautier, France.
- Robert Gentleman, Fred Hutchinson Cancer Research Center, USA.
- Jeff Gentry, Dana-Farber Cancer Institute, USA.
- Kurt Hornik, Wirtschaftsuniversität Wien, Austria.
- Torsten Hothorn, Institut fuer Medizininformatik, Biometrie und Epidemiologie, Germany.
- Wolfgang Huber, European Bioinformatics Institute, UK
- Stefano Iacus, Italy
- Rafael Irizarry, Department of Biostatistics (JHU), USA.
- Friedrich Leisch, Wirtschaftsuniversität Wien, Austria.
- Li Long, Swiss Institute of Bioinformatics, Switzerland.
- Ting-Yuan Liu, Fred Hutchinson Cancer Research Center, USA.
- James MacDonald, University of Michigan, USA.
- Martin Maechler, Federal Inst. Technology, Switzerland.
- Crispin Miller, Paterson Institute Bioinformatics Group, UK.
- Herves Pages, Fred Hutchinson Cancer Research Center, USA.
- Anthony Rossini, Biomedical and Health Informatics, University of Washington and Biostatistics, Fred Hutchinson Cancer Research Center, USA.
- Colin Smith, University of California, San Francisco, USA.
- Luke Tierney, University of Iowa, USA.

- Jean Yee Hwa Yang, University of California, San Francisco, USA.
- Jianhua (John) Zhang, Dana-Farber Cancer Institute, USA.

References

Andreas Buness, Wolfgang Huber, Klaus Steiner, Holger SÃŒltmann, and Annemarie Poustka. arraymagic: two-colour cdna microarray quality control and preprocessing. *Bioinformatics*, 21:554–556, Feb 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti052.

Vincent J. Carey, Jeff Gentry, Elizabeth Whalen, and Robert Gentleman. Network structures and algorithms in bioconductor. *Bioinformatics*, 21: 135–136, Jan 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth458.

Nema Dean and Adrian E. Raftery. Normal uniform mixture differential gene expression detection for cdna microarrays. *BMC Bioinformatics*, 6: 173, Jul 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-173.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21:3439–3440, Aug 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti525.

Fodor, Nelson, Alegria-Hartman, Robbins, Langlois, Turteltaub, Corzett, and McCutchen-Maloney. Statistical challenges in the analysis of two-dimensional difference gel electrophoresis experiments using decydertm. *Bioinformatics*, Aug 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti612.

Jelle J. Goeman, Jan Oosting, Anne-Marie Cleton-Jansen, Jakob K. Anninga, and Hans C. van Houwelingen. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21:1950–1957, May 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti267.

Qunhua Li, Chris Fraley, Roger E. Bumgarner, Ka Yee Yeung, and Adrian E. Raftery. Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, 21:2875–2882, Jun 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti447.

Claudio Lottaz and Rainer Spang. stam–a bioconductor compliant r package for structured analysis of microarray data. *BMC Bioinformatics*, 6:211, Aug 2005a. ISSN 1471-2105. doi: 10.1186/1471-2105-6-211.

Claudio Lottaz and Rainer Spang. Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics*, 21:1971–1978, May 2005b. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti292.

Henrik BjÃ?rn Nielsen, Laurent Gautier, and Steen Knudsen. Implementation of a gene expression index calculation method based on the pdnn model. *Bioinformatics*, 21:687–688, Mar 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti078.

Chiara Romualdi, Nicola Vitulo, Micky Del Favero, and Gerolamo Lanfranchi. Midaw: a web tool for statistical analysis of microarray data. *Nucleic Acids Res*, 33:W644–W649, Jul 2005. ISSN 1362-4962. doi: 10.1093/nar/gki497.

Stefanie Scheid and Rainer Spang. twilight; a bioconductor package for estimating the local false discovery rate. *Bioinformatics*, 21:2921–2922, Jun 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti436.

Denise Scholtens, Marc Vidal, and Robert Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21:3548–3557, Sep 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti567.

Mat Soukup, Hyungjun Cho, and Jae K. Lee. Robust classification modeling on microarray data using misclassification penalized posterior. *Bioinformatics*, 21 Suppl 1:i423–i423, Jun 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti1020.

Willenbrock and Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, Sep 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti677.

Claire L. Wilson and Crispin J. Miller. Simpleaffy: a bioconductor package for affymetrix quality control and data analysis. *Bioinformatics*, 21:3683–3685, Sep 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti605.

Witold E. Wolski, Maciej Lalowski, Peter Jungblut, and Knut Reinert. Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinformatics*, 6:203, Aug 2005. ISSN 1471-2105. doi: 10.1186/1471-2105-6-203.

Gunnar Wrobel, FrÃ?dÃ?ric Chalmel, and Michael Primig. gocluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, 21:3575–3577, Sep 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti574.