Bioconductor Annual Report 2007

Seth Falcon
Fred Hutchinson Cancer Research Center

October 2007

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. As the project has grown, the organizational and computational resources required to sustain the package build system have also increased.

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

We are also focusing resources on the underlying storage mechanism used for the annotation data packages. New high-throughput technologies such as SNP and exon arrays require significantly larger annotation libraries; the infrastructure requires improvement to support work with these emerging technologies.

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 150 active developers and 242 contributed packages in Bioconductor's development repository. The project also maintains 231 annotation data packages that aid in the analysis of data from microarray experiments. Table 1 tracks the growth of the project over the semi-annual releases.

Release	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Package Count	20	30	49	81	100	123	141	172	188	214

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 July, 2007 the list has 1694 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 July, 2007 this list has 351 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` averages over 21728 unique visitors and over 700GB of content per month. In June, 2007, the

[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	13
2003	232	47
2004	320	60
2005	352	61
2006	356	68
2007	407	80

Table 2: Monthly average number of posts and number of unique authors for the `bioconductor` mail list for the years 2002–2007.

site served 867GB of content of which 850GB (98%) corresponded to package downloads. The Biobase package was downloaded by 13683 unique IP addresses between January, 2007 and June, 2007. We host the website on a dual-Xeon 3.0GHz server with 2GB of RAM from Dell.

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 3 provides details on the systems we currently have available for the nightly build.

Platform	CPU	RAM	Build Time (Hours)
Linux 64-bit	quad-core Xeon 3.00 GHz	8 GB	7h + 1h
Windows 32-bit	dual-Xeon 2.80 GHz	2 GB	9h30 + 3h
Solaris 64-bit	4x - Sparc 1.3 GHz	16 GB	10h

Table 3: Servers used to build and test Bioconductor packages along with the number of hours required for a build/test cycle of all software packages (first number) and all experiment data packages (second number). The experiment data packages are not built on Solaris.

6 Accomplishments

6.1 Papers Citing Bioconductor

Based on a PubMed search for "bioconductor", there have been 27 publications citing Bioconductor to date in 2007. They are listed in the bibliography of this report.

6.2 Bioconductor Courses

The following Bioconductor training courses were held in 2007:

- Introductory level workshops on R, Bioconductor or Programming in R by Thomas Girke, UC Riverside.
- Statistical Analysis of Microarray Expression Data with R and Bioconductor - Copenhagen, DK - November 5-9, 2007
- BioC2007 - Seattle, WA, USA - August 2007
- Local Training - Seattle, WA - May 2007
- BioC Developers Meeting - Lausanne, Switzerland - April 4-5, 2007
- Advanced R Programming and Bioconductor - Hinxton, UK - 30 March - 1 April 2007
- Bioconductor Advanced Course - Seattle, WA, USA - January 2007
- Bioconductor is a central component of the Cold Spring Harbor Lab summer course on Integrative Data Analysis for High-throughput Biology (13-27 July 2007)

6.3 Sponsorships

- We provided travel expense and conference fee scholarships for attending the BioC2007 conference to four BioC package developers and three students for a total of seven scholarships.
- We provided $1500USD for student travel expenses to the DSC2007 conference in Auckland, NZ.
- Refreshments for two evening lab sessions at Cold Sping Harbor course

- Support for a last-minute substitution speaker, Alexandre Morozov, of Siggia Lab at Rockefeller University, at Interface Conference 2007, Philadelphia

6.4 BioC2006 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 *BioC2006: Where Software and Biology Connect* conference was held in Seattle at the Fred Hutchinson Cancer Research Center on August 3–4, 2006. Over 100 scientists attended. The conference consisted of 11 talks from leading researchers in computational biology and 13 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:

What was your main reason for coming to BioC2006?

> I use BioC packages for most of the data analysis I do. I know some packages pretty well, but I wanted to learn about other packages, as well as interact with people who are knowledgeable in the field.

Do you have any general comments about morning speakers?

> It was good to see the new issues and areas in genomics, both from a researcher's perspective and from the developer's perspective. I thought it was a good mix between biologists and biostatisticians.

What did you like best about the conference?

> Meeting and talking with the "names" on the BioC mailing list. The internet is great, but I just can't replace the ease of verbal communication, especially with a group. I had several important discussions which never would have happened on the mailing list.

Was two days enough time, too much time?

> I would have liked it to be a day longer, even if the 3rd day was all labs. You really could only go to 2 of the 14 lab sessions, and I would have liked to gone to 5 or 6 of them. Having access to all the lab's materials on-line is helpful - I can attempt to go through them on my own.

BioC2007 will take place in Seattle, August 6th-8th, 2007.

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 2007.

Seth Falcon Scientific programmer, project and release manager.

Martin Morgan Developer in charge of Biobase.

Nianhua Li Developer in charge of annotation data package construction and data management. In May, Nianhua left the group.

Marc Carlson Marc Carlson has joined the group to take over the management of the annotation data packages.

Herve Pages Developer in charge of the build and test system.

7.2 Harvard Medical School Members

Vincent Carey Co-investigator.

7.3 European Bioinformatics Institute Members

Wolfgang Huber Co-investigator.

7.4 Johns Hopkins University School of Hygiene and Public Health Members

Rafael Irizarry Co-investigator.

References

Oliver Bembom, Sunduz Keles, and Mark J van der Laan. Supervised detection of conserved motifs in DNA sequences with cosmo. *Stat Appl Genet Mol Biol*, 6:Article8, 2007.

Gianluca Bontempi. A blocking strategy to improve gene selection for classification of gene expression data. *IEEE/ACM Trans Comput Biol Bioinform*, 4(2):293–300, Apr 2007.

Michael Boutros, Ligia P Bras, and Wolfgang Huber. Analysis of cell-based RNAi screens. *Genome Biol*, 7(7):R66, 2006.

Harbron C, Chang KM, and South MC. RefPlus : an R package extending the RMA Algorithm. *Bioinformatics*, Jul 2007a.

Lottaz C, Toedling J, and Spang R. Annotation-based Distance Measures for Patient Subgroup Discovery in Clinical Microarray Studies. *Bioinformatics*, Jun 2007b.

Vincent J Carey, Martin Morgan, Seth Falcon, Ross Lazarus, and Robert Gentleman. GGtools: analysis of genetics of gene expression in bioconductor. *Bioinformatics*, 23(4):522–523, Feb 2007.

Benilton Carvalho, Henrik Bengtsson, Terence P Speed, and Rafael A Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8(2):485–499, Apr 2007.

Zhu D, Li Y, and Li H. Multivariate Correlation Estimator for Inferring Functional Relationships from Replicated Genome-wide Data. *Bioinformatics*, Jun 2007.

Diego Diez, Rebeca Alvarez, and Ana Dopazo. Codelink: an R package for analysis of GE healthcare gene expression bioarrays. *Bioinformatics*, 23 (9):1168–1169, May 2007.

S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan 2007.

Zhao H, Engelen K, De Moor B, and Marchal K. CALIB: a Bioconductor package for estimating absolute expression levels from two-color microarray data. *Bioinformatics*, May 2007.

Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, Nov 2006.

Bowen JM, Gibson RJ, Tsykin A, Stringer AM, Logan RM, and Keefe DM. Gene expression analysis of multiple gastrointestinal regions reveals activation of common cell regulatory pathways following cytotoxic chemotherapy. *Int J Cancer*, Jun 2007.

Ted Laderas and Shannon McWeeney. Consensus framework for exploring microarray data using multiple clustering methods. *OMICS*, 11(1):116–128, Spring 2007.

Nolwenn Le Meur, Anthony Rossini, Maura Gasparetto, Clay Smith, Ryan R Brinkman, and Robert Gentleman. Data quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry A*, 71 (6):393–403, Jun 2007.

Jiajun Liu, Jacqueline M Hughes-Oliver, and J Alan Jr Menius. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*, 23(10):1225–1234, May 2007.

Jun Lu, Joseph C Lee, Marc L Salit, and Margaret C Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, 2007.

Andrew McQuillin, Mie Rizig, and Hugh M D Gurling. A microarray gene expression study of the molecular pharmacology of lithium carbonate on mouse brain mRNA to understand the neurobiology of mood stabilization and treatment of bipolar affective disorder. *Pharmacogenet Genomics*, 17 (8):605–617, Aug 2007.

Dunning MJ, Smith ML, Ritchie ME, and Tavare S. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, Jun 2007a.

Okoniewski MJ, Yates T, Dibben S, and Miller CJ. An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. *Genome Biol*, 8(5):R79, May 2007b.

Davis S and Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, May 2007.

Robert B Scharpf, Jason C Ting, Jonathan Pevsner, and Ingo Ruczinski. SNPchip: R classes and methods for SNP array data. *Bioinformatics*, 23 (5):627–628, Mar 2007.

Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, May 2007.

Maria Stalteri and Andrew Harrison. Comparisons of annotation predictions for affymetrix GeneChips. *Appl Bioinformatics*, 5(4):237–248, 2006.

Maria A Stalteri and Andrew P Harrison. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8:13, 2007.

A L Tarca, V J Carey, X W Chen, R Romero, and S Draghici. Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6), Jun 2007. doi: 10.1371/journal.pcbi.0030116. URL `http://www.hubmed.org/display.cgi?uids=17604446`.

Joern Toedling, Oleg Sklyar, and Wolfgang Huber. Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, 8:221, 2007.

E S Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, Mar 2007.