New supporting members

Michael Bojanowski (Netherlands)
Caoimhin ua Buachalla (Ireland)
Jake Bowers (UK)
Sandrah P. Eckel (USA)
Charles Fleming (USA)
Hui Huang (USA)
Jens Oehlschlägel (Germany)
Marc Pelath (UK)
Tony Plate (USA)
Julian Stander (UK)

Kurt Hornik
Wirtschaftsuniversität Wien, Austria
Kurt.Hornik@R-project.org

News from the Bioconductor Project

by Hervé Pagès and Martin Morgan

Bioconductor 2.1 was released on October 8, 2007 and is designed to be compatible with R 2.6.0, released five days before Bioconductor. This release contains 23 new software packages, 97 new annotation (or metadata) packages, and many improvements and bug fixes to existing packages.

New packages

The 23 new software packages provide exciting analytic and interactive tools. Packages address Affymetrix array quality assessment (e.g., **arrayQualityMetrics**, **AffyExpress**), error assessment (e.g., **OutlierD**, **MCRestimate**), particular application domains (e.g., comparative genomic hybridization, **CGHcall**, **SMAP**; SNP arrays, **oligoClasses**, **RLMM**, **VanillaICE**; SAGE, **sagenhaft**; gene set enrichment, **GSEABase**; protein interaction, **Rintact**) and sophisticated modeling tools (e.g., **timecourse**, **vbmp**, **maigesPack**). **exploRase** uses the GTK toolkit to provide an integrated user interface for systems biology applications.

Several packages benefit from important infrastructure developments in R. Recent changes allow consolidation of C code common to several packages into a single location (**preprocessCore**), greatly simplifying code maintenance and improving reliability. Many new packages use the S4 class system. A number of these extend classes provided in **Biobase**, facilitating more seamless interoperability. The ability to access web resources, including convenient XML parsing, allow Bioconductor packages such as **GSEABase** to access important curated resources.

SQLite-based annotations

This release provides SQLite-based annotations in addition to the traditional environment-based ones. Annotations contain *maps* between information from microarray manufactures, standard naming conventions (e.g., Entrez gene identifiers) and resources such as the Gene Ontology consortium. Eighty-six SQLite-based annotation packages are currently available. The name of these packages end with `".db"` (e.g., **hgu95av2.db**). For an example of different metadata packages related to specific chips, view the annotations available for the hgu95av2 chip: `http://bioconductor.org/packages/2.1/hgu95av2.html`

New *genome wide* metadata packages provide a more complete set of maps, similar to those provided in the chip-based annotation packages. Genome wide annotations have an `"org."` prefix in their name, and are available as SQLite-based packages only. Five organisms are currently supported: human (**org.Hs.eg.db**), mouse (**org.Mm.eg.db**), rat (**org.Rn.eg.db**), fly (**org.Dm.eg.db**) and yeast (**org.Sc.sgd.db**). The ***LLMappings** packages will be deprecated in Bioconductor 2.2.

Environment-based (e.g., **hgu95av2**) and SQLite-based (e.g., **hgu95av2.db**) packages contain the same data. For the end user, moving from **hgu95av2** to **hgu95av2.db** is transparent because the objects (or *maps*) in **hgu95av2.db** can be manipulated as if they were environments (i.e., functions `ls`, `mget`, `get`, etc. . . still work on them). Using SQLite allows considerable flexibility in querying maps and in performing complex joins between tables, in addition to placing the burden of memory management and optimized query construction in sqlite. As with the implementation of operations like `ls` and `mget`, the intention is to recognize common use cases and to code these so that R users do not need to know the underlying SQL query.

Looking ahead

For the next release (BioC 2.2, April 2008) all our annotations will be SQLite-based and we will deprecate the environment-based versions.

We anticipate increasing emphasis on sequence-based technologies like Solexa (`http://www.illumina.com`) and 454 (`http://www.454.com`). The volume of data these technologies generate is very large (a three day Solexa run produces almost a terabyte of raw data, with 10's of gigabytes appropriate

for numerical analysis). This volume of data poses significant challenges, but graphical, statistical, and interactive abilities and web-based integration make R a strong candidate for making sense of, and making fundamental new contributions to, understanding and critically assessing this data.

The best Bioconductor packages are contributed by our users, based on their practical needs and sophisticated experience. We look forward to receiving your contribution over the next several months.

Hervé Pagès
Computational Biology Program
Fred Hutchinson Cancer Research Center
hpages@fhcrc.org

Martin Morgan
Computational Biology Program
Fred Hutchinson Cancer Research Center
mtmorgan@fhcrc.org

Past Events: useR! 2007

Duncan Murdoch and Martin Maechler

The first North American useR! meeting took place over three hot days at Iowa State University in Ames this past August.

The program started with John Chambers talking about his philosophy of programming: our mission is to enable effective and rapid exploration of data, and the prime directive is to provide trustworthy software and "tell no lies". This was followed by a varied and interesting program of presentations and posters, many of which are now online at http://useR2007.org. A panel on the use of R in clinical trials may be noteworthy because of the related publishing by the R Foundation of a document (http://www.r-project.org/certification.html) on "Regulatory Compliance and Validation" issues. In particular "21 CFR part 11" compliance is very important in the pharmaceutical industry.

The meeting marked the tenth anniversary of the formation of the R Core group, and an enormous blue R birthday cake was baked for the occasion (Figure 1). Six of the current R Core members were present for the cutting, and Thomas Lumley gave a short after dinner speech at the banquet.

Figure 1: R Core turns ten.

There were two programming competitions. The first requested submissions in advance, to produce a package useful for large data sets. This was won by the team of Daniel Adler, Oleg Nenadić, Walter Zucchini and Christian Gläser from Göttingen. They wrote the **ff** package to use paged memory-mapping of binary files to handle very large datasets. Daniel, Oleg and Christian attended the meeting and presented their package.

The second competition was a series of short R programming tasks to be completed within a time limit at the conference. The tasks included relabelling, working with ragged longitudinal data, and writing functions on functions. There was a tie for first place between Elaine McVey and Olivia Lau.

Three kinds of T-shirts with variations on the "R" theme were available: the official conference T-shirt (also worn by the local staff) sold out within a day, so the two publishers' free T-shirts gained even more attraction.

Local arrangements for the meeting were handled by Di Cook, Heike Hofmann, Michael Lawrence, Hadley Wickham, Denise Riker, Beth Hagenman and Karen Koppenhaver at Iowa State; the Program Committee also included Doug Bates, Dave Henderson, Olivia Lau, and Luke Tierney. Thanks are due to all of them, and to the team of Iowa State students who made numerous trips to the Des Moines airport carrying meeting participants back and forth, and who kept the equipment and facilities running smoothly–useR! 2007 was an excellent meeting.

Duncan Murdoch, University of Western Ontario
Duncan.Murdoch@R-project.org
Martin Maechler, ETH Zürich
Martin.Maechler@R-project.org