News from the Bioconductor Project

Bioconductor Team
Program in Computational Biology
Fred Hutchinson Cancer Research Center

We are pleased to announce Bioconductor 2.3, released on October 22, 2008. Bioconductor 2.3 is compatible with R 2.8.0, and consists of 294 packages. There are 37 new packages, and enhancements to many others. Explore Bioconductor at `http://bioconductor.org`, and install standard or individual packages with

```
> source("http://bioconductor.org/biocLite.R")
> biocLite() # install standard packages...
> biocLite("rtracklayer") # or rtracklayer
```

New packages

New to this release are powerful packages for diverse areas of high-throughput analysis. Highlights include:

Assay technologies such as HELP and MeDIP DNA methylation (**HELP**, **MEDME**), micro-RNAs (**microRNA**, **miRNApath**), and array comparative genomic hybridization (**ITALICS**, **CGHregions**, **KCsmart**).

Pathways and integrative analysis packages such as **SIM** (gene set enrichment for copy number and expression data), **domainsignatures** (InterPro gene set enrichment), **minet** (mutual information network inference) and **miRNApath** (pathway enrichment for micro-RNA data).

Interoperability with the ArrayExpress data base (**ArrayExpress**), the Chemmine chemical compound data base (**ChemmineR**), the PSI-MI protein interaction data base (**RpsiXML**) and external visualization tools such as the X:Map exon viewer (**xmapbridge**).

Algorithms for machine learning, such as **BicARE**, **dualKS**, **CMA**) and iterative Bayesian model averaging (**IterativeBMA**, **IterativeBMAsurv**).

Refined expression array methods including preprocessing (e.g., **multiscan**), contaminant and outlier detection (**affyContam**, **arrayMvout**, **parody**), and small-sample and other assays for differential expression (e.g., **logitT**, **DFP**, **LPEadj**, **PLPE**).

Annotations

Bioconductor 'Annotation' packages contain biological information about microarray probes and the genes they are meant to interrogate, or contain ENTREZ gene-based annotations of whole genomes. This release marks the completion of our transition from environment-based to SQLite-based annotation packages. SQLite annotation packages allow for efficient memory use and facilitate more complicated data queries. The release now supports a menagerie of 15 different model organisms, from Arabidopsis to Zebrafish, nearly doubling the number of species compared to the previous Bioconductor release. We have also increased the amount of information in each annotation package; the inst/NEWS file in **AnnotationDbi** provides details.

High-throughput sequencing

An ensemble of new or expanded packages introduces tools for 'next generation' DNA sequence data. This data is very large, consisting of 10s to 100s of millions of 'reads' (each 10s to 100s of nucleotides long), coupled with whole genome sequences (e.g., 3 billion nucleotides in the human genome). **BSgenome** provides a foundation for representing whole-genome sequences; there are 13 model organisms represented by 18 distinct genome builds, and facilities for users to create their own genome packages. Functionality in the **Biostrings** package performs fast or flexible alignments between reads and genomes. **ShortRead** provides tools for import and exploration of common data formats. **IRanges** offers an emerging infrastructure for representing very large data objects, and for range-based representations. The **rtracklayer** package interfaces with genome browsers and their track layers. **HilbertVis** and **HilbertVisGUI** provide an example of creative approaches to visualizing this data, using space-filling (Hilbert) curves that maintain, as much as possible, the spatial information implied by linear chromosomes.

Other activities

Bioconductor package maintainers and the Bioconductor team invest considerable effort in producing high-quality software. New packages are reviewed on both technical and scientific merit, before being added to the 'development' roster for the next release. Release packages are built daily on Linux, Windows, and Mac platforms, tracking the most recent released versions of R and the packages hosted on CRAN repositories. The active Bioconductor mailing lists (`http://bioconductor.org/docs/mailList.html`) connect users with each other, to domain experts, and to maintainers eager to ensure that their packages satisfy the needs of leading edge approaches. The Bioconductor community meets at our annual conference in Seattle; well over 100 individuals attended the July meeting for a combination of scientific talks and hands-on tutorials.

Looking forward

This will be a dynamic release cycle. New contributed packages are already under review, and our build machines have started tracking the latest development versions of R. It is hard to know what the future holds, but past experience points to surprise — at the creativity, curiosity, enthusiasm, and commitment of package developers, and at the speed of technological change and growth of data. In addition to development of high-quality algorithms to address microarray data analysis, we anticipate continued efforts to leverage diverse external data sources and to meet the challenges of presenting high volume data in rich graphical contexts.

High throughput sequencing technologies represent a particularly likely area for development. First-generation approaches with relatively short reads, restricted application domains, and small numbers of sample individuals are being supplanted by newer technologies producing longer and more numerous reads. New protocols and the intrinsic curiosity of biologists are expanding the range of questions being addressed, and creating a concomitant need for flexible software analysis tools. The increasing affordability of high throughput sequencing technologies means that multi-sample studies with non-trivial experimental designs are just around the corner. The statistical analytic abilities and flexibility of R and Bioconductor represent ideal tools for addressing these challenges.