An Introduction to R and Bioconductor

Robert Gentleman
Program in Computational Biology
Fred Hutchinson Cancer Research Center

© Copyright 2008, all rights reserved
Overview

• biology is a computational science
• problems of data analysis, data generation, reproducibility require computational support and computational solutions
• we value code reuse
  – many of the tasks have already been solved
  – if we use those solutions we can put effort into new research
• well designed, self-describing data structures help us deal with complex data
Goals

• Provide access to powerful statistical and graphical methods for the analysis of genomic data.

• Facilitate the integration of biological metadata (GenBank, GO, Entrez Gene, PubMed) in the analysis of experimental data.

• Allow the rapid development of extensible, interoperable, and scalable software.

• Promote high-quality documentation and reproducible research.

• Provide training in computational and statistical methods.
Bioconductor

• Bioconductor is an open source and open development software project for the analysis of biomedical and genomic data.
• The project was started in the Fall of 2001 and includes core developers in the US, Europe, and Australia.
• R and the R package system are used to design and distribute software.
• A goal of the project is to develop software modules that are integrated and which make use of available web services to provide comprehensive software solutions to relevant problems.
• ArrayAnalyzer: Commercial port of Bioconductor packages in S-Plus.
Why are we Open Source

- so that you can find out what algorithm is being used, and how it is being used
- so that you can modify these algorithms to try out new ideas or to accommodate local conditions or needs
- so that they can be used as components (potentially modified)
Bioconductor packages
Release 2.3, 294 Software Packages!

- **General infrastructure:**
  Biobase, Biostrings, biocViews
- **Annotation:**
  annotate, annaffy, biomaRt, AnnotationDbi ➔ data packages.
- **Graphics/GUIs:**
  geneplotter, hexbin, limmaGUI, exploRase
- **Pre-processing:**
  affy, affycomp, oligo, makecdfenv, vsn, gcrm, limma
- **Differential gene expression:**
  genefilter, limma, ROC, siggenes, EBArrays, factDesign
- **GSEA/Hypergeometric Testing**
  GSEABase, Category, GOstats, topGO
- **Graphs and networks:**
  graph, RBGL, Rgraphviz
- **Flow Cytometry:**
  prada, flowCore, flowViz, flowUtils
- **Protein Interactions:**
  ppiData, ppiStats, ScISI, Rintact
- **Sequence Data:**
  Biostrings, ShortRead, chipseq, rtracklayer, HilbertVis
- **Other data:**
  xcms, DNAcopy, PROcess, aCGH, rsbml, SBMLR, Rdisop
Component software

- most interesting problems will require the coordinated application of many different techniques
- thus we need integrated interoperable software
- web services are one tool
- well designed software modules are another
- you should design your piece to be a cog in a big machine
Data complexity

- Dimensionality.
- Dynamic/evolving data: e.g., gene annotation, sequence, literature.
- Multiple data sources and locations: in-house, WWW.
- Multiple data types: numeric, textual, graphical.

No longer $X_{nxp}$!

We distinguish between biological metadata and experimental metadata.
Experimental metadata

• Gene expression measures
  – scanned images, i.e., raw data;
  – image quantitation data, i.e., output from image analysis;
  – normalized expression measures,
  – Reliability/quality information for the expression measures.

• Information on the probe sequences printed on the arrays (array layout).

• Information on the target samples hybridized to the arrays.

• See Minimum Information About a Microarray Experiment (MIAME) standards and the MAGEML package.
Biological metadata

• Biological attributes that can be applied to the experimental data.
• E.g. for genes
  – chromosomal location;
  – gene annotation (Entrez Gene, GO);
  – relevant literature (PubMed).

• Biological metadata sets are large, evolving rapidly, and typically distributed via the WWW.

• Tools: annotate, annaffy, biomaRt, and AnnotationDbi packages, and annotation data packages.

• Please use the db versions of all meta-data packages.
Annotation packages: annotate, annafy, biomaRt, and AnnotationDbi

- Assemble and process genomic annotation data from public repositories.
- Build annotation data packages.
- Associate experimental data in real time to biological metadata from web databases such as GenBank, GO, KEGG, Entrez Gene, and PubMed.
- Process and store query results: e.g., search PubMed abstracts.
- Generate HTML reports of analyses.
Bioconductor developed a new documentation paradigm, the vignette.

A vignette is an executable document consisting of a collection of documentation text and code chunks.

Vignettes form dynamic, integrated, and reproducible statistical documents that can be automatically updated if either data or analyses are changed.

Vignettes can be generated using the Sweave function from the R tools package.
Short Courses/Conferences

• we have given many short courses
  – see bioconductor.org for more details on upcoming courses

• BioC2009 - Seattle, July 27-29
Bioconductor Software

• we concentrate our development on a few important aspects
• **Biobase**: core classes and definitions that allow for succinct description and handling of the data
• **annotate**: generic functions for annotation that can be specialized
• **genefilter**: fast filtering via virtually every mechanism
• **graph/Rgraphviz/RBGL**: code for handling graphs and networks
• **Biostrings/ShortRead** (and soon friends): string manipulations, sequence analysis
Biobase:ExpressionSet

- software should help organize and manipulate your data
- this was the intention of the original exprSet class
- the data need to be assembled correctly once, and then they can be processed, subset etc without worrying about them
- exprSet was too limited (and too oriented to single channel arrays)
- we developed the new ExpressionSet class
Microarray data analysis

Pre-processing
- CEL, CDF
  - affy
  - vsn
- .gpr, .Spot
  - marray
  - limma
  - vsn

ExpressionSet

Differential expression
- edd
- genefilter
- limma
- multtest
- ROC
- + CRAN

Graphs & networks
- graph
- RBGL
- Rgraphviz

Cluster analysis
- CRAN
  - class
  - cluster
  - MASS
  - mva

Prediction
- CRAN
  - class
  - e1071
  - ipred
  - LogitBoost
  - MASS
  - nnet
  - randomForest
  - rpart

Annotation
- annotate
- annaffy
- biomaRt
  + metadata packages

Graphics
- geneplotter
- hexbin
  + CRAN
Pre-processing oligonucleotide chip data:
• diagnostic plots,
• background correction,
• probe-level normalization,
• computation of expression measures.
graph and Rgraphviz
The Arp2/3 complex is a stable multiprotein assembly required for the nucleation of actin filaments in all eukaryotic cells and consists of seven proteins in human and yeast.

Quality Assessment using arrayQualityMetrics

• comprehensive QA assessment of microarrays (one color or two color)
• easy to browse HTML
Machine Learning

• A new machine learning package
  MLInterfaces
• goal is to provide uniform calling sequences and return values for all machine learning algorithms
• we have postpended a B (e.g. knnB)
• return values are of class classifOutput
• see the MLInterfaces vignette for more details
Publications


• Bioconductor Case Studies, *Springer*

• *R Programming for Bioinformatics*, Chapman Hall
References

• **R** [www.r-project.org](http://www.r-project.org), [cran.r-project.org](http://cran.r-project.org)
  – software (CRAN);
  – documentation;
  – newsletter: R News;
  – mailing list.

• **Bioconductor** [www.bioconductor.org](http://www.bioconductor.org)
  – software, data, and documentation (vignettes);
  – training materials from short courses;
  – mailing list (please read the posting guide)