Bioconductor: Annotation Package Overview

October 28, 2003

1 Overview

The annotation library provides an interface to relevant biological data. Assembling annotation data and associating it with the relevant experimental data is not a simple task. This package is likely to be in a constant state of flux for at least two reasons. First the available data itself will be in a state of flux for the forseable future. Our understanding of it and how it relates to the experimental data will also mature over time.

The annotation for Bioconductor is handled by two systems. One, *AnnBuilder* is a system for assembling and relating the data from various sources. It is much more *industrial* and takes advantage of many different non-R tools. These include Perl and a relational database. The second package is *annotate*. This package is designed to provide experiment level annotation data suitable for the analysis of individual experiments (or combinations of experiments).

Any given experiment typically involves a set of known identifiers (probes in the case of a microarray experiment). These identifiers are typically unique (for any manufacturer). This holds true for any of the standard databases such as LocusLink. However, when the identifiers from one source are linked to the identifiers from another there does not need to be a one–to–one relationship. For example, several different Affymetrix accession numbers correspond to a single LocusLink identifier. Thus, when going one direction (Affymetrix to LocusLink) we have no problem, but when going the other we need some mechanism for dealing with the multiplicity of matches.

Databases have a long history and there are some solutions available for the issues of one–to–many and many–to–many relationships. However, in *annotate* our approach is to use hash tables to provide translation. Thus we must either break the one–to–many relationships or have the stored value in a hash table be a vector of values.

While it would be reasonably straightforward to provide an implementation of *annotate* that used a relational database rather than hash tables we currently feel that this would be little used and hence not worth the effort. However, we would be happy to provide advice and support if an effort to build such an interface is attempted.

Experimental Data

We currently can access a great deal of data. Examples include chromosome number, chromosomal location (cytoband or bp). The Gene Ontology (GO) categorizations. Other information such as syntenic regions or orthologous grouping can also be obtained.

Since these data are likely to be constantly changing (and we will be constantly updating the data) we will provide some data with the annotate package but will also provide data for downloading (automatically) from the Bioconductor website (www.bioconductor.org). The data will be provided in two different forms. One is in files marked up in XML. The other will be as saved R data objects. The latter will be preferable for most users since they are relatively fast to load and provide the data in an internal format ready for use.

Researchers with special needs should feel free to contact us regarding the production of annotation data specialized to their needs.

2 Some examples

In the following example we show how to produce a simple web page with links to Locus Link (at NCBI) for genes that were selected according to some criteria.

```
> library(Biobase)

Welcome to Bioconductor
         Vignettes contain introductory material.  To view,
         simply type: openVignette()
         For details on reading vignettes, see
         the openVignette help page.

> library(genefilter)
> library(annotate)
> data(geneData)
> f1 <- kOverA(10, 500)
> ff <- filterfun(f1)
> whichg <- genefilter(geneData, ff)
> sum(whichg)

[1] 60

> genes.used <- geneData[whichg, ]
> mnexpr <- apply(genes.used, 1, mean)
> ord.mn <- order(mnexpr)
> genes.ord <- genes.used[ord.mn, ]
> data(hgu95All)
> llnames <- multiget(row.names(genes.ord), env = hgu95All)
> ll.htmlpage(llnames, "selectedG.html", "Top 60 Genes", row.names(genes.ord))
```

3 Translators Vignette

In this section we show how to build and use translators. Many of the basic data sets required for this step are distributed with the `annotate` library. Others are available from the Bioconductor web site.