BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 22 2011, pages 3204–3205
doi:10.1093/bioinformatics/btr529

Gene expression Advance Access publication September 21, 2011

inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO

Jonatan Taminau[1,*], David Steenhoff[1], Alain Coletta[2], Stijn Meganck[1,3], Cosmin Lazar[1], Virginie de Schaetzen[1], Robin Duque[2], Colin Molter[2], Hugues Bersini[2], Ann Nowé[1] and David Y. Weiss Solís[2]

[1]CoMo, Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2, [2]IRIDIA-CoDE, Université Libre de Bruxelles, Avenue F. D. Roosevelt 50 and [3]ETRO, Department of Electronics and Informatics, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Associate Editor: Janet Kelso

ABSTRACT

Microarray technology has become an integral part of biomedical research and increasing amounts of datasets become available through public repositories. However, re-use of these datasets is severely hindered by unstructured, missing or incorrect biological samples information; as well as the wide variety of preprocessing methods in use. The inSilicoDb R/Bioconductor package is a command-line front-end to the InSilico DB, a web-based database currently containing 86 104 expert-curated human Affymetrix expression profiles compiled from 1937 GEO repository series. The use of this package builds on the Bioconductor project's focus on reproducibility by enabling a clear workflow in which not only analysis, but also the retrieval of verified data is supported.

Availability: inSilicoDb is available as part of the Bioconductor project. There is a companion web interface that can be used for browsing available datasets before importing them into R/Bioconductor (http://insilico.ulb.ac.be).

Contact: jtaminau@vub.ac.be

Received on June 29, 2011; revised on September 1, 2011; accepted on September 15, 2011

1 THE INSILICO DATABASE: BUILDING ON GEO

With >500 000 genomic profiles freely available in NCBI GEO (Edgar *et al.*, 2002), there is a huge amount of genome-wide information available, which could contain the clues necessary to treat fatal diseases such as cancer. However, accessibility to these datasets requires complex and intensive computational steps. Additionally, manual parsing and compilation of experimental attributes and values is tedious and error-prone (Baggerly and Coombes, 2009). Also, the large number of normalization and preprocessing methods in use make the comparison of different existing studies difficult, or even impossible (Sims, 2009).

The inSilico project has uniformly compiled a large amount of human Affymetrix gene expression studies (Colletta *et al.*, in press) from publicly available datasets in GEO–*ca.* 1 billion gene expression measurements. Through the inSilicoDb package, the InSilico DB content is made available for enhanced programmatic access. inSilicoDb enables large-scale genome-wide analysis through automated scripting by seamless integration with the R/Bioconductor genome-wide datasets visualization and analysis platform (Gentleman *et al.*, 2004).

*To whom correspondence should be addressed.

2 USING THE INSILICODB PACKAGE

Other software packages to retrieve gene expression datasets in R/Bioconductor exist (Davis and Meltzer, 2007). However, the information about the samples is in a raw form requiring a manual curation step in transit between a data repository (e.g. GEO) and a data analysis platform (e.g. R/Bioconductor). In contrast, inSilicoDb streamlines this process by providing data manually verified by volunteers from the scientific community. Contributed content is entered through a spreadsheet-based online collaborative biocuration platform. This approach avoids the complexity of defining and using formal ontologies (Shah *et al.*, 2009), but requires trust in the contributions made by the community.

Accessing these data from the InSilico DB is simple and straightforward with the inSilicoDb package:

```
> library("inSilicoDb");
> eset = getDataset("GSE781", "GPL96");
> dim(eset);
Features Samples
   22283      17
```

The getDataset function queries the InSilico DB for a given series and platform identifiers and returns an `ExpressionSet` object, a standard R/Bioconductor data structure. In GEO, a *Series* is composed of samples assayed on one or more platforms, each platform containing tens of thousands of gene measurement probe sets. In InSilico DB, the series are conveniently represented by multiple samples **x** genes matrices. Two auxiliary functions to allow flexible management of studies with multiple platforms are provided: `getDatasets` to retrieve, for a given series, all gene expression matrices, and `getPlatforms` to retrieve all platforms:

```
> getPlatforms("GSE781");
[1] "GPL96" "GPL97"

> esets = getDatasets("GSE781");
> sapply(esets, annotation);
[1] "hgu133a" "hgu133b"
```

By default, numerical data downloaded is identical to the data originally provided by GEO. However, when combining different

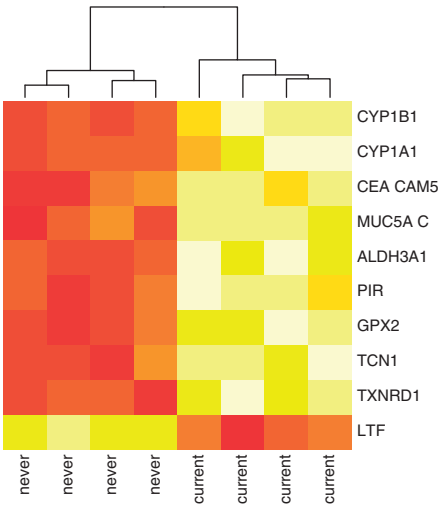

Fig. 1. Heatmap of discriminating genes of dataset `GSE4635` at adjusted $P < 10\%$. The samples cluster by 'Smoker' status (smokers : `current` and non-smokers: `never`).

studies, a consistent preprocessing is required and therefore all studies for which raw data exists are available re-normalized [fRMA normalization algorithm (McCall *et al.*, 2010)].

The unit of interest is usually the gene but there is no consensus on how to combine microarray probe measurements mapping to the same gene. The inSilicoDb offers the option to retrieve the probe set- or gene-level measurements. For the *probe-to-gene mapping*, the most recent versions of the R/Bioconductor platform annotation packages are used (`hgu133a.db`, `hgu133plus2.db`, …). The maximum probe set value is retained when multiple probe sets map to the same gene.

The following example illustrates the use of the `genes` parameter.

```
> eset = getDataset("GSE781", "GPL96",
                    genes=FALSE);
> dim(eset);
  Features  Samples
     22283       17

> eset = getDataset("GSE781", "GPL96",
                    genes=TRUE);
> dim(eset);
  Features  Samples
     12679       17
```

Both normalization and gene/probe options are precomputed for all datasets and are therefore as fast and easy to retrieve as the original data.

3 AN EXAMPLE

Once an expression set is retrieved, all available Bioconductor packages can be applied for further analysis. Executing the code below results in the clustered heatmap shown in Figure 1:

```
> library("limma")
> library("inSilicoDb")

> eset = getDataset("GSE4635", "GPL96",
                    norm="FRMA", genes=TRUE)

> labels = pData(eset)[ ,"Smoker"]
> design = model.matrix(~labels)

  # Find significant genes
> fit = eBayes(lmFit(eset, design))
> t = topTable(fit, coef=2)[,c("ID","logFC","adj.P.Val")]
> deg = t[sort(t[,"logFC"],index=TRUE,decr=TRUE)$ix,"ID"]

> heatmap(exprs(eset[deg,]), labCol=labels, Rowv=NA)
```

More information and examples to perform large-scale analyses can be found in the accompanying vignette of this package.

4 CONCLUSION

The inSilicoDb R/Bioconductor package provides an efficient means of performing large-scale genomic analysis on the large and growing amount of human Affymetrix gene expression profiles using automated scripting. The accompanying web interface (InSilico DB) allows search and browsing of curated datasets that can then be automatically retrieved, adding a means for reproducible data sourcing to the reproducible analysis platform R/Bioconductor.

Funding: Brussels-Capital Region, Innoviris in part.

Conflict of Interest: none declared.

REFERENCES

Baggerly,K.A. and Coombes,K.R. (2009) Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.*, **3**, 1309–1334.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*, **23**, 1846–1847.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

McCall,M.N. *et al.* (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

Shah,N.H. *et al.* (2009) Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, **10** (Suppl. 2), S1.

Sims,A.H. (2009) Bioinformatics and breast cancer: what can high- throughput genomic approaches actually tell us? *J. Clin. Pathol.*, **62**, 879–885.