BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 16 2005, pages 3439–3440
doi:10.1093/bioinformatics/bti525

Gene expression

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis

Steffen Durinck[1,2,*], Yves Moreau[1], Arek Kasprzyk[2], Sean Davis[3], Bart De Moor[1], Alvis Brazma[2] and Wolfgang Huber[2]

[1]Department of Electronical Engineering, ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, [2]EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Cancer Genetics Branch, National Human Genome Research Institute, National Institute of Health, 50 South Drive, Bethesda, MD 20892-8000, USA

Received on April 12, 2005; revised on May 25, 2005; accepted on May 31, 2005
Advance Access publication June 2, 2005

ABSTRACT

Summary: *biomaRt* is a new Bioconductor package that integrates BioMart data resources with data analysis software in Bioconductor. It can annotate a wide range of gene or gene product identifiers (e.g. Entrez-Gene and Affymetrix probe identifiers) with information such as gene symbol, chromosomal coordinates, Gene Ontology and OMIM annotation. Furthermore *biomaRt* enables retrieval of genomic sequences and single nucleotide polymorphism information, which can be used in data analysis. Fast and up-to-date data retrieval is possible as the package executes direct SQL queries to the BioMart databases (e.g. Ensembl). The *biomaRt* package provides a tight integration of large, public or locally installed BioMart databases with data analysis in Bioconductor creating a powerful environment for biological data mining.

Availability: http://www.bioconductor.org. LGPL

Contact: steffen.durinck@esat.kuleuven.ac.be

INTRODUCTION

Bioconductor is an open source and open development software project that provides a wide range of statistical and graphical tools based on R (Ihaka and Gentleman, 1996), for the analysis and comprehension of genomic data (Gentleman *et al*., 2004). These tools are distributed as separate but interoperable packages, each specializing in different subareas of analysis such as the 'affy' package to normalize Affymetrix chip data and the 'graph' package to handle graph data structures. BioMart (http://www.ebi.ac.uk/biomart) is a simple, federated query system designed specifically for use with large datasets. One of the major databases providing a BioMart database implementation is the Ensembl (Hubbard *et al*., 2005; Kasprzyk *et al*., 2004). Central in BioMart database systems is the concept of the star and the reverse-star schemas, of which the former consist of a single main table linked to different dimension tables and the latter is a variant (Kasprzyk *et al*., 2004). The overall simplicity of these schemas avoids complex joins and enables fast data retrieval. The *biomaRt* package is an add-on package for R that provides query ability to BioMart databases.

*To whom correspondence should be addressed.

DESCRIPTION

Our package currently covers four BioMart databases: Ensembl (Hubbard *et al*., 2005), a software system that produces and maintains automatic annotation on selected eukaryotic genomes; VEGA (Ashurst *et al*., 2005), the manually annotated Vertebrate Genome Annotation; dbSNP (Sherry *et al*., 2001), the Single Nucleotide Polymorfism database of NCBI and sequence mart, containing the Ensembl genome sequences. The package depends on the R package *RMySQL* and has been tested on Windows and Linux. After loading the library one can connect to either public BioMart databases or local installations of these. *biomaRt* offers several functions that enable the user to query these databases. One set of functions can be used to annotate identifiers such as Affymetrix, RefSeq and Entrez-Gene, with information such as gene symbol, chromosomal coordinates, OMIM and Gene Ontology. Alternatively, one can use a gene symbol as the starting point and query for the corresponding Affymetrix identifiers on a given chip. The queries can also have an inter-species nature and one can use an identifier of one type in species *a* to look up identifiers of the same or another type corresponding to homologs in species *b*. A second set of functions allow sequence-related data retrieval. Given a species and chromosome coordinates, one can retrieve genome sequences. This way a user can go directly from a set of differentially expressed genes to the upstream promoter sequences. Similarly, single nucleotide polymorphism (SNP) information can be retrieved. The SNP information is derived from dbSNP, which is mapped onto Ensembl.

USAGE

biomaRt provides documentation in the form of manual pages for every function and a vignette, which is an interactive document containing executable code chunks giving a more problem-oriented style of help.

EXAMPLES

A typical situation arising in the analysis of microarray data is that one has a list of identifiers corresponding to differentially expressed features on the array. In the example below, we first connect to the BioMart databases and retrieve gene information using an Affymetrix

identifier as the input. Then we use this information to retrieve the corresponding sequence.

```
mart <- martConnect()
gene <- getGene(id="1939_at",
        array="hg_u95av2", mart = mart)
seq <- getSequence(martTable = gene,
        mart = mart)
```

Another example is to sort different genes based on their chromosome coordinates, this could be used for investigation if the co-localized genes are also co-expressed.

A more advanced example could be a microarray analysis in *Drosophila*, where we want to focus on genes that have human homologs known to be involved in a certain disease. *biomaRt* enables one to first look up the human homologs, then using these homologs to query for OMIM identifiers. *Drosophila* genes that have human homologs with an OMIM identifier associated with them can then be selected for subsequent analysis.

DISCUSSION

The Bioconductor package *biomaRt* enables direct access from Bioconductor to BioMart databases such as Ensembl, creating a strong alliance for data analysis with biological databases. The current annotation packages available from Bioconductor are complementary to our package. They use precompiled annotation tables derived from the NCBI and stored as hashtables in R (Zhang *et al*., 2003). Precompiled annotation packages are convenient when working with one or a few array types with relative constant designs; however, this approach has limitations. When multiple chip designs are used in, for example, a meta-analysis study, different metadata packages need to be installed, that will contain redundant information. Very large gene sets make the metadata packages sizeable, while with *biomaRt*, only the annotation of the genes of interest is retrieved. *biomaRt* is more scalable as it gathers up-to-date information from BioMart databases. Fast data retrieval is possible as the *biomaRt* package executes direct SQL queries from R to the BioMart databases. Besides annotation information *biomaRt* also enables mapping of homologs and retrieval of sequence and SNP data, which can become part of a microarray data analysis. The biomaRt package will be further developed to include more BioMart databases and allow more complex types of queries. This tight integration of large public databases with data analysis in R provides a powerful platform for biological data mining.

ACKNOWLEDGEMENTS

The authors would like to thank Ewan Birney for the fruitful discussions on BioMart. FWO: PhD/postdoctoral grants, projects G.0115.01, G.0413.03, G.0388.03, G.0229.03, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-SQUAD, GBOU-ANA, GBOU-McKnow, STWW-Genprom; Belgian Federal Government: DWTC IUAP V-22; EU: FP5, CAGE, ERNSI; German Ministry for Education and Research through National Genome Research Network (NGFN) grant FKZ 01GR0450.

Conflict of Interest: none declared.

REFERENCES

Ashurst,J.L. *et al*. (2005) The Vertebrate Genome Annotation (VEGA) database. *Nucleic Acids Res.*, **33**, D459–D465.

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

Hubbard,T. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.

Kasprzyk,A. *et al*. (2004) Ensmart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Zhang,J. *et al*. (2003) An extensible application for assembling annotation for genomic data. *Bioinformatics*, **19**, 155–156.