BIOINFORMATICS

Importing MAGE-ML format microarray data into BioConductor

Steffen Durinck[†] [1], Joke Allemeersch[†] [1], Vincent J. Carey[2], Yves Moreau[1,3] and Bart De Moor[1]

[1] Department of Electronical Engineering, ESAT-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001, Leuven-Heverlee, Belgium, [2] Channing Laboratory, Brigham and Women's Hospital, 75 Francis Street, Boston, 02115, USA and [3] On leave at Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, Kemitorvet, Building 208, 2800, Lyngby, Denmark

ABSTRACT

The microarray gene expression markup language (MAGE-ML) is a widely used XML (eXtensible Markup Language) standard for describing and exchanging information about microarray experiments. It can describe microarray designs, microarray experiment designs, gene expression data, and data analysis results. We describe *RMAGEML*, a new Bioconductor package that provides a link between cDNA microarray data stored in MAGE-ML format and the Bioconductor framework for preprocessing, visualization, and analysis of microarray experiments.

Availability: http://www.bioconductor.org. Open Source.
Contact: joke.allemeersch@esat.kuleuven.ac.be; steffen.durinck@esat.kuleuven.ac.be
Keywords: cDNA microarray, MAGE-ML, Bioconductor, R

INTRODUCTION

Microarray data is generated in many different formats and often lacks standardized annotation and documentation. The MIAME (minimum information about a microarray experiment) guidelines define a set of fields for describing and annotating microarray data (Brazma et al., 2001). Major journals, such as *Nature, Cell,* and *the Lancet* have adopted these guidelines and have made submission of microarray expression data, coupled with MIAME-compliant documentation, compulsory for publication (DeFrancesco, 2002).

The MIAME recommendations define an information set, but do not address the formatting and exchangeability of microarray data. Formatting and exchangeability issues have been confronted in the microarray gene expression markup language (MAGE-ML), which is an XML standard for serializing data structured according to the MAGE object model (MAGE-OM) (Spellman et al., 2002). This data model incorporates the MIAME information and addresses many aspects of microarray experiment documentation and data encoding that are not covered by MIAME.

Bioconductor is an open source project that provides a framework for the statistical analysis of genomic data in R (Ihaka and Gentleman, 1996; Gentleman and Carey, 2003; Dudoit and Yang, 2003; Irizarry et al., 2003). Our Bioconductor package *RMAGEML* extracts information from MAGE-ML documents for cDNA microarray experiments and maps this information to Bioconductor R objects for the analysis of cDNA microarrays. The *RMAGEML* package implements a three-tiered architecture, transforming MAGE-ML to R objects via middleware that interfaces R to a Java software kit created expressly for working with MAGE-ML documents. The current version of *RMAGEML* transforms documents and data for cDNA experiments; work is in progress for a wider variety of microarray platforms.

DESCRIPTION

MAGE-ML is derived from the MAGE-OM model, which is an object model specification standardized by the Object Management Group (OMG) and maintained by the MGED Society. MAGE-OM consists of several classes of data structure. Examples of classes include `BioSequence`, `ArrayDesign`, `Array`, `Protocol` and `AuditAndSecurity`. The MGED society supplements the MAGE-OM with Software ToolKits (MAGEstk) in Java and Perl that specify data structure classes for programming within MAGE-OM, and serialization and deserialization between MAGE-OM structures and MAGE-ML.

The Bioconductor *RMAGEML* package is written in R and Java, and makes use of the Java-MAGEstk (http://mged.sourceforge.net/software/MAGEstk.php) application programming interface via the *SJava* interface (http://www.omegahat.org). *SJava* facilitates creation and manipulation of Java classes and methods in R as *references*. The overall architecture of *RMAGEML* minimizes the involvement of R with details of XML processing

04-104

Durinck S., Allemeersch J., Carey V.J., Moreau Y., De Moor B., ``Importing MAGEML format microarray data into BioConductor", *Bioinformatics*, vol. 20, no. 18, Dec. 2004, pp. 3641-3642., Lirias number: 72862.

and maximizes the use of MGED-sponsored software for coupling data structure and serialization processes to the MAGE-OM and MAGE-ML specifications. The *RMAGEML* package currently allows import to two Bioconductor data structures: `limma::RGlist` and `marrayClasses::marrayRaw`.

Information from different MAGE-ML packages is needed to create these objects. The DesignElement package contains a mapping of *Features* (which are the actual locations on the array) to *Reporters* (what is present at those locations). This package also provides a mapping from *Reporters* to their corresponding *BioSequence* references. These *BioSequences* are characterized by their name and database entries in the BioSequence package, which are mapped to the `genes` and `maGnames` slots of *limma* and *marray* respectively. The ArrayDesign package contains information on the layout of the array. From this package, we can derive the position of each *Feature* on the array in terms of *Zone* (block) and row and column within each *Zone*. This layout information is mapped to the `genes` slot of *limma* and a `maLayout` object of *marray*. The BioAssay package describes the different steps in the microarray experiment. The BioMaterial package describes the sample source and how it is labeled, this is mapped to a `maTargets` object of *marray* and the `targets` slot of *limma*. Finally, the BioAssayData package describes the feature references that were assayed and the available *QuantitationTypes*. It also contains the *BioDataCube*, which is a three dimensional matrix that stores the actual intensity data. This 3D-matrix is usually stored as slices of 2D-matrices in *ExternalData* files which, in the most commonly used ordering, contain the *DesignElementDimension* in the rows and *QuantitationTypes* in the columns. RMAGEML is capable of handling this data structure and maps it to the foreground and background intensity slots of BioConductor objects.

USAGE

The *RMAGEML* package has a manual and vignette describing its use. As an example, you can go to EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and query the database for the experiment with accession E-MEXP-5. Download the E-MEXP-5 MAGE-ML file, together with the MAGE-ML document describing the array used (accession A-MEXP-7). Place these files in a common directory `dir`, serving as the current working directory of R, with the *RMAGEML* package and all dependencies loaded. You can now import the E-MEXP-5 data to *limma* using the call `importMAGEML(dir, package="limma")`. An experiment of 18 hybridisations and 5184 spots takes 39 seconds to import on a 1.9GHz system with 256Mb RAM.

DISCUSSION

One of the first databases to make microarray data available in MAGE-ML format is ArrayExpress at the European Bioinformatics Institute (EBI) (Brazma et al., 2003). As MAGE-ML will become the standard format to exchange microarray data, development of tools that enable working with this format are highly in demand. RMAGEML will evolve such that it can exploit the huge amount of information contained in MAGE-ML format data and make possible to store analysis results back as MAGE-ML.

ACKNOWLEDGEMENTS

We thank Helen Parkinson for useful discussions on MAGE-ML and Kjell Petersen for help with the Java-MAGEstk API. Research supported by: Research Council KUL: GOA Mefisto 666, GOA-Ambiorics, IDO (IOTA Oncology, Genetic networks); Flemish Government: FWO: PhD/postdoc 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, Marie-Curie QLRI-1999-50595; USA NIH grant 1R33 HG002708.

REFERENCES

Brazma, A., P. Hingamp, J. Quackenbush, et al. (2001, Dec). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet 29*(4), 365–371.

Brazma, A., H. Parkinson, U. Sarkans, et al. (2003, Jan). ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res 31*(1), 68–71.

DeFrancesco, L. (2002, Oct). Journal trio embraces MIAME. *The Scientist*.

Dudoit, S. and J. H. H. Yang (2003). R packages for the analysis of cDNA microarray data. In G. Parmigiani, E. Garrett, R. Irizarry, and S. Zeger (Eds.), *The analysis of gene expression data: methods and software*, pp. 313–341. Springer, NY.

Gentleman, R. and V. J. Carey (2003). Visualization and annotation of genomic experiments. In G. Parmigiani, E. Garrett, R. Irizarry, and S. Zeger (Eds.), *The analysis of gene expression data: methods and software*, pp. 313–341. Springer, NY.

Ihaka, R. and R. Gentleman (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*(3), 299–314.

Irizarry, R., L. Gautier, and E. Cope (2003). An R package for oligonucleotide array analysis. In G. Parmigiani, E. Garrett, R. Irizarry, and S. Zeger (Eds.), *The analysis of gene expression data: methods and software*, pp. 313–341. Springer, NY.

Spellman, P. T., M. Miller, J. Stewart, et al. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol 3*(9), RESEARCH0046.