BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 2 2009, pages 279–280
doi:10.1093/bioinformatics/btn617

Gene expression

MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB

Tim F. Rayner[1,*], Faisal Ibne Rezwan[2], Margus Lukk[1], Xiangqun Zheng Bradley[1], Anna Farne[1], Ele Holloway[1], James Malone[1], Eleanor Williams[1] and Helen Parkinson[1]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton CB10 1SD and [2]University of Hertfordshire, College Lane Campus, Hatfield, Hertforshire AL10 9AB, UK

Received on October 7, 2008; revised and accepted on November 24, 2008

Advance Access publication November 27, 2008

Associate Editor: David Rocke

ABSTRACT

Summary: The MAGE-TAB format for microarray data representation and exchange has been proposed by the microarray community to replace the more complex MAGE-ML format. We present a suite of tools to support MAGE-TAB generation and validation, conversion between existing formats for data exchange, visualization of the experiment designs encoded by MAGE-TAB documents and the mining of such documents for semantic content.

Availability: Software is available from http://tab2mage.sourceforge.net/

Contact: tfrayner@gmail.com

1 INTRODUCTION

The data standards developed by the microarray community are among the most mature in the field of functional genomics. These standards include a convention on the disclosure of experimental data and annotation (MIAME; Brazma *et al*., 2001), an object model (MAGE; Spellman *et al*., 2002) and a data exchange format (MAGE-ML), and have been widely adopted and used successfully for several years. The MAGE-ML format is highly flexible; however, it is more complex than is typically required for data exchange between applications, and has not been widely accepted. The newly introduced MAGE-TAB format (Rayner *et al*., 2006), in contrast, is simpler, is human readable and can be opened and edited in any spreadsheet application. To support MAGE-TAB data exchange, demonstrate the utility of the format in dealing with high-throughput data, and make MAGE-TAB accessible to the bioinformatics community, we have developed a series of open-source Perl applications to address common MAGE-TAB use cases.

2 SOFTWARE OVERVIEW

As shown in Figure 1, MAGETabulator consists of a suite of tools that can be used singly or together to perform a variety of tasks: (i) preparation, syntactic and semantic validation of MAGE-TAB formatted data; (ii) visualization of investigation designs encoded in MAGE-TAB; (iii) conversion of MAGE-TAB documents to MAGE-ML format; (iv) conversion of NCBI Gene Expression Omnibus (GEO) SOFT records to MAGE-TAB format; and (v) support for automated *post hoc* addition of ontology terms to MAGE-TAB documents.

2.1 Preparation, validation and visualization of MAGE-TAB formatted data

The core components of MAGETabulator can be used to generate MAGE-TAB template documents which can be completed by the researcher, validate the syntax and check the content of the completed document, and parse the document and convert it into MAGE-ML. MAGETabulator thus facilitates the submission of MIAME-compliant data to public repositories which accept either MAGE-TAB or MAGE-ML documents.

2.1.1 Template generation and data submission Template generation is implemented using a MySQL database to store semantic relationships between standard experiment types from the MGED Ontology (Whetzel *et al*., 2006) and the sample annotation and experimental variables expected for each experiment type. The user can generate a MAGE-TAB template relevant to their particular organism, technology type and the relevant biological aspects of the system under study. For example, a template produced for an experiment studying development in mice would include fields for mouse strain, sex, developmental stage, age and so on. These relationships are fully configurable via the underlying database. MAGETabulator supplies a Ruby on Rails curation web interface to simplify the template configuration process. The database and template generation web forms can be used via ArrayExpress (at http://www.ebi.ac.uk/cgi-bin/microarray/magetab.cgi), or installed and used locally. Data submissions to ArrayExpress which use MAGE-TAB can accurately describe a wider range of experiment types compared with those submitted using the MIAMExpress web interface, and are usually quicker and easier to complete for larger experiments.

2.1.2 Experiment checker The MAGE-TAB specification describes a flexible format for its documents. MAGETabulator includes a validation script which confirms that a given MAGE-TAB document is syntactically correct, that the content is internally consistent and that it contains all the elements required by the MIAME guidelines. Any associated data files are also checked

[*]To whom correspondence should be addressed.

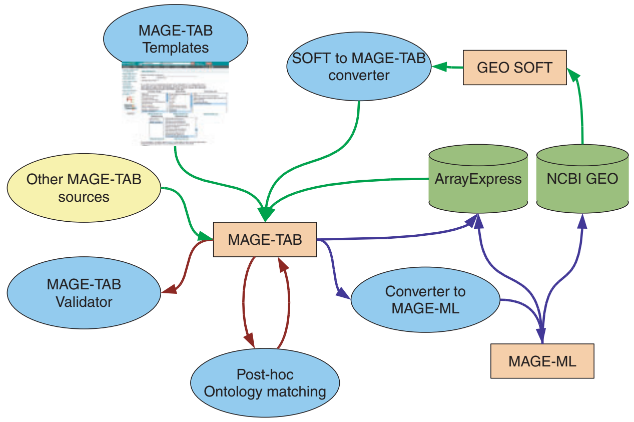

Fig. 1. Overview of MAGETabulator components. Green arrows represent processes generating MAGE-TAB, blue arrows show potential submission routes to data repositories and red arrows depict utility functions.

for errors. This validator script is written in modular object-oriented Perl, using a recursive-descent parser to decode the document. The code is readily extensible to support parsing of new data file types. Optionally, the script can also use Graphviz (http://www.graphviz.org/) to generate an experimental design graph to visualize the links between samples, hybridizations and data files.

2.2 Data exchange between applications

2.2.1 MAGE-TAB to MAGE-ML converter MAGETabulator provides a tool for converting a validated MAGE-TAB document into MAGE-ML format. The output is fully compliant with current best practices for encoding experimental metadata in MAGE-ML, and conforms to the ArrayExpress standard for data submissions. While the degree of semantic information expressable in MAGE-ML is greater than that for MAGE-TAB, in practice this extra flexibility is very rarely used. ArrayExpress uses MAGE-ML internally such that all annotation is retained from data submissions using either format.

2.2.2 SOFT to MAGE-TAB converter To promote data exchange between the GEO and ArrayExpress databases, we have created a pipeline application which can generate MAGE-TAB directly from a GEO Series record, as part of the GEOImport package. In its simplest form, fields in the SOFT file are mapped directly to their counterparts in a new MAGE-TAB document.

2.3 Semantic support for MAGE-TAB documents

The MAGE-TAB format consists entirely of tab-delimited text, and imposes no intrinsic restrictions on the terms used to annotate an experiment. We have therefore adopted a *post hoc* validation and ontology term matching strategy to provide machine-readable semantic content in the output documents.

2.3.1 GEOImport Annotation of sample characteristics and other experimental variables in the GEO database is heavily reliant on free text. To improve annotation consistency, MAGETabulator uses a tool that searches for candidate ontology terms within free text descriptions. The tool was implemented using the Java Finite Automata class library monq.jfa (http://www.ebi.ac.uk/Rebholz-srv/whatizit/software), and is incorporated into the GEO import pipeline.

2.3.2 Post hoc *additions to MAGE-TAB documents* The MAGE-TAB format supports encoding not only ontology term names, but also their identifiers/accession numbers and the ontologies from which they are derived. This information is optional, however, so MAGETabulator uses the 'Double Metaphone' phonetic algorithm (Philips, 990) to match terms used to annotate an experiment to terms from any ontology in OBO format (http://www.geneontology.org/GO.format.obo-1_2.shtml; Smith *et al*., 2007). The script reads in an input MAGE-TAB document and produces a new MAGE-TAB document including the matched term identifiers and ontology source.

3 DISCUSSION

The representation of high-throughput array-based data using simple spreadsheets has allowed us to develop a suite of tools that enable a biologist with Perl skills to manage these data. The tools have been developed, tested and used extensively at the ArrayExpress database and, since February 2008, have become the major route of submission to ArrayExpress.

The MAGETabulator project is publicly available on the SourceForge web site http://tab2mage.sourceforge.net, where the code is maintained in a Subversion repository and made available for download as part of the Tab2MAGE package. Researchers who generate and use array data will benefit from the free availability of tools that bring data submission within reach of a large segment of the community.

ACKNOWLEDGEMENTS

We acknowledge Midori Harris and Alvis Brazma for helpful comments on the article.

Funding: EMBL; EC grants EMERALD; MUGEN; FELICS; NHGRI; NIDDIB grants.

Conflict of Interest: none declared.

REFERENCES

Brazma,A. *et al*. (2001) Minimum information about a microarray experiment (MIAME) — toward standards for microarray data, *Nat. Genet.*, **29**, 365–371.

Philips,L. (1990) *Computer Language*, **7**.

Rayner,T.F. *et al*. (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data, *BMC Bioinformatics*, **7**, 489.

Smith,B.S. *et al*. (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat. Biotechnol.* **25**, 1251–1255.

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

Whetzel,P.L. *et al*. (2008) The MGED Ontology: a resource for semantics-based description of microarray experiments, *Bioinformatics* **22**, 866–873.