BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 10 2012, pages 1402–1403
doi:10.1093/bioinformatics/bts148

Data and text mining

Advance Access publication April 2, 2012

MageComet—web application for harmonizing existing large-scale experiment descriptions

Vincent Xue[1,*], Tony Burdett[2], Margus Lukk[3], Julie Taylor[2], Alvis Brazma[2] and Helen Parkinson[2,*]

[1]Department of Computer Science, Hunter College, City University of New York, NY 10065, USA, [2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, CB10 1SD and [3]Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Meta-analysis of large gene expression datasets obtained from public repositories requires consistently annotated data. Curation of such experiments, however, is an expert activity which involves repetitive manipulation of text. Existing tools for automated curation are few, which bottleneck the analysis pipeline.

Results: We present MageComet, a web application for biologists and annotators that facilitates the re-annotation of gene expression experiments in MAGE-TAB format. It incorporates data mining, automatic annotation, use of ontologies and data validation to improve the consistency and quality of experimental meta-data from the ArrayExpress Repository.

Availability and implementation: Source and tutorials for MageComet are openly available at goo.gl/8LQPR under the GNU GPL v3 licenses. An implementation can be found at goo.gl/IdCuA

Contact: parkinson@ebi.ac.uk or xue.vin@gmail.com

Received on August 26, 2011; revised on February 29, 2012; accepted on March 24, 2012

1 INTRODUCTION

The amount of experimental data in public repositories is accumulating at an ever increasing rate thanks to the advents of high-throughput technologies. Meta-analysis of these data are limited by the availability of harmonized, consistently annotated data in machine readable formats.

In the study of gene expression, MAGE-TAB is a commonly used file format that provides flexibility and structure for describing experimental data (Rayner *et al.*, 2006). Assays in the MAGE-TAB format can be explicitly annotated with characteristics and experimental variables, which in turn are referenced to ontologies to provide explicit meaning. Public repositories such as the ArrayExpress Archive (Parkinson *et al.*, 2009) house over 23 000 experiments in the MAGE-TAB format, which is also used by the TCGA project (Hampton, 2006).

The ability to perform meta-analyses requires considerable investment in human re-annotation, harmonization and mapping of data from different labs to single ontologies. Due to the spreadsheet nature of MAGE-TAB, it is a convenient bridge between readability and interoperability but meta-analysis is faster if formatting, validation, ontology enrichment and common editing tasks are addressed within the application. This addresses bottlenecks in the re-annotation process and allows users to extract implicit meta-data quickly from pre-existing MAGE-TAB documents and add to this in a convenient editing environment.

We present MageComet, a web application for curators that provides semi-automatic tools for annotating existing MAGE-TAB documents. We have used it to curate hundreds of datasets imported into ArrayExpress from GEO.

2 SOFTWARE COMPONENTS

2.1 Data input

MageComet is able to edit two of three files comprising a MAGE-TAB experiment. It accepts the IDF and SDRF files, which are both tab-delimited flat files containing high-level experiment information such as experiment design and variables and the SDRF containing per sample annotation often expressed as text. Curation begins in one of two ways: users can choose to upload an experiment's IDF and SDRF together, or load experiments via an ArrayExpress accession. Data is displayed to the user through MageComet, which functions as a spreadsheet editor.

2.2 Data summarization

MAGE-TAB files are conveniently compatible with any spreadsheet or text editor. However, these editors perform poorly when displaying the IDF and SDRF components in meaningful ways. Relevant experiment terms are burried in the text and considerable time is spent finding important fields, especially for large datasets with long text fields seen in GEO imported experiments.

MageComet handles this issue in two ways, the first being an ontology driven tag cloud. On loading an experiment, MageComet automatically data-mines the IDF and SDRF separately using a designated ontology as its dictionary. The default ontology used is the Experimental Factor Ontology (Malone *et al.*, 2010). It generates a tag cloud of terms weighted based on whether found in the IDF or SDRF (Fig. 1). The IDF contains less specific meta-data terms and therefore all terms mined are given a lower weight of one. Terms mined from the SDRF are given a weight of 2, and intersecting terms have a weight of 3, e.g. a submitter may provide the experimental context in the IDF by referring to a similar disease. If the disease

[*]To whom correspondence should be addressed.

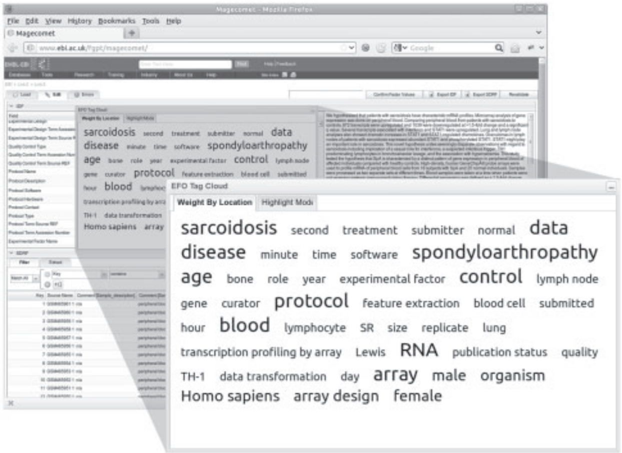
EFO Tag Cloud
Weight By Location
Highlight Modi
sarcoidosis second treatment submitter normal data
disease minute time software spondyloarthropathy
age bone role year experimental factor control lymph node
gene curator protocol feature extraction blood cell submitted
hour blood lymphocyte SR size replicate lung
transcription profiling by array Lewis RNA publication status quality
TH-1 data transformation day array male organism
Homo sapiens array design female

Fig. 1. An EFO tag cloud of accession E-GEOD-18781 as presented in MageComet. Useful terms such as sarcoidosis, spondyloarthropathy, blood, and Homo sapiens are emphasized to help users quickly annotate the text.

is found in the ontology and only referenced in the IDF, it will be represented in the tag cloud with a weight of one.

The tag cloud provides a visual semantic summary of the dataset and users grasp general idea of the experiment without in depth analysis. This is particularly useful when searching for similar experiments. Clicking on an element in the tag cloud highlights the corresponding term in the documents, allowing curators to confirm the context and explore the semantics of the meta-data independently of the document or ontology structure.

Second MageComet helps extract information via a multi-paneled view. SDRF and IDF editing panels are complemented by a third panel where values pertaining to the experimental design are grouped. MageComet hides columns that do not contain biological information, users can choose to unhide these. For example, the names of data files and protocols are hidden and biological annotations are prioritized.

2.3 Data extraction tools

MAGE-TAB can contain verbose comments or description columns—GEO derived data often contains factor values or sample annotations separated by a varied number delimiters, e.g. colons, semi-colons or commas. These require further curation to be informative for large-scale analyses. Separation of such columns is difficult to automate due to varied delimiters—a simplified ‘extract’ feature targets this by creating new columns based on what the user designates as the surrounding delimiters, without the user specifying a regular expression.

MageComet also has a specialized ‘filter and replace’ feature. Unlike common editors, MageComet gives users control of the source columns from which to filter from, and the target column to be replaced. This feature allows users to add factor values and characteristics to the SDRF based on existing annotations without repetitive text manipulation.

2.4 Validation

MageComet incorporates the Limpopo (limpopo.sf.net) library into its backend for syntactic and semantic validation. Curators can edit and quickly validate changes within the application, allowing for a seamless work environment.

2.5 Autocompletion

MageComet features an auto complete widget that helps curators tag samples with consistent annotations. Leveraging the extensive synonyms available in EFO, MageComet provides a query box which returns standard labels as defined in the ontology (i.e. Homo sapiens instead of human). With MageComet’s tight integration with ontologies, it can also complete the Term Source Ref and Term Source Number columns in the SDRF for richer annotations.

3 IMPLEMENTATION

MageComet was designed as a web application to provide cross platform compatibility without installation on client machines. As a web application, it is able to synchronize with the latest ontologies and validators to ensure consistent annotations. Its integration into the web browser makes external services easily accessible. As a service, MageComet provides validation through the Limpopo parsers. It provides data mining through Whatizit and interacts with ontologies using the OntoCat library (Adamusiak *et al.*, 2011).

4 CONCLUSION

MageComet is a web application created for harmonization of large-scale experiment annotations. It is designed to reduce the repetitive aspects of re-annotation and curation of pre-existing documents and serve as a consistent, up-to-date workspace across all platforms. Its features are designed to provide user-driven automation of repetitive tasks. It differs from standard editors through it’s close integration with the MAGE-TAB specification. The implementation and code can be found at goo.gl/IdCuA, and goo.gl/8LQPR, respectively.

Funding: The authors thank the Macaulay Honors College at CUNY Hunter College for funding for V.X., Duncan Odom from CRUK for funding M.L., and EMBL for J.T., H.P., and A.B.

Conflict of Interest: none declared.

REFERENCES

Adamusiak,T. *et al.* (2011) OntoCAT–simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics*, **12**, 218.

Hampton,T. (2006) Cancer Genome Atlas. *J. Am. Med. Assoc.*, **296**, 1958–1958.

Malone,J. *et al.* (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112.

Parkinson,H. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

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