Egas – Collaborative Biomedical Annotation as a Service

David Campos [1*], Joni Lourenço[1], Tiago Nunes[1], Rui Vitorino[2], Pedro Domingues[2], Sérgio Matos[1*] and José Luís Oliveira[1]

[1]IEETA/DETI, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[2]Chemistry Department, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

*Corresponding author: Tel: +351 234 370 500, E-mails: {david.campos,aleixomatos}@ua.pt

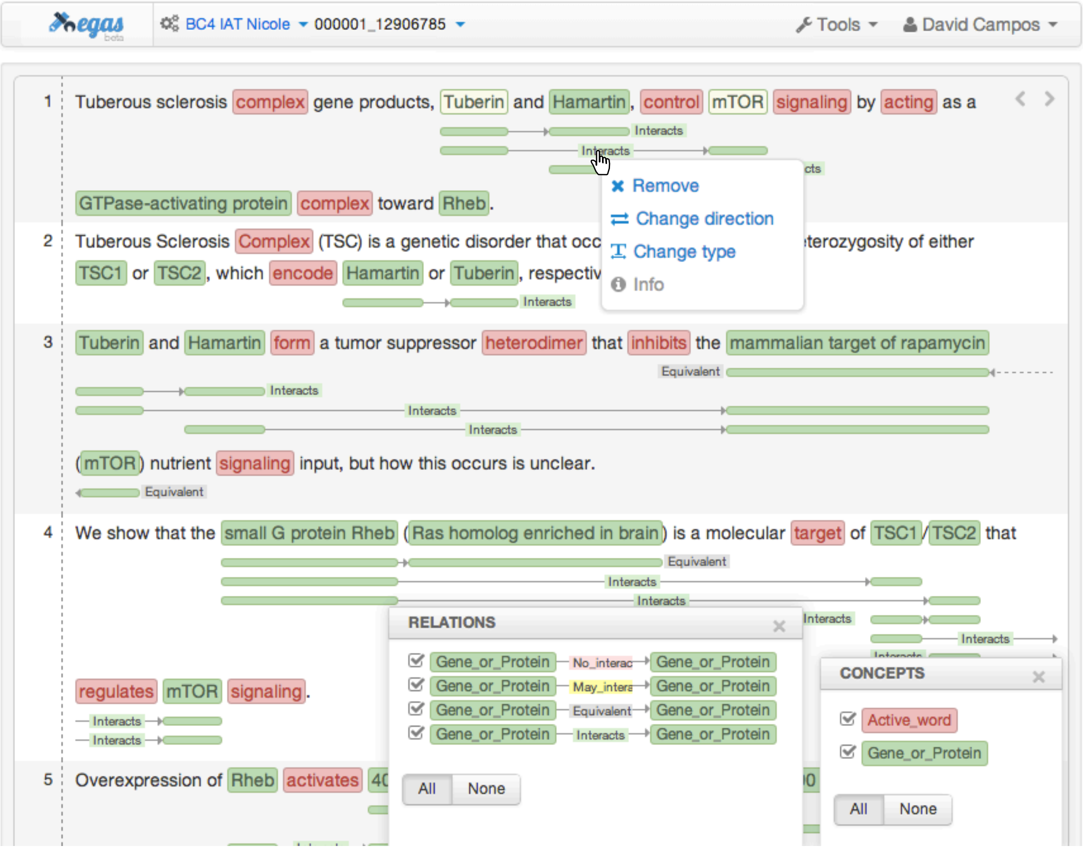
Abstract

In this paper we present Egas, a web-based platform for biomedical text mining and collaborative curation. The web tool allows users to annotate texts with concept occurrences as well as with relations between concepts. Annotations may be imported together with the documents using one of the accepted input formats, or may be added during the annotation process, either manually or by calling a document annotation service. Users can further inspect, correct or remove automatic text mining results, add new annotations, and export the results to standard formats.
The tool is available at http://bioinformatics.ua.pt/egas and is compatible with most recent versions of Google Chrome, Mozilla Firefox, Internet Explorer and Safari.

Introduction

Egas (**Figure** 1) is a web-based platform for biomedical text mining and collaborative curation. It allows users to annotate texts with occurrences of concepts and relations between these concepts. The annotation tool follows what we termed an “annotation-as-a-service” paradigm. Document collections, users, configurations, annotations, back-end data storage, as well as the tools for document processing and text mining, are all managed centrally. This way, a curation team can use the service, configured according to their requisites, taking advantage of a centrally managed pipeline.

The tool is based on the idea of *Projects*. A *Project* consists of a curation or document annotation task, performed on a collection of documents, by a team of (one or more) curators, and considering a pre-defined set of concept and relation types defined by the curation guidelines. A project administrator is responsible for managing the users (curators) associated to the project and the project characteristics, such as annotation guidelines and target concepts and relations. Projects may be public or private, in which case they are only accessible by users that have been added by the project manager.


egas beta
BC4 IAT Nicole
000001_12906785
Tools
David Campos
Interacts
Equivalent
Remove
Change direction
Change type
Info
RELATIONS
☑ Gene_or_Protein No_interac Gene_or_Protein
☑ Gene_or_Protein May_intera Gene_or_Protein
☑ Gene_or_Protein Equivalent Gene_or_Protein
☑ Gene_or_Protein Interacts Gene_or_Protein
All
None
CONCEPTS
☑ Active_word
☑ Gene_or_Protein
All
None

1 Tuberous sclerosis complex gene products, Tuberin and Hamartin, control mTOR signaling by acting as a
GTPase-activating protein complex toward Rheb.

2 Tuberous Sclerosis Complex (TSC) is a genetic disorder that occ terozygosity of either
TSC1 or TSC2, which encode Hamartin or Tuberin, respectiv

3 Tuberin and Hamartin form a tumor suppressor heterodimer that inhibits the mammalian target of rapamycin
(mTOR) nutrient signaling input, but how this occurs is unclear.

4 We show that the small G protein Rheb (Ras homolog enriched in brain) is a molecular target of TSC1/TSC2 that
regulates mTOR signaling.

5 Overexpression of Rheb activates 40

Figure 1: Egas main user interface.

To create the document collection for a project, three import options are available:

- Local – from a local collection of documents;
- Remote – from remote resources using a list of identifiers;
- Search – from remote resources by submitting search queries.

For each Project, the project administrator can freely define the relevant concept and relation types, according to the requisites of the task. To facilitate the annotation work, each different concept and relation type is associated with a markup color. A relation type is defined by specifying the types of the intervening concepts and assigning a name to the relation. For example, for protein-protein interactions, after defining a concept type “Protein”, an “Interacts” relation can be defined between two concepts of type “Protein”.

Curators can start from the raw text and add the concept and relation annotations as they review the documents, or they can start from preprocessed texts, containing automatically identified concepts and relations that they will revise. This can be achieved by importing a previously processed document collection or by using integrated concept and/or relation extraction services to pre-process a set of documents in the collection.

System Description

Project management

Project management allows project administrators to specify configurations of the annotation task, such as:

- Project: manage project information and annotation guidelines;
- Users: manage curators;
- Concepts: manage concepts to annotate;
- Relations: manage relations to annotate;

Through the project panel, the administrator can indicate which curators are allowed to annotate the documents associated with the project. Moreover, the project administrator can also provide a brief description of the annotation task, and upload documents describing the guidelines of the annotation task, which are accessible to the curators associated with the project.

Concept and relation management allows adding or removing target concept types and defining relations between those concepts, as well as selecting the associated markup colors.

Adding documents to a project

Importing documents from the client machine supports RAW, A1 [1] and BioC [2] formats. Regarding the A1 format, if corresponding annotation files are provided, both the text of the documents and any concept and relation annotations are imported to the database. Importing documents from remote resources supports both PubMed and PubMed Central, through a list of identifiers. The corresponding documents are loaded from the remote resource and displayed to the users, so they can select which ones to import. Likewise, users can submit a query to search either PubMed or PubMed Central. In this case, the search results are obtained from the remote resource and displayed to the users for selection.

Exporting project documents

Exporting project documents to an external resource supports both A1 and BioC formats. Such feature allows users to store the generated information locally, in order to add it to a local knowledge base or for using in text mining pipelines, for instance.

Automatic concept and relation mining

Egas provides an interface that allows using external automatic annotation tools that are available as web-services. For instance, users can automatically annotate a document with specific concepts and respective relations, and then correct the provided annotations. It currently integrates an automatic service for protein-protein interactions (PPIs) annotation, providing the following annotations: *a)* protein concepts; *b)* relations between proteins (PPIs); *c)* relations marking equivalent protein mentions (e.g. acronyms and long forms); and *d)* active words that may indicate the presence of PPIs. The service is implemented on top of Neji [3], using Gimli [4] to perform machine learning-based protein name recognition. BioThesaurus [5] is used to normalize recognized names, through the application of prioritized dictionary matching [3]. Equivalent protein relations are added using a simple abbreviation resolution technique, and PPIs are recognized through a rule-based approach using dependency-parsing trees.

Annotation interface

Figure 1 shows the tool's main user interface. The central box displays the content of the text being curated, showing the concepts and relations that have been identified. Concepts are shown as colored boxes, using the colors defined in the project configuration. Relations are shown as lines, tagged with the relation type. The colored boxes connected by the relation markup are placed under the concepts that participate in the relation, and are colored with the same color as the respective concept, making it easy to identify the entire relation. Moreover, hovering the cursor over the relation markup also highlights the involved concepts.

The boxes on the lower right corner allow curators to select the concept and relation types they want to appear highlighted in the text.

During the curation task, concepts and relations can be added, edited or removed. Hovering the mouse over an annotation shows the corresponding semantic type and, by right-clicking, a menu opens that allows removing the annotation. A new concept annotation is added by selecting a text span in the annotation window. This opens a concept type selection box for choosing the concept type for the new annotation.
To add a relation, the user clicks the first concept in the relation while pressing the "Alt" key, and then clicks the second concept also while pressing the "Alt" key. Relations are considered directional, so the order in which the concepts are added to a relation is important. For example, if a relation "promotes" is defined, the order needs to be considered. As for concepts, right-clicking over an existing relation allows removing it. In the same menu, the user can easily change the relation type and/or direction (Figure 1).

Implementation

Text-processing modules, such as the concept and relation annotation services, were implemented in Java, the article fetching modules were also built in Java, and the web interface was developed using HTML5, CSS3, and JavaScript, in order to allow fast processing of large documents and support mobile devices. The resulting information, such as annotations and relations, is stored in a relational database. Finally, all database operations are performed using secured RESTful web-services, allowing easy integration with mobile devices, such as smartphones and tablets.

Case Study – the BioCreative IV IAT task

The proposed curation task consists in the identification and extraction of biomolecular events described over PubMed abstracts related to neuropathological disorders, including protein-protein interactions, protein expression and post-translational modifications. To create the corpus for this task, a collection consisting of more than 135 thousand PubMed abstracts was first obtained with the PubMed search:

"Neurodegenerative Diseases"[MeSH Terms] OR "Heredodegenerative Disorders, Nervous System"[MeSH Terms] AND hasabstract[text] AND English[lang]

The documents were then ranked according to their relevance for extracting protein-protein interactions, using a SVM classifier trained on the BioCreative III PPI Article Classification Task data [6]. Finally, the top-ranked 100 documents were selected for the task.

Four curators were selected, and each was assigned 50 documents from the corpus to curate. Curators were asked to annotate 25 of their assigned documents using the available PPI annotation service described above, and the remaining 25 documents without using this service, in order to assess its impact on curation effort. In the first case, curators had to revise the automatically generated annotations, correcting any erroneous concept or relation annotation and adding missing ones. In the second case, curators had to annotated all mentions of protein names and all protein interactions described in each document The tool recorded the time taken by each curator to curate each document, as well as the number of annotated concepts and relations.

Conclusion

A tool for collaborative document annotation and curation is proposed. The tool allows teams of curators to work on a shared curation project, following a set of configurable concept and relation types. The curation task can be performed over a collection of raw text documents or by reviewing automatic concept and relation annotations, obtained either with the included concept and relation identification service or through external annotation tools. Documents can be imported in raw text, A1 and BioC formats, and the final annotations may be exported to A1 and BioC formats. Apart from the local import option, it is also possible to create a document collection by importing from PubMed and PubMed Central either through a list of identifiers or by submitting a search to these services.

We are currently working on adding real-time collaboration features, providing instant feedback of users' interactions within a document. Thus, multiple users can change a document at the same time, showing exactly who changed what. A project chat will also be available, allowing users to discuss details of the annotation task.

The tool is available at http://bioinformatics.ua.pt/egas and is in active development by the Bioinformatics group at the University of Aveiro, Portugal, aiming to provide an annotation-as-a-service solution through a flexible, configurable and user-friendly environment.

Funding

This work was supported by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-010029 (FCT reference PTDC/EIA-CCO/100541/2008), FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013. S. Matos is funded by FCT under the Ciência2007 programme.

References

1. Standoff format - brat rapid annotation tool [http://brat.nlplab.org/standoff.html].
2. Comeau DC, Islamaj Doğan R, Ciccarese P, Cohen KB, Krallinger M, Leitner F, Lu Z, Peng Y, Rinaldi F, Torii M, Valencia A, Verspoor K, Wiegers TC, Wu CH, Wilbur WJ: BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)* 2013, 2013:bat064.
3. Campos D, Matos S, Oliveira JL: A modular framework for biomedical concept recognition. *BMC Bioinformatics* 2013, 14:281.
4. Campos D, Matos S, Oliveira J: Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics* 2013, 14:54.
5. Liu H, Hu ZZ, Zhang J, Wu C: BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 2006, 22:103–105.
6. Krallinger M, Vazquez M, Leitner F: The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* 2011, 12:S3.