Textpresso text mining: semi-automated curation of protein subcellular localization using the Gene Ontology's Cellular Component Ontology

Textpresso Group
California Institute of Technology
Mail Code 156-29
Pasadena, CA 91125

Introduction

Manual curation of experimental data from the biomedical literature is expensive and time-consuming; however, most biological knowledge bases still rely heavily on manual curation for data extraction and entry. We have developed and actively use a category-based information retrieval and extraction system for curating *C. elegans* proteins to the Gene Ontology's Cellular Component Ontology. The system's core components are the Textpresso full text database and index, a support vector machine (SVM) classifier for the presence of expression pattern data in a paper as well as a specifically-designed semantic category search. All automatically extracted information is presented to a curator for validation in a user-friendly web-based interface.

System description

The method is successful because authors describe subcellular localization in a sufficiently stereotypical manner. Stereotypical language can be used to empirically create new Textpresso categories specific for retrieval of sentences relevant to GO cellular component curation. Details about this method can be found in (1).

Figure 1 describes the curation system in detail. The Textpresso database (2) is updated every night with the bibliography and full text of new papers pertaining to *C. elegans* research. They are subsequently classified for the existence of

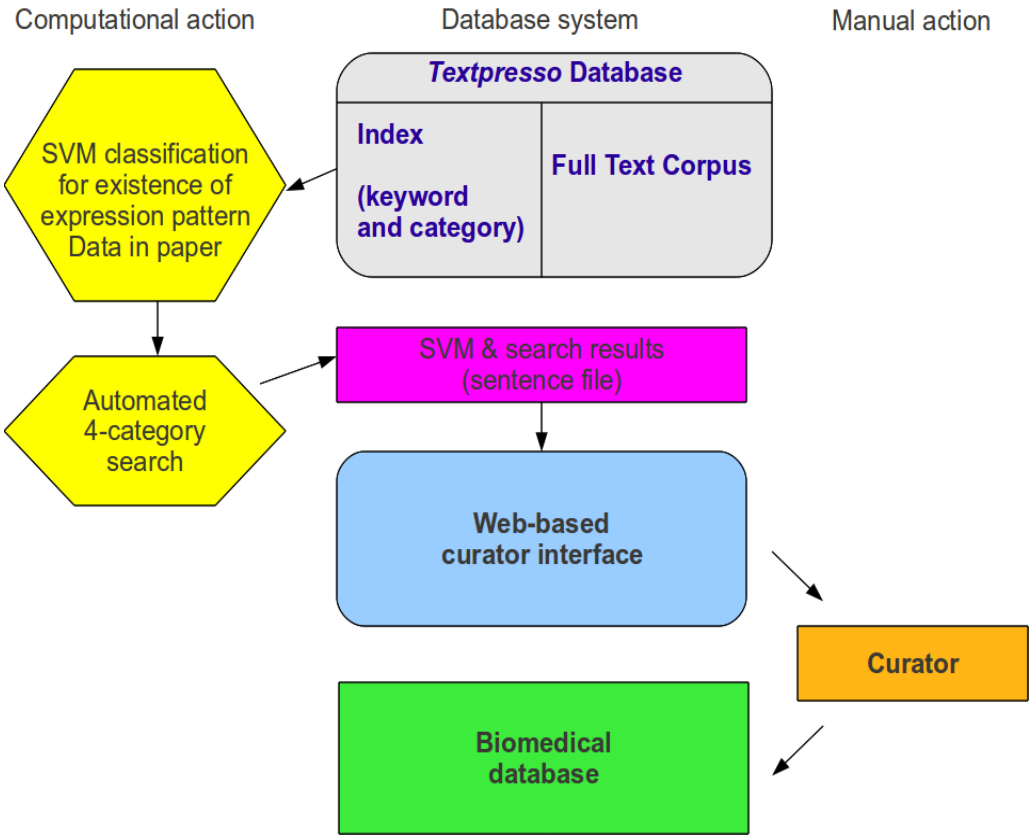

Figure 1: Schema of the curation system for GO Cellular Component Curation.

expression patterns by an SVM trained by a set of ~1000 *C. elegans* papers containing expression patterns. Papers predicted by the SVM to contain this data type are then subjected to a specialized Textpresso category query. Using a training set of sentences that describe results of localization experiments in the published literature, we generated three new curation-task specific categories (cellular components, assay terms, and verbs) containing words and phrases associated with reports of experimentally determined subcellular localization. The Textpresso query searches the full text of the pre-filtered articles for sentences containing terms from each of the three new categories plus the name of a *C. elegans* protein. The results of this query are then stored in a file that contains the sentences that match the query, the paper ID and matched query items (protein, cellular component, assay term and verb). The curator can access the results stored in the file via a web-based curation interface, which is displayed in Figure 2. The curation form allows for inspection of the retrieved sentences and pre-populates data fields with entries extracted from the sentence. If a similar annotation has been made in the past, the form suggests GO annotation terms. The curator can take several steps, from adding an annotation to the biomedical database to marking a falsely made extraction as false-positive, as not go-curatable (e.g. the sentence describes localization in a mutant background), as a scrambled sentence (an artifact from erroneous pdf to text conversion) or as already curated.

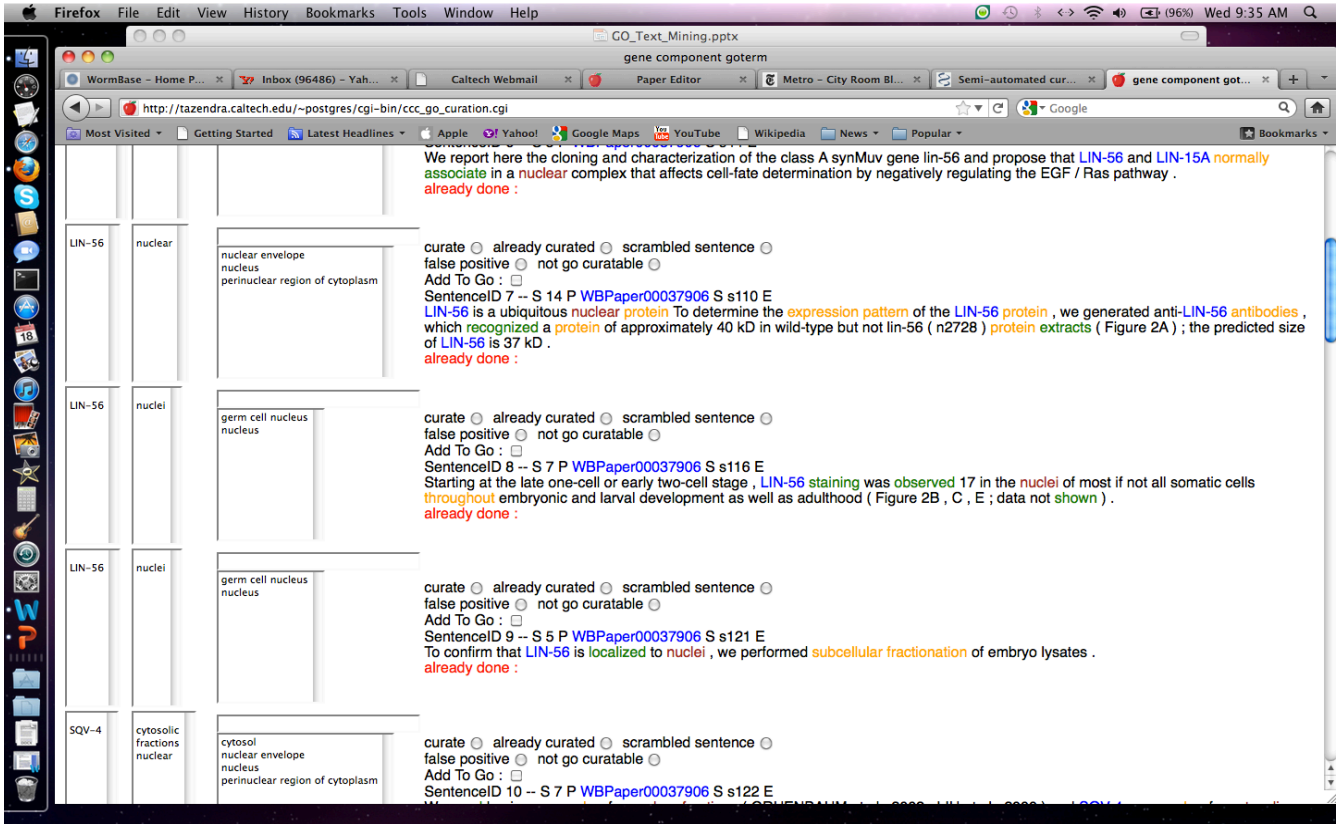

Figure 2: Curation interface to validate and extract GO Cellular Component Curation.

Relevance and impact

The system is currently being used by WormBase for GO cellular component curation. An automated pipeline sends the curator an e-mail if new sentences have been identified by SVM and the 4-category Textpresso query. Other groups using this approach include the Arabidopsis Information Resource (TAIR), who use this curation workflow for the same curation purposes on a corpus that is updated every 6 months, and plans are underway to implement GO cellular component curation for FlyBase and dictyBase. WormBase extended this approach to mining macromolecular interactions as well as finding orthologs of human disease genes.

Adaptability

As the system is already used by other model organism databases and for other data types, it has proven its adaptability. The biggest challenge in adapting the curation workflow lies in the fact that the specific Textpresso categories need to be modified to retrieve sentences of interest for (a) a new model organism or (b) a different data type. New training sentences

need to be retrieved manually and analyzed for word and phrase frequency so meaningful categories and corresponding lexica can be formed.

Interactivity

All curator activities are performed via a web-interface. Should further information be required to make a curational decision, the Textpresso search engine can be accessed via an interface to post additional keyword and/or category queries to the full text of all or specific papers. All other steps in the pipeline are automated via cronjobs, but could theoretically also be controlled via a simple web-interface.

Performance

We evaluated the system on three levels; on the document level, the Textpresso only system (no SVM) yields a recall of 95.2% and a precision of 66.7%. In this case, a true positive is a paper that contains a sentence describing subcellular localization of a protein, although the machine may or may not have picked the correct sentence for curation. When SVM is included in the process, recall drops to 76%, but precision increases to 80%. When evaluating the system on a sentence level (i.e., a true positive is a correctly identified sentence describing subcellular localization), the precision is 80.1%, but recall is only 30%. However, this recall is rescued to 66.2% when looking at the annotation level, i.e., how many of all possible annotations could be made from a set of papers. This is because the same information gets repeated across a paper or a corpus. There are no data yet on how the SVM step changes recall and precision on a sentence level.

Benchmark

For evaluation of the system we will provide a set of 50 documents possibly containing protein subcellular localization information. We intend to split up the set into two subsets of 25 documents each. The first subset will be curated by biocurators in a purely manual manner, measuring precision and recall on document, annotation and sentence level. We will measure the time it takes a biocurator to finish curating a paper to show how the system can improve curation efficiency. For the second subset, we will provide the same interface that WormBase curators use, except that it will be loaded with sentences that have been retrieved through the SVM classifier and the Textpresso 4-category search of this subset. Again, the curator will validate or reject the computational output with the help of the interface, and the time it takes to do this will be measured. Together with recall and precision of the output of this interactive step we will be able to compare recall, precision and curation efficiency (in terms of time) with the purely manual curation of the first subset. In all cases the required output will be the paper IDs of all papers containing protein subcellular localization, the sentences of each paper from which an annotation can be made as well as the annotation itself.

References

1. Van Auken K, Jaffery J, Chan J, Müller HM, Sternberg PW. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation (2009). BMC Bioinformatics. 2009 Jul 21;10:228.

2. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature (2004). PLoS Biol. 2004 Nov;2(11):e309. Epub 2004 Sep 21.