Data and text mining

Applying GIFT, a Gene Interactions Finder in Text, to fly literature

Núria Domedel-Puig* and Lorenz Wernisch

School of Crystallography, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, UK

Received on July 21, 2004; revised on July 4, 2005; accepted on July 6, 2005
Advance Access publication July 12, 2005

ABSTRACT

Summary: A number of freely available text mining tools have been put together to extract highly reliable Drosophila gene interaction data from text. The system has been tested with The Interactive Fly, showing low recall (27–34%), but very high precision (93–97%).

Availability: The extracted data and a web interface for submission of texts to GIFT are available at http://gift.cryst.bbk.ac.uk/gift

Contact: n.domedel_puig@cryst.bbk.ac.uk

Supplementary information: Additional documentation, such as the dictionaries and the reference sets, are available at the GIFT website.

1 INTRODUCTION

Genetic networks can be inferred from experimental data (Davidson et al., 2002) using various statistical and mathematical techniques (Pournara and Wernisch, 2004). However, for the successful reconstruction of large networks, prior information on gene interactions is often required. We present a strategy for extracting reliable gene interaction information from scientific literature. The data retrieved can then be incorporated as prior knowledge in gene network reconstruction. Achieving high precision is the key in this context. The biological text mining community has traditionally focused on Medline abstracts (Blaschke et al., 1999). For GIFT, The Interactive Fly, a hand-curated repository of high quality fly gene information (Brody, 1999), has been used.

2 METHODS

Our approach can be divided in the following steps:

(1) Initial text filtering. The Interactive Fly (hereafter referred to as TIF) was downloaded from the FlyBase site (The FlyBase Consortium, 2003) and transformed into plain text. Sentence boundaries were annotated, and those containing conditionals or negations were discarded to avoid confusion.

(2) Part-of-speech (POS) tagging. A POS tagger is a tool that identifies the grammatical categories of words in a corpus. These were required in two different steps of this application: during the gene annotation process and later, in the information extraction step. Here, the probabilistic tagger, TreeTagger (Schmid, 1994) was used.

(3) Gene annotation. Although other work has focused on extracting interactions involving user-specified genes or proteins (Blaschke et al., 1999), this tool aims at extracting interactions involving any two known fly genes. Fly gene name recognition faces a number of problems: Drosophila gene names do not follow a standard notation, each gene can be referred to using different synonyms; gene names are often identical to English words or widely used acronyms, and it was not possible to rely on the convention that genes are written in lowercase italics.

In this project, two different gene dictionaries were utilized. First, a gene list derived from the gene description file headers in TIF, where each gene alias uniquely maps to one gene ID. Second, a larger dictionary was obtained by merging all fly gene synonyms available at FlyBase to the TIF list. Ambiguous gene aliases do occur in this dictionary, and thus annotated genes often show a one-to-many relationship with gene IDs. Gene names in the corpus were detected with a method that resolved semantic ambiguity according to the POS of the words. For example, English words were only accepted as gene names when they have an appropriate POS (e.g. not verbs, not prepositions). Words that erroneously passed this filter usually failed to satisfy the requirements described in step 4 below. Examples illustrating this process, as well as more details about the dictionaries are available at the GIFT website.

(4) Corpus querying by CBW. The IMS Corpus Workbench (Christ, 1994) is a package for full-text retrieval from large textual resources. It provides a query language, CQP, capable of performing pattern matching on one or more words, taking into account their POS tags and their lemmas (i.e. their roots). A set of queries was created to extract gene interaction information from TIF. In particular, queries were designed to match any substring of the corpus that satisfied the following conditions: the information (a) is found within a sentence, (b) has a particular grammatical structure, (c) includes words from a pre-specified list of verbs and nouns and (d) involves at least two gene names. The three main grammatical structures allowed were, the active and passive forms of lexical verbs, and the use of the verb to be, in a number of tenses. The pre-specified list of allowed words consisted of both verbs and nouns frequently found in interaction descriptions, such as activate, inhibit, regulate, bind and inducer, repressor, mediator, target, amongst others. This list was drawn from the manual inspection of scientific texts by molecular biology experts. The queries were also designed to discriminate between four broad interaction types: those denoting activation (e.g. A activates B) and inhibition (e.g. A inhibits B), those referring to a neutral relation (e.g. A regulates B) and those showing a direct interaction between elements (e.g. A is a target of B). These lead to a total of 22 different queries, which are available in two versions (relaxed or stringent) at the GIFT website.

(5) Data storage and public access: the extracted information was stored in a PostgreSQL database. This can be accessed through the web interface at the GIFT website.

3 RESULTS

A total of 4010 matches to queries were found applying the relaxed query set to the text annotated with the short, specific dictionary.
These matches describe 1284 unique interactions, each consisting of a pair of genes with an interaction type. The large dictionary-annotated text, in combination with the stringent query set, lead to 4450 matches to queries. A number of unique interactions cannot be provided in this case owing to the ambiguous nature of the dictionary.

To assess the performance of GIFT, recall and precision were calculated comparing the results of the method against a gold standard. This reference set is a manual compilation of the sentences that contain interaction information from the original TIF text of four different genes, namely CycE, teashirt, arrow and futsch, and is available as Supplementary material. Initially, sentences containing more than two gene names were excluded. The recall (precision) values obtained with the short dictionary-annotated corpus are 34% (97%) with the relaxed query set, and 25% (100%) with the stringent query set. The corresponding values for the text annotated with the ambiguous gene dictionary are 37% (86%) and 27% (93%). For more details on the performance, also on sentences with more than two gene names, see the Supplementary material.

4 DISCUSSION

The major bottlenecks found in text mining are illustrated in this work. First, owing to the lack of gene name standardization, high quality named entity recognition methods are needed. Two different dictionaries have been used here, which unexpectedly lead to very similar recall values: a short, TIF-specific dictionary, and a large, ambiguous dictionary. The former would most probably prove inadequate for parsing fly texts from other sources. The second dictionary should be used for such corpora. The ambiguous gene aliases it contains were not filtered out because this step eliminates a number of important gene names [e.g. e2f] is a synonym for both e2f (FBgn0011766) and e2f/2 (FBgn0024371), despite usually referring to FBgn0011766]. Rather, each annotated alias is automatically linked to all gene IDs that share the same alias. Many English words (e.g. cell, cycle, complex, arrow) and acronyms (e.g. ap as the gene aperous or the acronym for antero-posterior) become potential gene names when using this large dictionary, a phenomenon that confuses the parser. This requires a more efficient gene annotation and information extraction process which is achieved here by using a more stringent query set, i.e. a query set where the context in which gene names are allowed is considerably restricted.

Second, there is a trade-off between query stringency and recall. As expected, very general queries increase recall at the expense of an increased false positive rate. Relaxed queries work reasonably well only when using a highly specific gene dictionary, such as the one from TIF. However, more specific queries are required in order to keep the precision high when the dictionary is less specific. In general, the approach is conservative, and relies on the fact that, if there is a significant support for an interaction, it will appear repeatedly throughout the corpus and chances are high that it will use one of the simple structures represented in the queries. In the special case of poorly documented genes, the user is always encouraged to switch to the low stringency query set.

Unfortunately, the recall and precision obtained with different methods are not directly comparable (BioCreAtIvE, 2004). Moreover, the interaction data extracted in this work are the result of two different tasks: gene name annotation and information extraction. For reference, a more sophisticated approach to extract gene interaction information from FlyBase was implemented by Proux et al. (2000), obtaining 44% recall and 81% precision. Although the conservative strategy in GIFT shows a lower recall, we are able to successfully keep precision >93% (for the large dictionary with stringent queries), which will allow us to use these interactions in network reconstruction.

Finally, this method ignores across-sentence information and anaphors. The biological context in which interactions occur is not explicitly retrieved either. To help validate the extracted information, each interaction is linked to the original sentence and, therefore, a final validation is always left to user judgement. Future work will focus on extending the current method to cope with information from other species, and by parsing other sources of published information. Finally, we will consider the use of fly-specific named entity recognition methods. A POS tagger specifically trained to deal with biological texts will be implemented as soon as they allow for XML annotation and offer high speed tagging.

ACKNOWLEDGEMENTS

We thank Thomas Brody for providing information that greatly facilitated parsing The Interactive Fly. N.D. is funded by a Functional Genomics programme grant of the Wellcome Trust.

Conflict of Interest: none declared.

REFERENCES