ABSTRACT

Summary: BioIE is a rule-based system that extracts informative sentences relating to protein families, their structures, functions and diseases from the biomedical literature. Based on manual definition of templates and rules, it aims at precise sentence extraction rather than wide recall. After uploading source text or retrieving abstracts from MEDLINE, users can extract sentences based on predefined or user-defined template categories. BioIE also provides a brief insight into the syntactic and semantic context of the source-text by looking at word, N-gram and MeSH-term distributions. Important applications of BioIE are in, for example, annotation of microarray data and of protein databases.

Availability: http://umber.sbs.man.ac.uk/dbbrowser/bioie/

Contact: divoli@bioinf.man.ac.uk

INTRODUCTION

Owing to the large volume of biomedical literature and its continuous fast growth, the need for text-mining tools has become increasingly important. In recent years, several different systems have been developed: some aim to detect interactions among proteins, genes or both (Wong, 2001; Hoffmann and Valencia, 2004); others specifically detect protein and gene names (Hirschman et al., 2002; Yu et al., 2002); other, more specialized, systems extract information relating to, for instance, gene expression profiling (MedMiner; Tanabe et al., 1999), drugs and genes relevant to cancer (EDGAR; Rindflesch et al., 2000), signal-transduction pathways and associated drugs and diseases (GENIES; Friedman et al., 2001) and c-DNA clones (FACTS; Nagashima et al., 2003).

In spite of the range of text-mining tools available, what has been lacking is a tool providing generic, easily customizable information extraction across a range of subjects of general interest to biologists and specifically to database annotators. Here we describe a new system that allows different types of sentence extraction, using predefined categories of interest relating to proteins, plus custom extraction around different entities and concepts, together with statistical feedback on the source and extracted text.

IMPLEMENTATION AND ARCHITECTURE

BioIE is a rule-based system, implemented in Perl and CGI, and accessible via the Web. Its architecture is shown in Figure 1.

Fig. 1. Overview of BioIE's architecture.
are also reported in the descending order of the number of times they occur in the text. Finally, for retrieved MEDLINE abstracts, the MeSH terms for which the abstracts are indexed are reported in descending order of the number of times they are found, together with their frequency per 100 MeSH terms. These options provide some insight into the syntactic and semantic content of the source-text and may be valuable for further more detailed studies.

**Sentence extraction**

Sentence extraction is based on manually defined templates and rules. Templates may be single words, word pairs or small phrases (which may or may not be contiguous). Templates and rules were chosen carefully, using domain-specific knowledge, aiming at high precision rather than recall. Currently, BioIE uses five predefined categories of interest relating to proteins: structure, function, diseases and therapeutic compounds and localization and familial relationships.

BioIE was designed to extract sentences because they are grammatically complete entities that are usually more informative than windows of words around target words/phrases: e.g. researchers may want to know about an event, but also about the conditions under which the event takes place—whole sentences are more likely to contain such information, and are small enough to be checked quickly.

Currently, there are three extraction options: the first extracts sentences containing the selected templates; the second extracts sentences containing one or more user-specified terms in addition to the templates; the third allows users to provide their own keywords or phrases, for those interested in information beyond the predefined templates. All returned sentences carry the PMID (when the source is from abstracts) or the line number (when users have uploaded their own text), linking them to the original source.

The extracted sentences are ranked in the order of importance, according to the number and the type of template words and phrases they contain; these and user-specified words, are highlighted in the output, making it easy for users to evaluate the results. The word, filtered word and N-gram distributions of the extracted text may also be calculated; this makes it easy to compare word usage in template-extracted sentences and the initial source-text, which could be useful for more in-depth linguistic analysis.

**APPLICATIONS**

BioIE provides several extraction options. The predefined categories, based on protein families, their structure, function and disease relationships, together with the custom extraction option, make it useful both as a generic text-mining tool and as an annotation tool, e.g. for microarray data or protein databases (InterPro, Mulder et al., 2003; PRINTS, Attwood et al., 2003), annotators of which currently have to manually trawl the literature to be able to compose an abstract for each family.

**FUTURE WORK**

BioIE has been designed as a decision support tool rather than to be fully automated. It is thus fast and simple to use. The system is highly interactive and users can customize it to best suit their needs. For instance, for the top ranked results of the user-specified extraction, BioIE achieves 100% precision for most of the protein entities. Currently, the system does not yet deal adequately with synonyms and homonyms, but we plan to provide options to return synonyms for some protein entities in future. We will also further revise the templates and rules, and add new categories of interest, and are exploring the ways to provide summarization options.

**ACKNOWLEDGEMENTS**

This work was funded by Ipharmatica Ltd. and BBSRC. We are grateful to Neil Maudling and Alex Mitchell for their suggestions on the system implementation.

**REFERENCES**

Friedman, C. et al. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics, 17(Suppl. 1), S74–S82.