

*Data and text mining***ALIBABA: PubMed as a graph**Conrad Plake¹, Torsten Schiemann¹, Marcus Pankalla², Jörg Hakenberg^{1,*} and Ulf Leser¹¹Knowledge Management in Bioinformatics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany and ²Department of Mathematics and Computer Science, Free University Berlin, Arnimallee 2-6, 14195 Berlin, Germany

Received on April 28, 2006; revised on May 31, 2006; accepted on July 22, 2006

Advance Access publication July 26, 2006

Associate Editor: Satoru Miyano

ABSTRACT

The biomedical literature contains a wealth of information on associations between many different types of objects, such as protein–protein interactions, gene–disease associations and subcellular locations of proteins. When searching such information using conventional search engines, e.g. PubMed, users see the data only one-abstract at a time and ‘hidden’ in natural language text. ALIBABA is an interactive tool for graphical summarization of search results. It parses the set of abstracts that fit a PubMed query and presents extracted information on biomedical objects and their relationships as a graphical network. ALIBABA extracts associations between cells, diseases, drugs, proteins, species and tissues. Several filter options allow for a more focused search. Thus, researchers can grasp complex networks described in various articles at a glance.

Availability: <http://alibaba.informatik.hu-berlin.de/>**Contact:** hakenberg@informatik.hu-berlin.de**1 INTRODUCTION**

Most information on biological entities and their interactions is available only in textual form. Since searching a text database is less precise than searching a structured database, many efforts are under way to automatically analyze texts to identify and extract relevant facts. The extracted information may be used to answer precise queries, to summarize multiple texts in a single representation and to connect to other sources of knowledge. Although the level of detail and accuracy of the extracted data currently cannot reach that of the original text, automatic information extraction is a valuable tool for navigating text databases and for offering quick overviews, both important tasks in many stages of research (Jensen *et al.*, 2006).

Since the most complete source of citations in biomedicine is PubMed, a number of projects use PubMed abstracts as an input for the analysis. iHOP offers access to the underlying literature by means of a network of concurring genes and proteins (Hoffmann and Valencia, 2005). Users access the information by searching for gene names. In contrast, ALIBABA evaluates arbitrary queries. EBIMed provides a quick overview of co-occurrences of a variety of entities: proteins, species, drugs and gene ontology (GO) terms. It searches all PubMed abstracts that fit an arbitrary user query and presents the resulting associations in tabular form (Kirsch *et al.*, 2005). ALIBABA provides a graphical view and offers more advanced association mining. GoPubMed searches GO terms in PubMed abstracts and links them to the GO hierarchy, which can then be used to navigate the result set (Doms and Schroeder, 2005).

All of the above applications present their results in the form of hyperlinked texts or tables. With ALIBABA, we present a system that graphically visualizes information on associations between biological entities extracted from a PubMed search result. Another distinctive feature is its text mining method for identifying and classifying associations that yields higher precision than a pure collocation analysis.

ALIBABA uses Java Web Start to launch a client from any web browser. This client handles only the visualization of the network. It sends a user’s query to the server, which forwards the query to PubMed, retrieves the matching abstracts and processes them. The server then returns the annotated abstracts to the client, which builds a network out of all annotations.

2 USING ALIBABA

ALIBABA’s screen consists of three regions, as shown in Figure 1. The upper horizontal input field accepts queries to PubMed. As an option, the results can be limited to a maximum number of citations. The ordering within a result set is the same as retrieved from PubMed. It is also possible to append the results of a query to those of previous queries, which can be used to build a network incrementally.

The large window shows the graph as it results from parsing the abstracts returned for the query. Nodes represent biological entities, with different colors for different classes. Edges represent associations between two entities. Whenever ALIBABA was able to assign a source/target-dependency to the relation, this is indicated by directed edges; arrows point from source to target. Undirected edges represent associations for which ALIBABA could not identify such a dependency (Section 3). The gray value of an edge correlates with its assigned confidence score, where darker edges represent more confident relations. The search field at the bottom helps to find specific objects within the graph.

The right-hand window contains more detailed information on all items in the graph. The upper part shows all extracted entities and their interaction partners categorized by biological classes (proteins, cells, tissues, etc.), arranged in a tree. After clicking a node, more information on this entity is shown in the lower part, including synonyms encountered in the abstracts with links to external databases (UniProt, MeSH, NCBI Taxonomy, MedlinePlus, PubMed), as well as sentences from abstracts mentioning this entity.

To view information on a specific relation, users select an associated partner of an entity from the upper tree view. The lower part then contains information on both entities and detailed information on the selected relation including its specific type and subtype (such as modification or activation). It will also show the textual evidence, i.e. all

*To whom correspondence should be addressed.

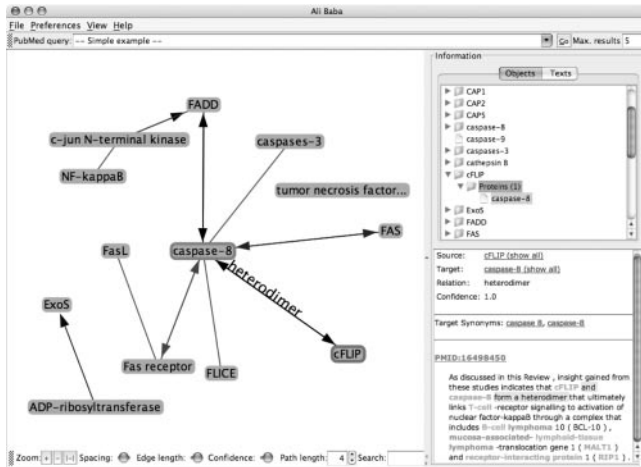


Fig. 1. (Parts of) the graph resulting from five PubMed abstracts for the query ‘FADD’. Information on the selected protein ‘caspase-8’ is given in the right panel, for instance, association partners and evidence texts.

sentences found in the abstracts that discuss the picked relation. Relevant objects in the text are highlighted, and all sentences are back-linked to the respective PubMed abstract.

Filtering result sets

The complete graph view can be altered by setting filter options from the preferences menu. ALiBABA offers to choose which entity classes to display, to hide unconnected entities and to set a minimum confidence value of visible associations. In addition, the user can choose to aggregate cells associated with proteins. Aggregated cells are visualized as bubbles containing their subcellular proteins as shown in Figure 2. Furthermore, associations can be restricted to co-occurrences and/or matching patterns, where patterns would overwrite co-occurrences between the same entities.

3 INFORMATION EXTRACTION

ALiBABA uses a dictionary-based approach for recognizing biomedical objects (Kirsch *et al.*, 2005). Dictionaries consist of regular expressions depicting terms and spelling variations. We collected the dictionaries from different sources (aforementioned databases). To find associations between entities, ALiBABA uses two different techniques in parallel: pattern matching and co-occurrence filtering. Pattern matching uses language patterns extracted from annotated, task-specific corpora (Hakenberg *et al.*, 2005). Such language patterns resemble regular expressions using tokens, part-of-speech tags and entity classes. The pattern matching algorithm also provides a confidence score for each relation, depending on the quality of the match between the sentence and a pattern. Furthermore, it identifies the type of the association and, in many cases, the direction. ALiBABA uses such patterns to extract protein–protein interactions and cellular locations of proteins. The extraction module achieves an F1-measure of 61% (maximum recall of 52% at 75% precision), as evaluated on the SPIES corpus (Hao *et al.*, 2005). To ensure higher recall, we also search for concurring entities, i.e. entities co-occurring in the same sentence. ALiBABA currently finds associations between two proteins (or genes), proteins and cells, diseases, species and tissues, as well as drugs and diseases. On average, ALiBABA parses one abstract per second.

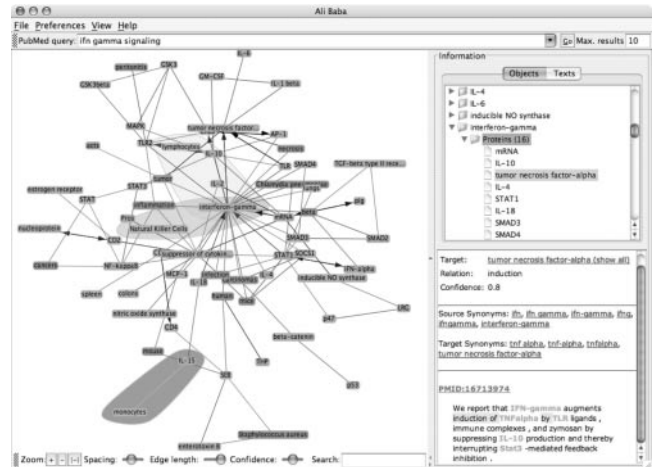


Fig. 2. A more complex query for ‘ifn gamma signaling’. Cells are represented as bubbles that contain their respective associated partners. For more examples ‘see the ALiBABA website’, including explanations.

4 DISCUSSION

ALiBABA is an easy-to-use application for browsing biological networks extracted on-the-fly from results of PubMed queries. We chose to forward queries to PubMed due to its elaborated search options and the familiarity of many biologists with PubMed searches. The main features of our systems are (1) the automatic summarization of information from all abstracts matching a query into a network; (2) the graphical and interactive display of the extracted information; (3) the provision of links to external databases; and (4) text mining methods more sophisticated than co-occurrence filtering. All these features build on an information extraction module using dictionaries and pattern matching, identifying a variety of different biological entities and relations between them. To leverage the sometimes erroneous results of the extraction module, users can filter results based on objects and relationships, confidence scores and extraction methods.

ACKNOWLEDGEMENTS

The Knowledge Management in Bioinformatics Group is a member of the Berlin Center for Genome Based Bioinformatics (BCB). This work is supported by the German Federal Ministry of Education and Research (BMBF) under grant contract 0312705B.

Conflict of Interest: none declared.

REFERENCES

Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33**, W783–W786.
 Hakenberg,J., Plake,C., Leser,U., Kirsch,H. and Rebholz-Schuhmann,D. (2005) Genic interaction extraction with alignments and finite state automata. In *Proceedings of the Learning Language in Logic Workshop (LLL’05)*, Bonn, Germany.
 Hao,Y. *et al.* (2005) Discovering patterns to extract protein–protein interactions from the literature: Part II. *Bioinformatics*, **21**, 3294–3300.
 Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**, ii252–ii258.
 Jensen,L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
 Kirsch,H. *et al.* (2005) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Inform.*, **75**, 496–500.