BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 7 2009, pages 977–978
doi:10.1093/bioinformatics/btp081

Data and text mining

OnTheFly: a tool for automated document-based text annotation, data linking and network generation

Georgios A. Pavlopoulos[1,*], Evangelos Pafilis[1], M. Kuhn[1], Sean D. Hooper[2] and Reinhard Schneider[1]

[1]Structural and Computational Biology Unit, EMBL Meyerhofstrasse 1, Heidelberg, Germany and [2]Department of Energy Joint Genome Institute (DOE-JGI), Genome Biology Program, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

Received on November 19, 2008; revised on February 2, 2009; accepted on February 7, 2009

Advance Access publication February 17, 2009

Associate Editor: Jonathan Wren

ABSTRACT

OnTheFly is a web-based application that applies biological named entity recognition to enrich Microsoft Office, PDF and plain text documents. The input files are converted into the HTML format and then sent to the Reflect tagging server, which highlights biological entity names like genes, proteins and chemicals, and attaches to them JavaScript code to invoke a summary pop-up window. The window provides an overview of relevant information about the entity, such as a protein description, the domain composition, a link to the 3D structure and links to other relevant online resources. OnTheFly is also able to extract the bioentities mentioned in a set of files and to produce a graphical representation of the networks of the known and predicted associations of these entities by retrieving the information from the STITCH database.

Availability: http://onthefly.embl.de, http://onthefly.embl.de/FAQ.html

Contact: pavlopou@embl.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Finding information about a biological entity is a step tightly bound to molecular biology research. Despite ongoing efforts, this task is both tedious and time consuming, and tends to become a challenge as the amount of information increases steadily. Currently available systems like Whatizit (Rebholz-Schuhmann *et al.*, 2008), allow the user to paste a free text section into a web page, which then links the recognized biological entities to existing databases. Our aim is to assist researchers with an easy-to-use interface which provides them with summary information about the biological entities mentioned in commonly used document types. In this applications note, we present OnTheFly, a tool that allows automated tagging of proteins, genes and chemicals and interaction network generation from widely used files like PDF, Microsoft Office files, as well as plain text files. In the following sections, we describe the functionality and the architecture of OnTheFly and we comment on its performance. We then demonstrate the functionality using a full text PDF article as an example.

2 FUNCTIONALITY

OnTheFly is a service to automatically annotate document files such as Microsoft Word, Excel, Power Point, PDF or plain text files. After submitting the files to the service, the system returns a tagged HTML version of the documents. Gene, protein and chemical names are highlighted and by clicking on them the user activates a pop-up window which contains relevant information about the entity. The presented information includes domains, sequence, organism, sub-cellular localization for proteins, formula for chemicals and protein–chemical and chemical–chemical interactions for both entity types. This functionality is provided by the Reflect server (http://reflect.ws).

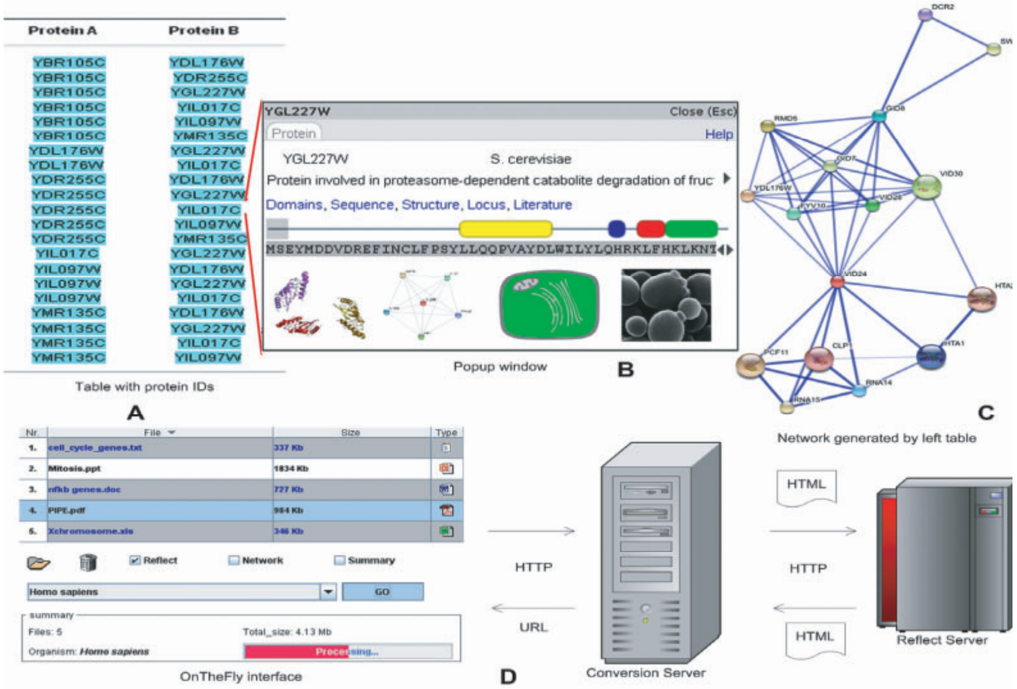
OnTheFly can furthermore generate interaction networks for a set of bioentities (genes, proteins, chemicals) extracted from the STITCH database (Kuhn *et al.*, 2008). The user can select the preferred organism whose protein aliases will be used for the tagging and network generation; the default organism is set to *Homo sapiens*. The size of the network and the number of interactors per recognized entity can be manually defined by the user. The network generation is not restricted to one document but can be applied to a set of documents simultaneously.

Lists, summarizing the identified bioentities are also generated. These lists contain the ID of the bioentities together with the organism and description. These summary results contain information about bioentities found in the set of the selected files.

The performance of the service can be assessed in a number of ways, such as the quality of the document conversion, the time required to tag a document and the accuracy of the annotation. The used file converters are able to maintain most of the layout of the documents, including column separation, tables and figures. The time to process a full text article of about 15 pages with images and tables ranges typically between 15 to 20 s. This time includes the whole process including the communication with the server.

The name tagging performance of the Reflect server is comparable to other available methods. More information can be found under the FAQ section on the web server.

[*]To whom correspondence should be addressed.



Protein A	Protein B
YBR105C	YDL176W
YBR105C	YDR255C
YBR105C	YGL227W
YBR105C	YIL017C
YBR105C	YIL097W
YBR105C	YMR135C
YDL176W	YGL227W
YDL176W	YIL017C
YDR255C	YDL176W
YDR255C	YGL227W
YDR255C	YIL017C
YDR255C	YIL097W
YDR255C	YMR135C
YIL017C	YGL227W
YIL097W	YDL176W
YIL097W	YGL227W
YIL097W	YIL017C
YMR135C	YDL176W
YMR135C	YGL227W
YMR135C	YIL017C
YMR135C	YIL097W

Table with protein IDs

YGL227W Close (Esc)

Protein Help

YGL227W S. cerevisiae

Protein involved in proteasome-dependent catabolite degradation of fruc

Domains, Sequence, Structure, Locus, Literature

MSEYMDDVDREFINCLFPSYLLQQPVAYDLWILYLQHRKLFHKLKNT

Popup window

Nr.	File	Size	Type
1.	cell_cycle_genes.txt	337 Kb	
2.	Mitosis.ppt	1834 Kb	
3.	nfkb genes.doc	727 Kb	
4.	PIPE.pdf	984 Kb	
5.	Xchromosome.xls	346 Kb	

☑ Reflect ☐ Network ☐ Summary

Homo sapiens GO

summary

Files: 5 Total_size: 4.13 Mb

Organism: *Homo sapiens* Processing...

OnTheFly interface

Fig. 1. The Figure shows an annotated table (**A**) of an PDF full text article (Pitre *et al*., 2006), the generated pop-up window with information about the protein YGL227W (**B**) and an automatically generated protein–protein interaction network (**C**) of associated entities for the proteins shown in part (**A**). For demonstration purposes, we isolated the table from the pdf file and processed the table separately. (**D**) The architecture and the functionality. Files get uploaded to OnTheFly server and they get converted into HTML. Reflect server annotates the HTML file and sends back the annotated HTML to OnTheFly server. A user can drag and drop files in the OnTheFly applet. The 'GO' button sends the selected documents to the conversion server that converts the according file formats into HTML pages, which will then be sent to the tagging server. A URL pointing to the generated HTML document is returned. The organism selection drop-down list enables users to define a species protein dictionary to be used by default. The 'Network' and 'Summary' option will extract the STITCH derived networks of associations of the recognized entities in the document(s) and produce a summary page listing the recognized entities.

To demonstrate the functionality of OnTheFly a full text article on protein–protein interaction predictions (Pitre *et al*., 2006) stored locally as a PDF file, has been processed. Figure 1A below shows a table section of the resulting HTML file with the tagged protein identifiers. Figure 1C shows the corresponding automatically retrieved association network of these entities using the STITCH database.

3 METHODS

OnTheFly uses the client-server architecture shown in Figure 1. The front end is an Applet written in Java 1.5. It can be accessed directly either from the server web page or as a Java desktop application. The Applet technology was chosen as it allows file drag-and-drop functionality and thus maximizes the ease of use. The server side components consist of a set of document converters along with the software modules required to invoke the Reflect service for tagging. The commercially available document converters currently employed (ultrashareware 2008; verypdf 2008) were selected based on their ability to maintain the layout of the original document, availability of a command line modus and their processing speed. Any data exchange is based on the HTTP protocol.

4 CONCLUDING REMARKS

We hope the *OnTheFly* service provides a powerful tool for researchers in the life science field. It is an attractive tool, not only for readers of scientific literature, but also for data annotators and experimentalists who want to link their in-house documents to literature and other biological databases.

Conflict of Interest: none declared.

REFERENCES

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

Pitre,S. *et al.* (2006) PIPE: a protein–protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.

Rebholz-Schuhmann,D. *et al.* (2008) Text processing through Web services: calling Whatizit. *Bioinformatics* **24**, 296–298.

ultrashareware. (2008) Ultra PPT To HTML Converter. Available at http://www.ultrashareware.com/Ultra-PPT-To-HTML-Converter.htm.

verypdf. (2008) PDF to HTML v2.0. Available at http://www.verypdf.com/pdf2htm.