OnTheFly 2.0: a tool for automatic annotation of files and biological information extraction.

Evangelos Pafilis, Georgios A. Pavlopoulos, Venkata P. Satagopam, Nikolas Papanikolaou, Heiko Horn, Christos Arvanitidis, Lars Juhl Jensen, Reinhard Schneider, Ioannis Iliopoulos

***Abstract*— Retrieving all of the necessary information from databases about bioentities mentioned in an article is not a trivial or an easy task. Following the daily literature about a specific biological topic and collecting all the necessary information about the bioentities mentioned in the literature manually is tedious and time consuming. OnTheFly 2.0 is a web application mainly designed for non-computer experts which aims to automate data collection and knowledge extraction from biological literature in a user friendly and efficient way. OnTheFly 2.0 is able to extract bioentities from individual articles such as text, Microsoft Word, Excel and PDF files. With a simple drag-and-drop motion, the text of a document is extensively parsed for bioentities such as protein/gene names and chemical compound names. Utilizing high quality data integration platforms, OnTheFly allows the generation of informative summaries, interaction networks and at-a-glance popup windows containing knowledge related to the bioentities found in documents. OnTheFly 2.0 provides a concise application to automate the extraction of bioentities hidden in various documents and is offered as a web based application. It can be found at: http://onthefly.embl.de, http://onthefly.med.uoc.gr or http://onthefly.hcmr.gr**

The project was financed by the European Molecular biology Laboratory (EMBL-Heidelberg), Heidelberg, Germany. EP has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 264089 (MARBIGEN project), HH by the Novo Nordisk Foundation Center for Protein Research and GAP by the European Commission FP7 programme 'Translational Potential' (TransPOT; EC contract number 285948).

Evangelos Pafilis is a postdoctoral researcher at the Institute of Marine Biology, Biotechnology and Aquacultures (IMBBC), Hellenic Center for Marine Research (HCMR), Heraklion Crete, Greece (phone: +30 6977 057150; e-mail: pafilis@hcmr.gr).

Georgios A. Pavlopoulos is a postdoctoral researcher at the University of Crete, Medical School, Heraklion, Crete, Greece (e-mail: g.pavlopoulos@med.uoc.gr).

Venkata P. Satagopam is a bioinformatician at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg (e-mail: venkata.satagopam@uni.lu).

Nikolas Papanikolaou works at the University of Crete, Medical School, Heraklion, Crete, Greece (e-mail: papnikol@med.uoc.gr).

Heiko Horn works at the The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark (e-mail: heiko.horn@cpr.ku.dk)

Christos Arvanitidis is a principal investigator at the Institute of Marine Biology, Biotechnology and Aquacultures (IMBBC), Hellenic Center for Marine Research (HCMR), Heraklion Crete, Greece (e-mail: arvanitidis@hcmr.gr).

Lars Juhl Jensen is a Professor at the The Novo Nordisk Foundation Center for Protein Research, Uni of Copenhagen, Denmark (e-mail: Lars.juhl.jensen@cpr.ku.dk).

Reinhard Schneider is the head of the bioinformatics core facility at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg (e-mail: reinhard.schneider@uni.lu).

Ioannis Iliopoulos is a Lecturer at the University of Crete, Medical School, Heraklion, Crete, Greece (e-mail: iliopj@med.uoc.gr).

I. Introduction

VERY often, researchers need to cope with the task of manually sifting through large amounts of text in the form of published articles in order to obtain pertinent information regarding proteins, genes and compounds. The manual approach is summarized as follows; i) read through the article text, ii) detect bioentities of interest, and iii) query one or more databases to get the relevant information. Such a procedure can be very tedious and time consuming. Due to the exponential growth of literature abstracts and data files produced by experiments, automated extraction of knowledge about bioentites stored in texts or abstracts is not a trivial task [1, 2].

Most of the current text mining applications are based on a top-to-bottom approach, meaning that vast amounts of text, such as PubMed abstracts, are initially parsed for interesting phrases or words and then disseminated downwards through databases or software such as STRING [3], STITCH [4, 5], Biotextquest [6], Bioalma [7], Whatizit [8, 9] or iHop [10]. Other tools, reviewed in [11], try to analyze literature using semantic concepts, clustering algorithms or ranking algorithms and most generic attempts such as [12] go one step further by using semantic web concepts to further enrich and annotate terms found in literature.

OnTheFly 2.0 together with its ancestor [13] aim to simplify the extraction of knowledge from vast amounts of literature by automating the following steps. First, a user loads one or more files into the application. The content of these files is subsequently extracted and searched against an extensive dictionary of known bioentities such as gene/protein names, drug names or Wikipedia terms, recording any matches. Finally, biological information for these bioentities is retrieved from relevant databases. The results are presented as a richly annotated and interactive HTML version of the original article.

OnTheFly 2.0 utilizes a drag-and-drop interface to maximize ease of use. The application currently supports the following file formats; Microsoft Word, Excel, Adobe PDF and plain text. OnTheFly 2.0 provides a bridge between the various identifiers used by different databases, greatly reducing the need for manual cross-referencing. The application is designed with the help of biologists and is easy to use, offering non-obtrusive and practical help for computer-based scientific reading. It is available at http://onthefly.embl.de, http://onthefly.med.uoc.gr or http://onthefly.hcmr.gr.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

II. METHODS AND RESULTS

A. *Functionality*

The OnThefly service is an online web application to automatically annotate and extract information from document files such as Microsoft Word, Excel, Power Point, PDF or plain text files. While OnTheFly is based on its sister project Reflect [14] and was initially presented as an application note in [13], it currently comes with significant advantages. The functionality of OnTheFly 2.0 service is described below in steps and is extensively presented in Figure 1.

1) Initially a user drags and drops a single file or a set of documents in the web interface.

2) Commercially available converters such as *ultrashareware* (http://www.ultrashareware.com/Ultra-PPT-To-HTML-Converter.htm) and *verypdf: PDF to HTML v2.0*. (http://www.verypdf.com/pdf2htm), convert the documents into HTML, trying to keep the original formatting.

3) The content of the converted documents is queried word-by-word against a dictionary based on the STRING database [15], which hosts the relevant information about the respective gene, protein and chemical names and their relevant identifiers in public databases.

4) After the information about the bioentities is collected, OnTheFly 2.0 offers functionality to generate interaction networks through the STITCH [16] database, informative summaries about the bioentities based on a local database, interactive summary popup windows from Reflect [14] service and linked information from the unpublished BioCompendium platform (http://biocompendium.embl.de).

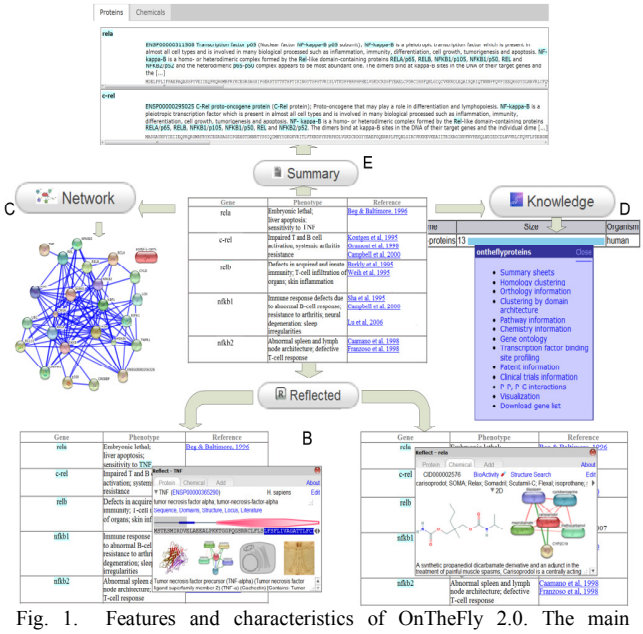

Fig. 1. Features and characteristics of OnTheFly 2.0. The main functionality of the OnTheFly 2.0 service. A. Document file containing a Nf-κB regulatory network. B. **Reflected Button:** generation of popup windows through the Reflect service. C. **Network Button:** protein-protein interaction network generation through the STITCH database. D. **Knowledge Button:** collection of information from public databases through the Biocompendium service. E. **Summary Button:** informative summaries containing short descriptions about the bioentities.

The number of files per session for OnTheFly 2.0 is currently limited to 15 documents. In accordance with the supported file formats, we accept the file extensions *.cvs, .txt, .pdf, .doc(x), .xlsl(x)*. Each file should have a unique file name and a size not larger than 2MB. Users can select one or more files per session and perform different analyses as explained below:

Reflect option: individual tabs in the default browser contain the annotated content of the document in HTML format. While mouse-hovering, users can see at-a-glance information about the highlighted bioentities in a popup window.

Summary option: a new tab containing information about all of the proteins (description, annotation and sequence) or small molecules (full names and descriptions) mentioned in the selected documents will appear as a list.

Network option: a protein-protein, protein-chemical or chemical-chemical interaction network is generated according to the information available in the STITCH [16] database.

Knowledge option: OnTheFly is currently compatible with the unpublished BioCompendium data integration platform. The recognized protein, gene and small molecule names from one or more documents are subjected to further bioinformatics analysis. Such analysis includes:

- Comprehensive knowledge collection from different biological databases for a given list(s) of genes.
- Search interface to the knowledge collection to find information like gene annotations, disease associations, sequences domain architectures, interfering chemicals and involved pathways.
- Enrichment analysis for Gene Ontology terms, diseases, pathways and other biological concepts.
- Extraction of the protein-protein, protein-chemistry interactions networks.
- Compilation of clusters based on sequence homology & sequence domain architectures in a given list(s) of genes.
- Analysis and clustering of transcription factor binding site (TFBS) profiles.
- Access to orthology information, clinical trial and patent information.
- Comparison of results derived from different experimental conditions, time series or treatments.

B. *Architecture*

OnTheFly 2.0 uses a client-server architecture as described in Figure 2. The main front page was implemented using HTML/Javascript while the back-end client-server communication is based on Java servlets. It is currently compatible with most of the widely used browsers such as Chrome (version >6) and Firefox (version >7) whereas there are still some compatibility issues with Safari (version >5). Notably, the OnTheFly application is not supported by the Internet Explorer browser.

The overall performance of the OnThefly 2.0 service can

be assessed separately in three different parts. The initial uploading of the documents is limited by the upload speed and thus the size of the documents. The second step refers to the conversion time cost, as each of the documents needs to be converted into a HTML file. The last part refers to the annotation time where each word is queried against the dictionary. As a rough estimation, a full text article of about 15 pages with images and tables, takes typically between 15-20 seconds from the initial upload to the final presentation of results.

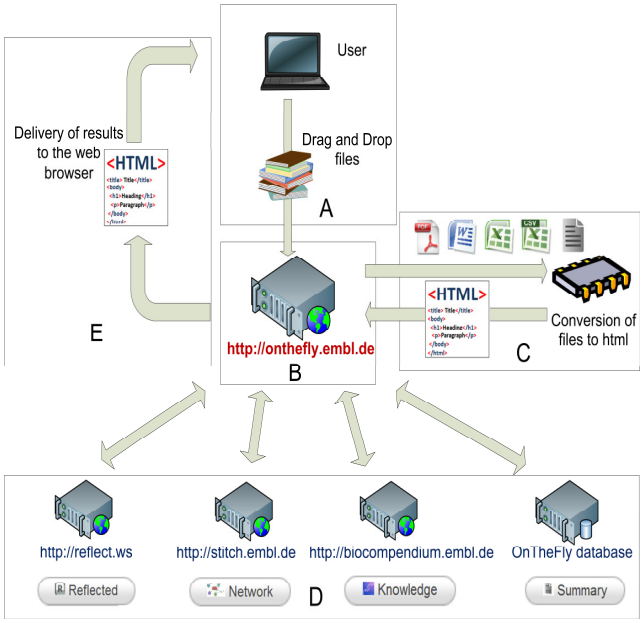

Fig. 2. An overview of the architecture of OnTheFly 2.0. Document annotation and bioentity recognition is performed by the Reflect server. Networks of protein-protein and protein-chemical interactions are generated by the STITCH server. Functional annotation of the terms is performed by the BioCompendim server. Summaries including protein and chemical descriptions and protein sequences are stored in a local server,

C. API

Currently, OnTheFly 2.0 comes with its own Application Programming Interface (API). Initially, OnTheFly assigns a session ID for each job and the user can then select an organism of choice and subsequently upload a file programmatically. After the file gets annotated and the genes, proteins, and chemicals are identified and linked to a corresponding database depending on the organism, several options are offered for further analysis as shown below. We offer a Perl module, simplifying the programmatically access to the OnTheFly API.

- *"http://onthefly.embl.de/sessionID/filename/entities"*
- *"http://onthefly.embl.de/sessionID/filename/stitch.input"*
- *"http://onthefly.embl.de/sessionID/filename/biocompendium.input"*
- *"http://onthefly.embl.de/Session"*
- *"http://onthefly.embl.de/Organisms"*

To give an example, let's assume that a user wants to detect the human proteins in a text file. After annotation, he will be provided with a list of identifiers such as "*9606.ENSP00000164227*" or "*-1.CID000002173*". In the case of a protein or a gene, the first part consists of the NCBI taxonomy identifier for the organism (i.e. *9606* for *Homo Sapiens*) whereas in this example, the remaining part consists of the ENSEMBL identifier [17]. Depending on the selected organism, the protein identifier will be database specific, following the STRING protocol [18]. In the case of a chemical, the first part of the identifier is marked with "*-1*" whereas the second part (i.e. *CID000002173*) indicates the PubChem record. When the URL ends with "/entities", the list which is generated, provides identifiers for the bioentities against all of the organisms stored in the database. Thus, if a protein exists both in *human* and *mouse* (i.e. *P53*), the list will contain both the human and the mouse identifiers (according to the NCBI Taxonomy database), followed by their ENSEMBL identifier (i.e. *9606.ENSP00000269305* for *human* and *10090.ENSMUSP00000005371* for *mouse*). When the URL ends with "/stitch_input", the generated list will be organism v specific and can be used directly as an input to the STITCH database. Finally when the URL ends with "/biocompendium.input", the list can be used directly as an input for the unpublished BioCompendium service for further extensive analysis. Figure 3, demonstrates an example of OnTheFly's API interface. A pdf document has been created manually and most of the terms mentioned in the document were annotated. The figure demonstrates the relevant lists produced by OnTheFly service and their respective URLs. Notably, due to data privacy, the articles are immediately deleted from the server after a complete analysis whereas the API URLs containing the results, remain alive for 30 minutes.

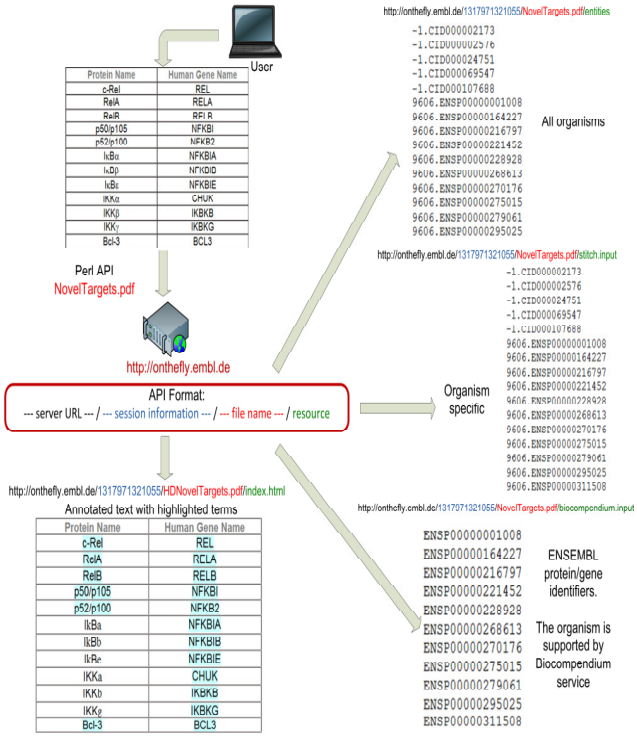

Fig. 3. An example of the OnTheFly API. A client can be used to programmatically annotate documents. A file path is provided to the OTF 2.0 server and several URLs are sent back containing HTML with different types of information (ENSEMBL and PubCheM identifiers/organism and non-organism specific).

D. OnTheFly 1.0 vs OnTheFly 2.0

OnTheFly 2.0 comes with significant advantages compared to its predecessor OnTheFly 1.0. A point-by-point comparison between the two versions concerning the interface, the functionality, the bioentity enrichment, the architecture and the performance is presented in Table 1.

TABLE I
COMPARISON BETWEEN THE TWO VERSIONS OF ONTHEFLY SERVICE

Features	OnTheFly v1.0	OnTheFly v2.0
Informative popup windows	X	X
Genes, Proteins, Chemicals	X	X
Support of .pdf, .doc(x), .xls(x) and .txt	X	X
Multiple selection of files	X	X
Wikipedia terms		X
Manual enrichment of Reflect's dictionary for aliases and synonyms		X
Support of .csv		X
Unlimited size for generated networks		X
630 organisms		X
Advanced Summaries		X
Highly interactive web interface		X
Application Programming Interface(API)		X
Integration with Biocompendium service		X
Better performance		X
Advanced client-server architecture		X

III. DISCUSSION

The OnTheFly web service is a widely used application with hundreds of visits per day from all over the world. It is designed to aid researchers and curators in increasing their productivity by offering an easy-to-use interface to provide candidate database identifiers and useful supplementary information of entities that are referred in a text. While most of the available tools orient their functionality towards similar approaches, most of them are able to only analyze and annotate sets of PubMed abstracts. To our knowledge, OnTheFly is the first available tool that goes one step further by providing full-text analysis in locally stored documents. Notably, disambiguation issues are addressed by the Reflect service and information about the quality and the accuracy of the dictionary can be found in the corresponding publication [14]. OnTheFly provides a bridge between identifiers of different databases, greatly reducing the need for manual cross-referencing and is designed with the help of biologists, while it offers a non-obtrusive and practical help for screen-based scientific reading. Constant maintenance and exception handling are manually ministered.

REFERENCES

[1] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: critical assessment of information extraction for biology," *BMC bioinformatics,* vol. 6 Suppl 1, pp. S1, 2005.

[2] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge," *Genome biology,* vol. 9 Suppl 2, pp. S1, 2008.

[3] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Res,* vol. 33, no. Database issue, pp. D433-7, Jan 1, 2005.

[4] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer, and P. Bork, "STITCH 2: an interaction network database for small molecules and proteins," *Nucleic acids research,* vol. 38, no. Database issue, pp. D552-6, Jan, 2010.

[5] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic acids research,* vol. 36, no. Database issue, pp. D684-8, Jan, 2008.

[6] N. Papanikolaou, E. Pafilis, S. Nikolaou, C. A. Ouzounis, I. Iliopoulos, and V. J. Promponas, "BioTextQuest: a web-based biomedical text mining suite for concept discovery," *Bioinformatics,* vol. 27, no. 23, pp. 3327-8, Dec 1, 2011.

[7] R. A. Erhardt, R. Schneider, and C. Blaschke, "Status of text-mining techniques applied to biomedical text," *Drug discovery today,* vol. 11, no. 7-8, pp. 315-25, Apr, 2006.

[8] D. Rebholz-Schuhmann, A. Jimeno-Yepes, M. Arregui, and H. Kirsch, "Measuring prediction capacity of individual verbs for the identification of protein interactions," *Journal of biomedical informatics,* vol. 43, no. 2, pp. 200-7, Apr, 2010.

[9] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno, "Text processing through Web services: calling Whatizit," *Bioinformatics,* vol. 24, no. 2, pp. 296-8, Jan 15, 2008.

[10] R. Hoffmann, and A. Valencia, "A gene network for navigating the literature," *Nature genetics,* vol. 36, no. 7, pp. 664, Jul, 2004.

[11] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database (Oxford),* vol. 2011, pp. baq036, 2011.

[12] J. L. Fink, P. Fernicola, R. Chandran, S. Parastatidis, A. Wade, O. Naim, G. B. Quinn, and P. E. Bourne, "Word add-in for ontology recognition: semantic enrichment of scientific literature," *BMC bioinformatics,* vol. 11, pp. 103, 2010.

[13] G. A. Pavlopoulos, E. Pafilis, M. Kuhn, S. D. Hooper, and R. Schneider, "OnTheFly: a tool for automated document-based text annotation, data linking and network generation," *Bioinformatics,* vol. 25, no. 7, pp. 977-8, Apr 1, 2009.

[14] E. Pafilis, S. I. O'Donoghue, L. J. Jensen, H. Horn, M. Kuhn, N. P. Brown, and R. Schneider, "Reflect: augmented browsing for the life scientist," *Nat Biotechnol,* vol. 27, no. 6, pp. 508-10, Jun, 2009.

[15] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D808-15, Jan, 2013.

[16] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein-chemical interactions," *Nucleic Acids Res,* vol. 40, no. Database issue, pp. D876-80, Jan, 2012.

[17] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp, "The Ensembl genome database project," *Nucleic Acids Res,* vol. 30, no. 1, pp. 38-41, Jan 1, 2002.

[18] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "STRING 8--a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Res,* vol. 37, no. Database issue, pp. D412-6, Jan, 2009.