SHORT COMMUNICATION

iMole, a web based image retrieval system from biomedical literature.

Manuela Giordano[1], Massimo Natale[1,2], Moreno Cornaz[1], Andrea Ruffino[1], Dario Bonino[1], Enrico M. Bucci[1,3]

[1] BioDigitalValley S.r.l., 11026 Pont Saint Martin, Italy

[2] Department of Control and Computer Engineering, Politecnico di Torino, 10129 Torino, Italy

[3] Istituto di Biostrutture e Bioimmagini, 80134 Napoli, Italy

Corresponding Author:
Enrico M. Bucci
via Carlo Viola ,78
11026 Pont Saint Martin (AO)
enrico.bucci@biodigitalvalley.com
t +39 0125 344249
f +39 0125 808487

Keywords: two-dimensional electrophoresis, image database, meta-analysis, text mining, web application

Total number of words: 2.488

Received: November 30, 2012; Revised: March 15, 2013; Accepted: March 30, 2013

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/elps.201300085

ABSTRACT

iMole is a platform that automatically extracts images and captions from biomedical literature. Images are tagged with terms contained in figure captions by means of a sophisticate text mining tool. Moreover iMole allows the user to upload directly their own images within the database and manually tag images by curated dictionary. Using iMole the researchers can develop a proper biomedical image database, storing the images extracted form paper of interest, image found on the web repositories and their own experimental images. In order to show the functioning of the platform, we used iMole to build a Two-dimensional gel electrophoresis (2DE gels) database (DB). Briefly, tagged 2DE gel images were collected and stored in a searchable 2DE gel database, available to users through an interactive web-interface. Images were obtained by automatically parsing 16 608 proteomic publications, which yielded more than 16 500 images. The database can be further expanded by users with images of interest trough a manual uploading process. iMole is available with a preloaded set of 2DE gel data at http://imole.biodigitalvalley.com.

Authors of biomedical articles do often use images to present experimental results, or to communicate important concepts by means of appropriate charts. In the biomedical research community, much attention is drawn by figures, since figures often summarize the findings of the research work under consideration [1]. There is growing evidence of the need for automated systems that would help biologists in finding the information contained in images and captions quickly and satisfactorily [2]. Different figure mining systems have indeed been proposed, including the Subcellular Location Image Finder (SLIF) system [3], BioText [4] and Yale Image Finder (YIF) [5]. All these systems store the processed information in a web-accessible, searchable database. Interestingly, these instruments allow researchers to structurally browse, in a precise and rapid way, otherwise unstructured knowledge. In particular, the development of proteomic image databases (2DE gel DB), with tools for the comparison of proteomic experimental maps, is widely encouraged [6]. The possibility of using text mining and image processing technologies to create an interactive 2DE gel repository could therefore be of interest to the field. Furthermore, the 2DE gel DB are often used to support experimental identification of spots, as obtained by mass spectrometry analysis [7].

Since it was launched in 1993, the ExPASy website (http://www.expasy.org/) has been a reference in the proteomics community. Through the World-2D PAGE Portal, users can access more than 60 federated databases containing nearly 420 2DE gel images [8]. The Proteome Experimental Data Repository (PEDRo), in addition, provides access to a collection of descriptions of proteomic protocols [9]. The Integrated Proteomics Exploring Database (IPED), based on the schema of PEDRo, was introduced in 2009 with the aim of standardize, store and visualize data obtained, not only from proteomics experiments but also from mass spectrometry (MS) assays [10]. However, current 2DE gel databases do not provide tools for high-throughput data uploading, and they lack convenient tools for data access.

Here we discuss iMole, an innovative system for automated image extraction, annotation and storage into a database, which is able to parse biological journal articles. Using BFO Java library (http://bfo.com), iMole is capable of parsing biomedical articles and to automatically extract images and captions from Portable Document Format (PDF) documents. iMole allows the user to upload experimental gel images adding them to literature-retrieved images. Furthermore, the user can add and modify various information to each 2D gel image as detailed spots annotations, isoelectric point (p*I*) and protein molecular weight (MW) landmarks, free text description and controlled vocabulary terms. Finally, the user can create a personal set of tagged 2DE images.

Interestingly, this instrument allows researchers to browse the vast amount of unstructured data in a precise and rapid way. In order to retrieve images, users can query the database using terms belonging to different ontologies.

iMole has been built according to the principles of Expasy federated database [11], and it provides: (i) gel images accessible directly on the iMole website; (ii) annotated gels with clickable protein spots, that if selected, display basic protein information; (iii) spot linked to other databases (i.e. Uniprot, Gene); (iv) keyword search query.

As a pilot demonstration, we used iMole to build a 2DE gel electrophoresis database. We used the software ProteinQuest (PQ) (www.proteinquest.com) [12] to search all free full text articles with available captions stored in Medline®. For all these target scientific articles we retrieved the PDF files and we extracted the images from these files by BFO, as described by Natale et al. [12].

In order to identify the captions referring to 2DE gel experiments, we tested two different classifiers: a supervised in house developed indexing method (iMole text mining tool) and a supervised learning neural network (NN) developed using Weka framework. We hand-

selected two independent sets of 2000 and 981 captions for training and testing steps, respectively. To train the multilayer perceptron of Weka we used the training set of 2000 captions (100 concerning 2DE gel and 1900 not concerning 2DE gel). The multilayer perceptron of Weka was implemented with three hidden levels and ten nodes in each level.

iMole text mining tool parsed all captions searching for terms related to the "2-Dimensional Gel Electrophoresis" technique and its alias (e.g. 2D-GE, DIGE, bi-dimensional electrophoresis, etc.); these terms belong to the methods dictionary. Ambiguities in the terminology are resolved using multiple searches for more than one alias, as well as the co-occurrence of specific words which can either deny or force the tagging [14].

To check the efficiency of the two classifiers, we analyse the control set of 981 captions (391 concerning 2DE gels and 590 not concerning 2D-gel). Our tool wrongly identifies as positive 97 captions (false positives, fp), and it correctly classifies, as negative, all captions that do not concern 2DE gels. The NN wrongly identifies 104 captions as positives (fp), and 91 as negative (false negative, fn) (Table 1). The comparison between the two systems reveals that our classification tool gives more reliable results than the NN system. In fact, using iMole the fraction of retrieved captions that are correct are 80.12 % (Precision), instead of 74.26 % obtained with the NN classification. Furthermore the fraction of correct captions retrieved by iMole tool, are 100.00 % (Recall) instead of 76.73 % performed by the NN system. The results were checked manually.

Other existing classifiers which can be used for retrieval of 2D gel images include BioText [4] and YIF [5], which are based on full-featured text search engine, such as Apache Lucene [15]. With respect to the implementation of these software, we indexed paper images with the help of semantic dictionaries describing biological entities, thus resolving ambiguities in a better way (an example is shown in Supporting Information: Figure 1S).

In order to compare our classifier to existing tools, we tested side by side iMole and YIF (http://krauthammerlab.med.yale.edu/imagefinder/). We queried YIF using "2-Dimensional Gel Electrophoresis" and all the aliases obtained from our dictionary, finding 20 196 images vs 16 752 obtained by the same query in iMole (data are shown in supplementary Table S1). To allow for a direct, manual comparison of the results, we restricted the search, filtering for the protein ApoE and analyzing all the images returned by iMole and YIF. The results shown in Table 3 demonstrate that iMole retrieves a higher number of correct images than the YIF platform. The higher accuracy of the process (no false positive or false negative images retrieved by iMole) is demonstrated by the precision index calculated on the iMole results (Table 4). In this particular case, iMole reached a precision index of 100.00%, whereas the

precision index calculated on YIF results, did not reach 10%. By repeating the experiment with other proteins, we retrieve similar results, demonstrating a better retrieval accuracy for iMole.

Images positively identified, by iMole, are tagged as 2DE gels and stored. In addition, the software associates other tags to the extracted images using several other ontologies, referring to diseases, proteins, tissue, cell type and organisms. Extra tags can be manually added using an input box with an auto-complete function, so to be compliant with a precompiled dictionary derived from Entrez MeSH with some manually editing. Images of 2DE gels maintain a link to the original publication through a PubMed identifier (ID) tag, that allows researchers to retrieve the original article containing them.

In order to fine tune the evaluation of the specificity of the images collected in the 2DE gels database, we verify how many captions, that contain the term "2DE gel" (or its alias), are really associated to 2DE gel images. Actually, we controlled 22 366 images that iMole identify as 2DE gels and we found that 17 238 are 2DE gels images and 5128 images are other than 2DE gels. The precision calculated on these data is 77.07 %. Hence, out of a hundred captions that contain the term "2DE gel", 77 are actually associated with a 2DE gel images. Taking into consideration these results as a whole, the iMole image selection system returns much more relevant images than irrelevant ones.

Furthermore, iMole can also support the manual images uploading performed by users. These images can be manually tagged and named by the user, and are associated to a free text description instead of a caption.

Supplementary information, such as p*I* and MW data, can be added to all images stored in the database. Furthermore, the users can add protein names to spots not yet annotated. Possible ambiguities are avoided at this stage by means of the auto-complete text function, which will force the user to select among official Entrez protein symbols. All protein annotations are automatically linked to Entrez, so to have a rapid access to protein information.

Additionally, users can associate further information with the gel, like 2DE landmarks or any terms of provided dictionaries. All image information that results from paper processing and from user uploading, is available through an interactive web-interface. Users can query the corresponding database by searching for the image names or for terms contained into the preloaded dictionaries to retrieve a specific image.

iMole database was developed in a MySQL environment on a Red-Hat server. The iMole web interface is implemented with servlets and Java Server Pages (JSP) technologies, utilizing Apache Tomcat as the JSP engine. Users can access the application after personal

authentication. In this way, they can easily manage and recover their own data, without interfering with other users.

At present, our pilot 2DE gel database houses more than 16 752 published and some users' uploaded experimental gels, and therefore is, to the best of our knowledge, the most comprehensive 2DE gel database ever. Of these images, 216 are accompanied by various information such as detailed spots annotations; all the rest is annotated anyway with indications on the type of biological samples analyzed and information about diseases associated to proteins of interest (an example is shown in Supporting Information: Figure 2S, 3S, and 4S).

Here we presented iMole, a new web platform, that retrieves images and parses their associated captions in biomedical publications. This project demonstrates how an advanced text mining tool can be used to annotate a huge database containing images related to a specific experimental technique such as 2DE gel electrophoresis. By using iMole, users can upload their own experimental gel images adding them to literature-retrieved images, so that they can also create a personal set of tagged 2DE images. This feature allows researchers to study new experimental protocols, to evaluate protein post-translational modifications or to match protein spots in experimental gels. 2DE images annotation is compliant with the federated 2DE database concepts as specified in the introduction.

To the best of our knowledge, iMole is the first example of a bioinformatic platform able to automatically feed an image database parsing PDF files and processing the information contained in figure captions.

AKNOWLEDGEMENTS

The project which led to IMOLE and ParIS was partially funded under the "Legge 84" regional funding program of the Valle d'Aosta Region. We thank Dr. Chiara Abrescia for her useful suggestions during the editorial revision of this manuscript.

REFERENCES

[REFERENCES

[1] Kim, D., Yu,,H. PloS one, 2011, 6, e15338.

[2] Divoli, A., Wooldridge, M.A., Hearst, M.A. PloS one, 2010, 5, e9619.

[3] Ahmed, A., Arnold, A., Coelho, L.P., Kangas, J., Sheikh, A.S., Xing, E., Cohen, W., Murphy. R.F., Web Semant., 2010, 8, 151-154.

[4] Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A., Ye, J. Bioinformatics, 2007, 23, 2196-2197.

[5] Xu, S., McCusker, J., Krauthammer, M. Bioinformatics, 2008, 24, 1968-1970.

[6] Drews, O., Görg, A. Nucleic Acids Res., 2005, 33, D583–D587.

[7] López-Farré, A.J., Mateos-Cáceres, P.J., Sacristán, D., Azcona, L., Bernardo, E., de Prada, T.P., Alonso-Orgaz, S., Fernández-Arquero, M., Fernández-Ortiz, A., Macaya, C. J Proteome Res., 2007, 6, 2481-2487.

[8] Hoogland, C., Mostaguir, K., Appel, R.D., Lisacek, F. J. Proteomics, 2008, 71, 245-248.

[9] Taylor, C.F., Paton, N.W., Garwood, K.L., Kirby, P.D., Stead, D.A., Yin, Z., Deutsch, E.W., Selway, L., Walker, J., Riba-Garcia, I., Mohammed, S., Deery, M.J., Howard, J.A., Dunkley, T., Aebersold, R., Kell, D.B., Lilley, K.S., Roepstorff, P., Yates, J.R., Brass, A., Brown, A.J., Cash, P., Gaskell, S.J., Hubbard, S.J., Oliver, S.G. Nat Biotechnol., 2003, 21, 247-254.

[10] Zheng, G., Li, H., Wang, C., Sheng, Q., Fan, H., Yang, S., Liu, B., Dai, J., Zeng, R., Xie, L. Acta Biochim Biophys Sin., 2009, 41, 273-279.

[11] Hoogland, C., Mostaguir, K., Appel, R.D. Methods Mol. Biol., 2009, 519, 533-539.

[12] Gatti, S., Leo, C., Gallo, S., Sala, V., Bucci, E., Natale, M., Cantarella, D., Medico, E., Crepaldi, T. Transgenic Res., 2012, 6, doi: 10.1007/s11248-012-9667-2.

[13] Natale, M., Bonino, D., Consoli, P., Alberio, T., Ravid, R.G., Fasano, M., Bucci, E.M. Bioinformatics, 2010, 26, 946-52.

[14] Alberio, T., Bucci, E.M., Natale, M., Bonino, D., Di Giovanni, M., Bottacchi, E., Fasano, M. J Proteomics 2013, doi: 10.1016/j.jprot.2013.01.025.

[15] Ghosh, P., Antani, S. , Long, L.R., Thoma, G.R. Computer-Based Medical Systems (CBMS), 2011, 1-6.

Table 1. Results of the text mining engine: the Weka supervised learning neural network (NN), and iMole text matching method.

	NN text mining tool	iMole text mining tool
true positive	300	391
false positive	104	97
true negative	486	493
false negative	91	0

Table 2. Results of the analysis described in the text:
precision (true positive / (true positive + false positive)), recall (true positive / (true positive + false negative)) and F1 score (2 x (precision x recall/precision + recall).

	precision	recall	F1 score
NN	74.26%	76.73%	0.754717
iMole	80.12%	100.00%	0.889647

Table 3. Results of the query of 2DE gel AND "apoe" using Yale Image Finder and iMole.

	YIF ("apoe" filter)	iMole ("apoe" filter)
true positive	60	84
false positive	606	0
true negative	0	622
false negative	24	0

Table 4. Results of the comparison between Yale Image Finder and iMole:
precision (true positive / (true positive + false positive)), recall (true positive / (true positive + false negative)) and F1 score (2 x (precision x recall/precision + recall).

	precision	recall	F1 score
YIF	9.01%	71.43%	0.16
iMole	100.00%	100.00%	1