BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 16 2009, pages 2082–2084
doi:10.1093/bioinformatics/btp318

Data and text mining

Figure mining for biomedical research

Raul Rodriguez-Esteban[1,*] and Ivan Iossifov[2]

[1]Systems Biology, Pfizer Inc., Cambridge, MA 02139 and [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742, USA

Received on March 2, 2009; revised and accepted on May 11, 2009
Advance Access publication May 13, 2009
Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Figures from biomedical articles contain valuable information difficult to reach without specialized tools. Currently, there is no search engine that can retrieve specific figure types.

Results: This study describes a retrieval method that takes advantage of principles in image understanding, text mining and optical character recognition (OCR) to retrieve figure types defined conceptually. A search engine was developed to retrieve tables and figure types to aid computational and experimental research.

Availability: http://iossifovlab.cshl.edu/figurome/

Contact: raul.rodriguez-esteban@pfizer.com

1 INTRODUCTION

Biomedical articles are not just text. They include tables, graphs, charts, pictures and multimedia files. If we consider supplementary information, databases and web sites, an article's content might be scattered in numerous locations and formats. While the properties of an article's text have been dilated upon, the focus of the present study is on the figures and tables. Figures store experimental data often not available anywhere else. Due to their limit in number and costly production, figures are the subject of carefully curated content, including protocols or descriptions of biological processes. Tables store experimental results and data summaries in an aggregate manner, and they are sometimes represented using images or embedded within figures, blurring the difference between tables and figures. Tables can be mined for data sharing with technologies like the semantic web (Pivk *et al.*, 2007).

Finding figures or tables with tools not designed for the task is not trivial. Three previous studies have considered the issue of figure retrieval: Sub-cellular Location Image Finder (SLIF; Murphy *et al.*, 2004), BioText (Hearst *et al.*, 2007a, b) and Yale Image Finder (YIF; Xu *et al.*, 2008a). Both BioText and YIF perform text queries across different article sections, and YIF additionally searches through text embedded in the image. SLIF retrieves fluorescence micrograph images from a single-journal corpus. Micrograph images have regular properties that allow high-retrieval performance; therefore the design may not be flexible enough to be adapted to other figure types.

Figures convey information in similar ways by using a recognizable language, which is arguably part of the scientific language. The degree of coherence and articulation reflects the grammar that regulates the representation of layout and pictorial elements (e.g. text, colors, lines, shapes). Despite the regularities, there exists ample variability in figure design to hamper efforts at categorization. A figure may belong to several figure types (or none) by virtue of its variegated facets or its constituent sub-figures. This variability prevents applying simple descriptions to define figures. Rafkind *et al.* (2006) proposed a classification system that divided figures into five sets according to their coherence and frequency, similar to the division by Shatkay *et al.* (2006). We propose classes driven by research needs and we show that the flexibility of our methods allow for a less restricted approach than Rafkind *et al.* (2006) used for classification.

2 METHODS

The object of the study was a set of 80 949 articles from the digital archive PubMed Central (PMC). Some 233 395 images and 86 625 tables were extracted from this set and indexed using the open source Lucene technology.[1] The text indexed came from tables, table captions, figure captions and text within figure images. Additionally, the full text of articles was indexed separately to allow for more focused searches, e.g. searching for figures about the protein *p*53 within articles mentioning colon cancer.

Four figure types were selected in consultation with computational and experimental biologists in the area of systems biology. The types selected were *gel*, *pathway*, *structure* and *time*. Gel figures were defined as those depicting gel electrophoresis experimental results. Pathway figures were defined as diagrams representing interactions involving at least one protein. Structure figures described or depicted primary, secondary, tertiary and quaternary protein structures. Time figures plotted or listed data values over time.

A machine learning algorithm was trained to automatically annotate figures with the types defined. First, sets of randomly selected figures were manually annotated (Table 1). Then, figure descriptors were generated drawing from both text and image. The text was extracted from figure captions and from within figure images. Extracting text embedded in figure images required preliminary processing (Kou *et al.*, 2007; Li *et al.*, 2008; Xu *et al.*, 2008b). This included cleaning non-textual elements using properties that characterize horizontal text objects: alignment, height–width proportion, character separation and character connectedness. Text with angle different from horizontal was not considered. After optical character recognition (OCR) using ABBYY FineReader 9.0, the text extracted and the caption text were tokenized and encoded separately with a method called set-of-words which consists in representing every distinct text token as a feature with value one. The figure image was processed to generate several feature sets (Kalva *et al.*, 2007; Ritter and Wilson, 1996): color histograms, shape distributions and texture descriptors (similar to Rafkind *et al.*, 2006). Color histograms were built from gray-scale, black-and-white, RGB and HSV color space frequencies. Shape distributions were measured with

[*]To whom correspondence should be addressed.

[1]http://lucene.apache.org

Table 1. Figure type frequency within the annotated samples

Figure types	Frequency	Percentage
Gel	353/4147	8.5
Pathway	204/10029	2.0
Structure	139/4005	3.5
Time	287/2001	14.3

The number of samples annotated was different for each type. The frequency of figure types changed as shown by Percentage, time figures being the most frequent and pathway figures the least. The Percentage column represents the same values as the Frequency column.

Sobel operators. Texture descriptors were created from spatial gray level dependence statistics. All features were binarized using median values. Text and image features were combined in one vector per figure.

Support vector machines (SVMs) were selected as machine learning method (Chapelle *et al.*, 1999). SVM implementations (Joachims, 1999) handle weighted cost functions, which can better deal with unbalanced datasets (i.e. datasets with classes distributed non-uniformly) (Morik *et al.*, 1999). While classifiers are frequently designed to improve class-assignment accuracy, retrieval performance depends on precision and recall. In retrieval, maximizing accuracy is an optimal strategy only for balanced datasets. To deal with this problem, several strategies have been proposed (Chawla *et al.*, 2004), such as increasing the penalty for misclassification of positive samples. In SVMs, this penalty can be reflected in the cost-factors for positive and negative samples (C_+ and C_-, respectively). For example, for $C_+ = 2C_-$ a misclassified positive example is as pernicious as two misclassified negative examples. The smaller the proportion of positive examples the larger C_+ needs to be to compensate. A typical choice for the ratio C_+/C_- is the ratio of negative training samples over positive training samples.

SVM training and test sets were generated by 10-fold random sampling without replacement. Precision, recall, micro-averaged F-measure (van Rijsbergen, 1979) and area under the ROC curve (Fawcett, 2006) were used for evaluation. While F-measure is the commonest metric for retrieval performance, the area under the ROC curve is insensitive to class skew and therefore directly comparable between sets with different class distribution.

3 RESULTS AND DISCUSSION

Table 2 shows the performance of the trained algorithms. Retrieval of time figures yielded the lowest value, reflecting the difficulty of the task. Feature sets were of varying importance for each type, reflecting their different properties. For time figures, text within the figures was crucial because axis labels or measure units can be highly predictive. Image features had lower importance, being most useful for structure figures. Caption text features were the most important overall. Features derived from the text immediately before and after the figure, and from full text, did not improve performance (data not shown). Table 3 shows query examples. Search options allow filtering for single journals and full text.

The tool presented can take the role of a specialized image repository, with the added capability of query searching. The repository of pathway drawings BioCarta[2] contains 354 reference pathways, and the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) contains 345. In comparison, we predicted more than 4500 pathway figures in our set.

Conflict of Interest: none declared.

Table 2. Recall, precision, F-measure (F) and area under the ROC curve (AUC) for each figure type and combination of feature sets: *image* for features extracted from the figure image, OCR for features derived from the figure image OCR text and *text* for features from the figure caption.

Type	Features	Recall	Precision	F	AUC
Gel	Image	0.80	0.16	0.27	0.74
	OCR	0.55	0.45	0.50	0.79
	Text	0.81	0.82	0.81	0.97
	Image + OCR	0.65	0.30	0.41	0.84
	Image + text	0.83	0.77	0.80	0.97
	OCR + text	0.83	0.82	0.83	0.97
	Image + OCR + text	0.83	0.84	0.84	0.97
Pathway	Image	0.70	0.04	0.08	0.69
	OCR	0.36	0.31	0.33	0.77
	Text	0.57	0.65	0.61	0.93
	Image + OCR	0.50	0.14	0.22	0.85
	Image + text	0.71	0.61	0.66	0.97
	OCR + text	0.59	0.75	0.66	0.97
	Image + OCR + text	0.64	0.69	0.67	0.96
Structure	Image	0.81	0.08	0.14	0.76
	OCR	0.36	0.14	0.20	0.69
	Text	0.66	0.75	0.71	0.93
	Image + OCR	0.58	0.16	0.25	0.82
	Image + text	0.72	0.68	0.70	0.95
	OCR + text	0.64	0.78	0.70	0.94
	Image + OCR + text	0.83	0.83	0.83	0.98
Time	Image	0.85	0.21	0.33	0.67
	OCR	0.55	0.68	0.61	0.78
	Text	0.43	0.48	0.46	0.79
	Image + OCR	0.66	0.39	0.49	0.81
	Image + text	0.59	0.45	0.51	0.83
	OCR + text	0.57	0.69	0.62	0.86
	Image + OCR + text	0.63	0.62	0.63	0.88

Table 3. Comparison between keyword and type search

Query	Type	Retrieved	FP	Precision
p53 pathway	–	36	12	0.75
p53	Pathway	73	9	0.89
pi3k time	–	28	6	0.82
pi3k	Time	35	9	0.80
jnk western	–	78	2	0.98
jnk gel	–	10	2	0.83
jnk	Gel	126	5	0.96
Thrombin sequence	–	9	3	0.75
Thrombin structure	–	5	0	1.00
Thrombin	Structure	10	0	1.00

Example queries and number of retrieved hits, false positives (FP) and precision.

REFERENCES

Chapelle,O. *et al.* (1999) Support vector machines for histogram-based image classification. *IEEE T. Neural Networ.*, **10**, 1055–1064.

Chawla,N.V. *et al.* (2004) Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.*, **6**, 1–6.

Fawcett,T. (2006) An introduction to ROC analysis. *Patt. Rec. Lett.*, **27**, 861–874.

Hearst,M.A. *et al.* (2007a) BioText search engine: beyond abstract search. *Bioinformatics*, **23**, 2196–2197.

[2] http://www.biocarta.com

Hearst,M.A. *et al.* (2007b) Exploring the efficacy of caption search for bioscience journal search interfaces. In *Proceedings of the BioNLP Workshop*, The Association for Computational Linguistics, Prague, Czech Republic, pp. 73–80.

Joachims,T. (1999) Making large-Scale SVM learning practical. In Schölkopf,B. *et al.* (eds) *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, Cambridge, MA.

Kalva,P.R. *et al.* (2007) WEB image classification based on the fusion of image and text classifiers. In *Proceeding of the 9th International Conference on Document Analysis and Recognition*, Computer Society Press, Curitiba, Brazil, pp. 561–568.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kou,Z. *et al.* (2007) A stacked graphical model for associating sub-images with subcaptions. *Pac. Symp. Biocomput.*, 257–268.

Li,L. *et al.* (2008) A figure image processing system. Graphics recognition, recent advances and new opportunities. Vol. 5046 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 191–201.

Morik,K. *et al.* (1999) Combining statistical learning with a knowledge-based approach a case study in intensive care monitoring. In *Proceedings of the 16th International Conference on Machine Learning*, Morgan Kaufmann, Bled, Slovenia, pp. 268–277.

Murphy,R.F. *et al.* (2004) Extracting and structuring sub-cellular location information from on-line journal articles: the sub-cellular location image finder. In *Prococeedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE-2004)*, ACTA Press, St. Thomas, US Virgin Islands, pp. 109–114.

Pivk,A. *et al.* (2007) Transforming arbitrary tables into logical form with TARTAR. *Data Knowl. Eng.*, **60**, 567–595.

Rafkind,B. *et al.* (2006) Exploring text and image features to classify images in bioscience literature. In *Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL*, The Association for Computational Linguistics, New York, New York, pp. 73–80.

Ritter,G.X. and Wilson,J.N. (1996) *Handbook of Computer Vision Algorithms in Image Algebra*, 1st edn. CRC Press, Boca Raton, FL.

Shatkay,H. *et al.* (2006) Integrating image data into biomedical text categorization. *Bioinformatics*, **22**, e446–e453.

van Rijsbergen,C.J. (1979) *Information Retrieval*, 2nd edn. Butterworths, London.

Xu,S. *et al.* (2008a) Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics*, **24**, 1968–1970.

Xu,S. *et al.* (2008b) Improving OCR performance in biomedical literature retrieval through preprocessing and postprocessing. In *Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine (SMBM'08)*, pp. 161–164.