

(Computationally) Glancing at Biomedical Images

Hagit Shatkay¹ and Dorothea Blostein²

¹ The Computational Biology and Machine Learning Lab

² The Software Technology Lab

School of Computing, Queen's University

Kingston, Ontario, Canada K7L 3N6

email: shatkay@cs.queensu.ca

Keywords: biomedical image data, text mining, figures, database curation, categorization, classification, relevance evaluation, computational-glancing

Abstract

Quickly browsing through many biomedical papers for relevant information is part of the daily work of scientists and, to an even larger extent, of biomedical database curators (e.g. FlyBase, Mouse Genome Informatics, SwissProt, and others). As extensive literature search is a laborious manual task, automated text-categorization and text mining tools are slowly making their way into the process.

Several community efforts such as the KDD Cup 2002 (The ACM Knowledge Discovery and Data Mining challenge), the BioCreative workshops, and the TREC Genomics tracks, all paused challenges aiming to develop automated means for quickly determining the relevant content of scientific papers. As we noted in the KDD'02 Cup [1], using the text of figure captions significantly improved performance in deciding the relevance of documents for the KDD'02 task (identifying papers discussing wild-type gene expression in *Drosophila*).

Underlying this successful approach is our key observation that the figures themselves are highly informative, and are actively used as relevance-cues by curators, and by researchers who quickly browse through the literature. That is, a brief glance at several images helps to quickly determine the relevance of the whole paper to the curation task or to the research at hand.

While much work has been dedicated to mining the text of biomedical literature, scant little has been done in terms of using image data within that same literature. Our work for aims to develop computational means for utilizing images within documents, similar to the way curators and other readers use them, and to integrate these tools into biomedical literature mining and categorization. That is, we are working on introducing computational ways to "quickly glance at images".

We introduce the challenge, describe methods we have developed to represent images and to use them in ways that support the computational-glancing idea [2,3], and present our experiments to date where we incorporate these ideas into biomedical text categorization and database curation tasks.

References

- [1] Regev Y. *et al.* *Rule-based Extraction of Experimental Evidence in the Biomedical Domain – the KDD Cup (Task 1)*, ACM SIGKDD Explorations, 4(2). 90-91. 2002
- [2] Chen N. *et al.* *Use of Figures in Literature Mining for Biomedical Digital Libraries*. Proc. of the IEEE Int. Conf. on Document Image Analysis for Libraries (*DIAL'06*). 180-197. 2006
- [3] Shatkay H. *et al.* *Integrating Image Data into Biomedical Text Categorization*. *Bioinformatics* 22(14). e446-53. 2006