# Extracting Structured Information from Text and Images in On-line Journal Articles for Localization Proteomics

Robert F. Murphy, Zhenzhen Kou, Juchang Hua, Matthew Joffe and William W. Cohen
Departments of Biological Sciences and Biomedical Engineering and Center for Automated Learning and Discovery, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA/U.S.A.
murphy@cmu.edu, zkou@andrew.cmu.edu, juchangh@andrew.cmu.edu, mjoffe@andrew.cmu.edu, wcohen@cs.cmu.edu

There is extensive interest in automating the collection, organization and summarization of biological data. Data presented in literature, which is often in the form of images and accompanying captions, presents special challenges for such efforts. To extract structured information from text and images in online journals, we have built a system, SLIF (for Subcellular Location Image Finder), which extracts information on one particular aspect of biology from a combination of text and image in journal articles. Our poster will present an overview of the system.

SLIF applies both image analysis and text interpretation to the figure and caption pairs harvested from on-line journals, so as to extract assertions such as "Figure N depicts a localization of type L for protein P in cell type C". The protein localization pattern L is obtained by analyzing the figure, the protein name and cell type are obtained by analysis of the caption. Figure 1 illustrates an overview of the steps in the current SLIF system, with reference to publications in which they are described in more detail.

SLIF starts from decoding journal articles in PDF or XML format and extracting all figure caption pairs. Figure caption pair analysis involves several distinct tasks:

- Split the figure into panels (independently meaningful subfigures)
- Identify panels that depict fluorescence microscope images using a learned classifier[1,2]
- Calculate numerical features that adequately capture information about subcellular location using techniques described in [1,2]
- Extract protein names and cell types from captions[3,5]
- Map the information extracted from the caption to the right sections of the figure[4]

Once the information extraction is accomplished, it is stored in an SQL database. The SLIF database can be searched by standard SQL queries. We have implemented a number of common queries using Java Server Pages (JSP). Figure 2 shows an example query.

Current work is focused on generating summary reports using confidence estimates for the various processing steps, as well as combining the SLIF results with information from the protein databases.

## Reference

[1] R.F. Murphy, M. Velliste, J. Yao, & G. Porreca, Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Locations, 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001), Bethesda, MD, USA, 2001, 119-128.

[2] Z. Kou, W.W. Cohen, & R.F. Murphy, Extracting information from text and images for location proteomics, Proc 3rd ACM SIGKDD Workshop Data Mining Bioinformatics (BIOKDD03), 2003, 2-9.

[3] W.W. Cohen, R. Wang, & R.F. Murphy, Understanding Captions in Biomedical Publications., Proc 9th ACM SIGKDD (KDD-2003), 2003, 499-504.

[4] R. F. Murphy, Z. Kou, J. Hua, M. Joffe, and W. W. Cohen (2004) Extracting and Structuring Subcellular Location Information from On-line Journal Articles: The Subcellular Location Image Finder. Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE 2004), pp. 109-114.

[5] Zhenzhen Kou, William W. Cohen & Robert F. Murphy (2005): High-Recall Protein Entity Recognition Using a Dictionary in ISMB-2005 (forthcoming).

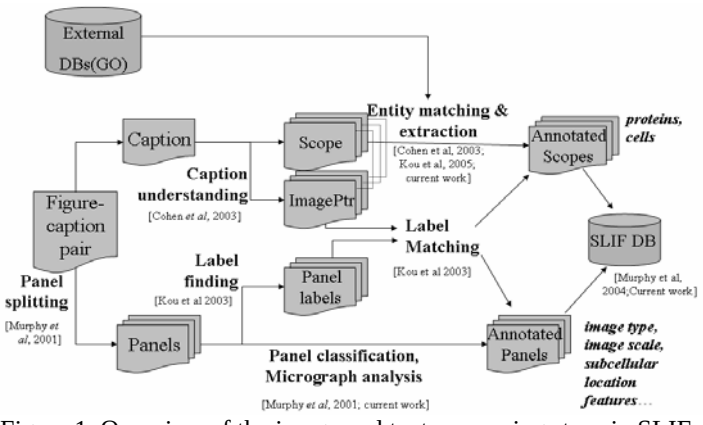


Figure 1. Overview of the image and text processing steps in SLIF.

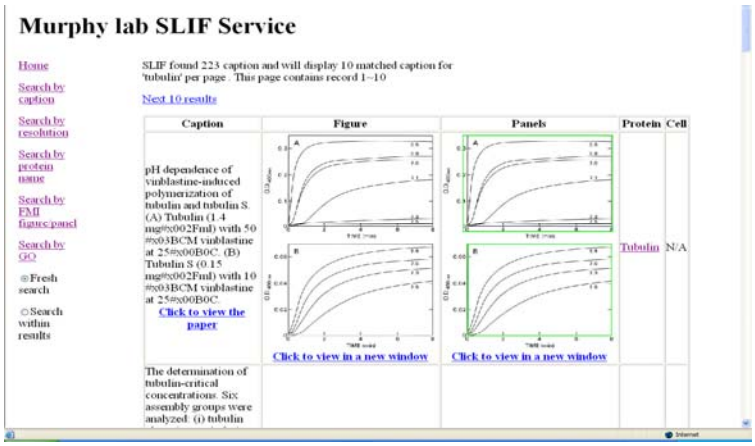


Figure 2. Example SLIF webpage.