EXTRACTING AND STRUCTURING SUBCELLULAR LOCATION INFORMATION FROM ON-LINE JOURNAL ARTICLES: THE SUBCELLULAR LOCATION IMAGE FINDER

Robert F. Murphy murphy@cmu.edu
Zhenzhen Kou zkou@andrew.cmu.edu
Juchang Hua juchangh@andrew.cmu.edu
Matthew Joffe mjoffe@andrew.cmu.edu
William W. Cohen wcohen@cs.cmu.edu
Departments of Biological Sciences and Biomedical Engineering and Center for Automated Learning and Discovery
Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA/U.S.A

ABSTRACT

Previous applications of information extraction methods to articles in biomedical journals have predominantly been based on interpreting article text. This often leads to uncertainty about whether statements that are found are attempts at reviews or summaries of data in other papers, conjectures or opinions, or conclusions from evidence presented in the paper at hand. The ability to extract information from the primary data presented in an article, which is often in the form of images, would allow more accurate information to be extracted. Towards this end, we have built a system that extracts information on one particular aspect of biology from a combination of text and image in journal articles. The design and performance of this system are described here, along with conclusions about possible improvements in the scientific publishing process that we have drawn from our implementation process.

KEY WORDS
Knowledge and Information Retrieval, Multimedia Databases, Data and Text Mining, Location Proteomics

1. Introduction

Biomedical research has undergone a major paradigm shift from consisting primarily of projects in which an individual investigator studies many aspects of a single gene, protein or process to increasingly consisting of projects in which teams of investigators study a single aspect of all genes, proteins or processes in a given cell type, tissue or organism. The successful completion of various genome projects, with their focus on obtaining the sequence of all genes in a particular organism, led this paradigm shift. In general, the results from these projects are objective (at least in the sense that the criteria for decisions are clearly specified independently from the data), systematic, widely useful, and well-suited to delivery via structured databases. While remarkable insights into a wide range of biological phenomena were achieved before the advent of genomics (and of course such insights continue to be achieved), results of traditional biological research are most commonly communicated via journal articles in which raw data, methods, processed results and conclusions are mixed. In addition, the writing styles, vocabularies, and assumptions used for interpretation vary widely from paper to paper.

Thus, there is a dramatic contrast in the ease with which results from the two paradigms can be organized and communicated. This has created a critical need for approaches that can bridge between the systematic, structured information in biological databases and the idiosyncratic, unstructured information in journal articles. This is often posed as a need for automated annotation of gene and protein sequences, but there are at least two significant ways in which the general need differs from the specific approaches taken to sequence annotation. The first is that the need extends to extracting information from literature about biological phenomena at the molecule, cell, tissue and organism level that do not relate directly to sequence. The second is that most prior annotation work has focused on extracting information from the *text* in abstracts (or more rarely, journal articles), but not from supporting published *data* that is often in the form of images.

To illustrate the initial feasibility of addressing these broader needs, we have developed a prototype system that can extract structured information from images and text in journal articles. The focus of this system is on one class of images, those produced by fluorescence microscopy, that capture information about the distribution of proteins and other biological macromolecules inside cells. It builds on our prior work demonstrating the feasibility of fully automated recognition of the distributions characteristic of the major structures that comprise a eukaryotic cell. The work to date not only provides a usable resource for biologists, but also reveals the most difficult challenges for building systems for other categories of biological figures. The work further suggests some alterations in scientific publishing practices that could facilitate automated interpretation without putting undue demands on authors or interfering significantly with the traditional appearance of articles.

2. Automated Interpretation of Fluorescence Microscope Images

The tens of thousands of proteins that make up eukaryotic cells all have specific places in which they carry out their roles. Some proteins play structural roles by working together to create the specialized structures and organelles that cells require, and others orient themselves in relation to those structures to perform specific tasks. Knowledge of these subcellular location patterns, and of how they change due to disease or to alterations in the cell environment) is critical for understanding how all of the proteins in a cell work together and for designing disease diagnostics and therapies. The most common approach used to study location patterns is to label the molecules of a particular protein with a fluorescent probe and collect cell images using a fluorescence microscope. Historically, the analysis of the resulting fluorescence micrographs has been carried out visually by biologists trained to recognize the patterns characteristic of the major cell structures.

As an alternative to visual analysis, our group has designed and implemented automated systems that can carry out this task with greater sensitivity, reproducibility and objectivity [1-4]. The heart of each system is a set of numerical features that captures the important aspects of the subcellular pattern in an image without being overly sensitive to the position, rotation and shape of each cell. These are combined with tools to select a discriminative subset of the features and to train a machine classifier.

Our subcellular location classifiers have been applied primarily to collections of images that we have generated under highly controlled conditions. An important issue is whether and with what modifications they can be applied to sets of images from diverse sources. Since journal articles that describe subcellular location frequently include supporting figures of microscope images, on-line journals represent a potential source of images to test the degree to which our approaches can be generalized. More significantly, the demonstrated feasibility of automated interpretation systems for fluorescence micrographs raises the possibility of building systems for extracting highly structured information on subcellular location from the combination of images and text in on-line journals.

3. Subcellular Location Image Finder

Towards this end, we initially constructed a web agent, which we termed SLIF (for Subcellular Location Image Finder) that could find fluorescence micrographs in articles in Pubmed Central [5]. This agent downloaded PDF files for articles that matched a text-based query to the Pubmed search engine. These PDF files were processed to find pairs of figures and captions, and then the figures were processed to identify "panels" within each figure. (A "panel" is an independently meaningful part of a figure; it is common in biological journals to create composite figures by combining related images and/or graphs.) Each panel was then classified as to whether it contained a fluorescence microscope image, and, if so, it was ranked by the degree to which matched a specific query pattern (using a neural network trained with images of HeLa cells labelled with antibodies against a specific protein). The system was evaluated by reading the caption and examining the image for panels it returned. For example, when a query for articles containing the protein name "tubulin" was combined with a neural network trained to recognize tubulin patterns, eight of the top ten ranked patterns actually contained a tubulin (and one was an artifact of panel segmentation that was subsequently corrected). The ability to use a totally automated approach to find sets of images highly enriched for a desired pattern was extremely encouraging.

4. Structuring Information Extracted from Figures and Captions

Based on the success of this initial effort, we next sought to optimize the steps required for figure processing and to explore extracting additional information from the associated captions. The initial SLIF system had been implemented using a web search approach in which only articles relevant to a specific query were downloaded and processed. To make repeated analysis and testing of our methods simpler, we changed to a model in which a specific large corpus of articles would be fully analyzed and indexed. For this purpose, we have used a collection of over 15,000 articles from the Proceedings of the National Academy of Sciences generously provided for testing purposes (referred to below as the PNAS test set). These articles were provided in an XML format in which figures and associated captions were explicitly identified.

Fig. 1 shows an overview of the steps in the current SLIF system, with references to publications in which they are described in more detail. The components that carry out these steps have been integrated into a file-based, light-weight blackboard system with a declarative control system to specify the inputs and outputs of each step [6]. Each step will be briefly described below, and improvements since the last description of the system [7] will be described more fully.

One of these recent improvements is the addition of an interface to an SQL database that allows the results of analysis steps to be stored as traceable assertions. Each category of assertion is stored in its own table that has an implied relationship between its entries. This facilitates more rapid and complex querying of the results than can be accomplished with the blackboard system alone. The two highest level tables are ones that link a paper to its source (i.e., a URL, DOI, and/or Pubmed ID) and a *figure*

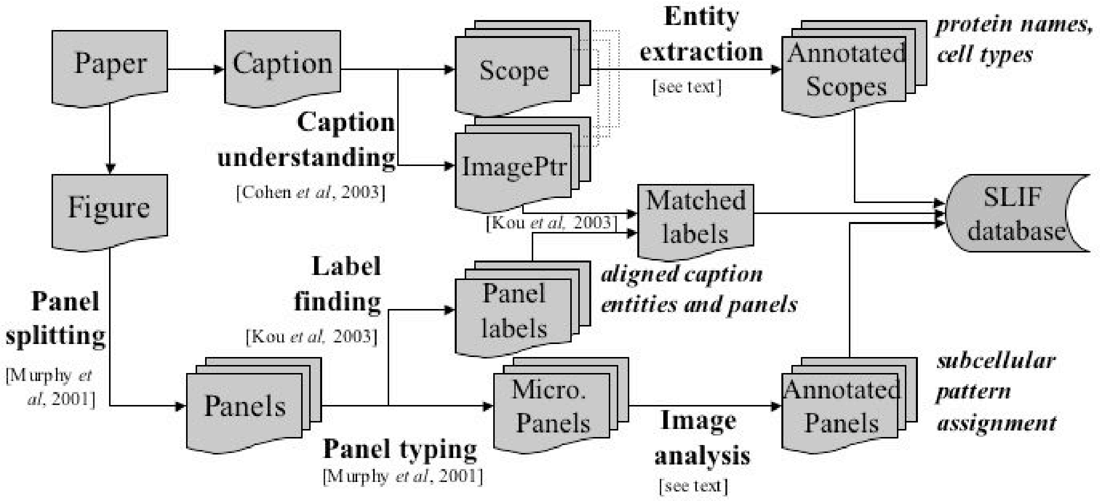


Figure 1. Overview of the image and text processing steps in SLIF.

table that has an entry for each of the figures in a given paper.

4.1 Figure processing

Panel splitting: As discussed above and seen in the lower path in Fig. 1, the processing of figures involves first splitting a figure into its component panels. For figures composed of multiple micrographs (which usually have a dark background with light areas showing where fluorescence was detected), panel splitting is straightforward. We have described a recursive algorithm for finding the light boundaries between micrographs even when the panels are not arranged in a symmetric pattern [5]. This algorithm achieved a precision of 73% and a recall of 60% on an arbitrarily-chosen collection of 100 figures from Pubmed Central articles. We have more recently made some slight adjustments in the algorithm and evaluated it again using 25 figures from articles in the PNAS test set. The results were a precision of 76% and a recall of 60%.

Two factors complicate the panel splitting process in the general case of a journal figure. The first is the mixing of different types of panels in the same figure, such as a graph (with dark lines on a white background) next to a micrograph. Our recursive algorithm that looks for nearly-white or nearly-black horizontal or vertical regions to cut along usually results in separation of the graph into multiple "panels" (most often, one for the body of the graph and one for each of the axis legends). Since our algorithm works well for separating micrographs from the remainder of the figure, this is not a major problem for SLIF but will need to be addressed in the future for more general article processing systems. The second factor is absence of standards for the placement of panel labels (i.e., letters linking the panel to information in the caption) relative to the panels themselves. When panel labels are contained within the panel (as is most common for the micrographs sought by SLIF), the label remains associated with the panel after splitting. However, when the labels are adjacent to the panel they are removed during splitting, and assigning them to their associated panel is non-trivial given the absence of conventions for their placement.

The result of panel splitting is a set of image files for each of the pieces of the original figure and an assertion of the form "Figure *f* contains panels *1-6*" stored as entries in a *panel table*. Additional entries in the *panel table* are populated below.

Recognizing fluorescence micrographs: After panels have been isolated, the next task is to recognize those that contain fluorescence micrographs. This is done using gray level frequencies and a k-nearest neighbor classifier that achieved a 100% precision and 90% recall in testing on Pubmed Central articles [5]. The result is stored as a Boolean field in the *panel table*.

Analyzing and removing panel annotations: Fluorescence micrographs typically have three types of annotations in their body. The first is a label (usually a single letter) that connects the panel to information in the caption. The second is a scale bar whose length is typically defined in the caption. The third is text or symbols that call attention to specific locations in the figure, usually in association with an explanation in the caption or body text. All of these annotations must be removed from the image before the pattern it contains can be analyzed, and the first two types of labels must be recognized and interpreted before removal. We have described image processing steps for finding the panel labels and matching them with parts of the caption [7] and for determining the panel scale (in microns per pixel) by finding the scale bar in the figure and the length definition in the caption [5]. The results of these steps are also stored in the *panel table*.

Subcellular pattern classification: As discussed above, we have described automated systems that can recognize major subcellular patterns. These were restricted to analyzing images containing single cells, but many (if not most) micrographs in journal articles contain images of more than one cell. The initial version of SLIF therefore used watershed segmentation in an attempt to segment each panel into single cell regions. While this worked well for protein patterns that are distributed through most of the cell (such as tubulin), it often failed for proteins that are not. Recently, we have described a subset of the single cell features that do not require single cell segmentation and showed that these can be used to classify multicell images with high accuracy [8]. We have therefore recently incorporated the multicell classifier into SLIF. This allows each panel to be classified based on the major subcellular locations it contains. Note that this may fail if the panel actually contains a mixture of different patterns.

4.2 Caption processing

Named entity recognition: The initial version of SLIF focused on finding micrographs that depicted a particular pattern, but could not associate that pattern with a specific protein. Information on the protein depicted in a given figure should be provided in its caption, but the structure of captions can be quite complex (especially for multi-panel figures). We therefore implemented a system for processing captions with three goals: identifying the "image pointers" (e.g., "(A)") in the caption that refer to panel labels in the figure [9], dividing the caption into fragments (or "scopes") that refer to an individual panel or the entire figure, and recognizing protein and cell names.

The next step is to match the image pointers to the panel labels found during image processing. The accuracy of this matching can be reduced by errors in optical character recognition, but we can compensate for at least some of these errors by using regularities in the arrangement of the labels (such as the likelihood that if the letters A through D are found as image pointers and if the panel labels are recognized as A,B,G and D, then the G should be corrected to a C). Using the PNAS test set, the precision of the final matching process was found to be 83% and the recall to be 74% [7].

The recognition of named entities (such as protein and cell names) in free text is a difficult task that may be even more difficult in condensed text such as captions. In the current version of SLIF, we have implemented two schemes for recognizing protein names. The first uses prefix and suffix features along with immediate context to identify candidate protein names. This approach has a low precision but an excellent recall (which is useful to enable database searches on abbreviations or synonyms that might not be present in structured protein databases). The second approach (Kou, Murphy & Cohen, in preparation) uses a dictionary of names extracted from protein databases in combination with soft match learning methods to obtain a recall and precision above 70%. The protein names found by this approach are entered in the *protein table*, along with a link to the supporting dictionary entry. The occurrences of the names found in the captions are stored in the *protein_in_figure table* and the *protein_in_panel table*, depending on the scope in which the protein name was found.

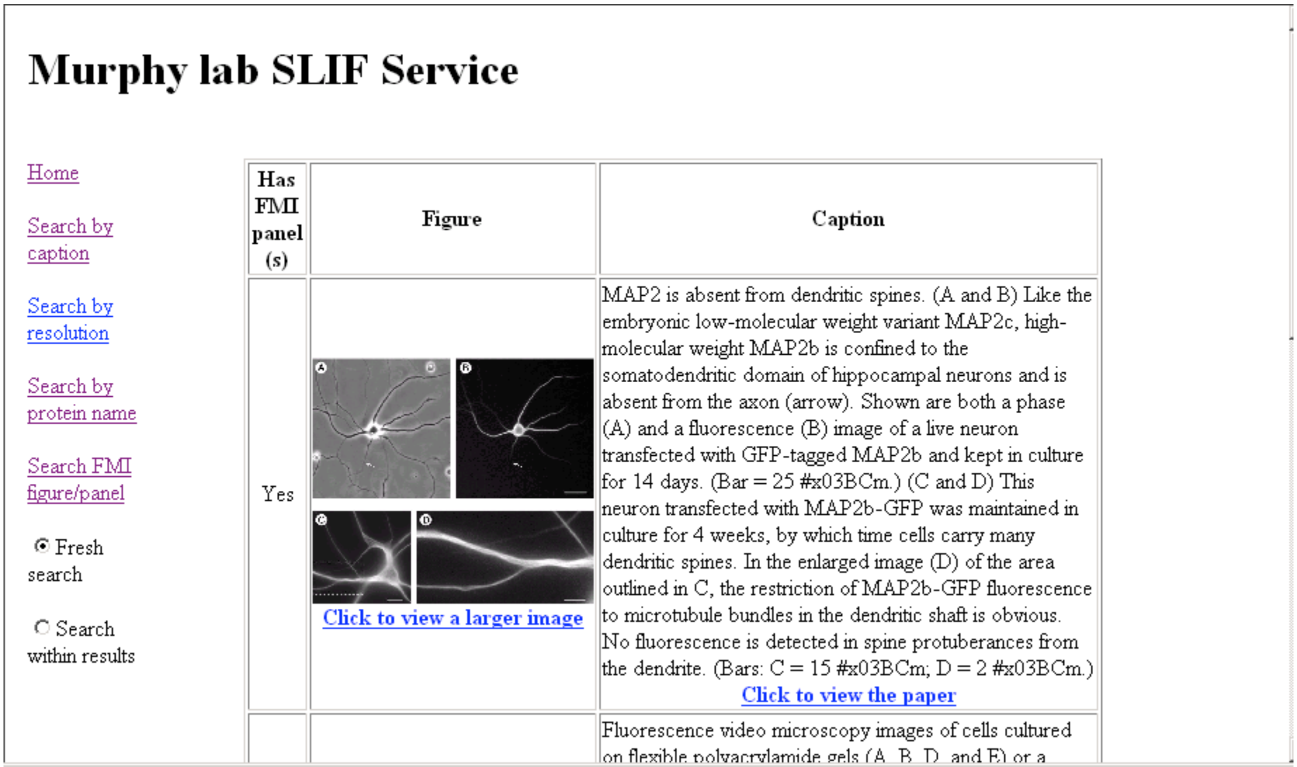
4.3 Database searching

The SLIF database resulting from processing of a corpus of articles with the above methods can be searched by standard SQL queries. We have implemented a number of common queries using Java Server Pages (see Figure 2). Examples include searching for figures or panels with a specific protein name, subcellular pattern, or microscope images with a particular spatial resolution (pixel size in the sample plane). 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.

5. Implications for Publishing Practices

The difficulties encountered (and in some cases only partially overcome) so far during the development of SLIF suggest a number of ways in which the publication of electronic journal articles could be modified to facilitate automated information extraction. These can be implemented with no or minimal impact on the printed journal article or electronic copies intended for human viewing. This can be done by incorporating them into XML structures without having them be visible upon normal display of the article. The improvements include:

- Specification of the coordinates of each panel as pixel numbers within the figure
- Specification of the type (and sub-type) of each panel (e.g., graph, picture:micrograph, picture:gel image)
- Placement of all panel annotations in a separate image layer
- Development of conventions for scale bar use or, preferably, inclusion of information on microns per pixel for each panel
- Inclusion of a URL to get an uncompressed figure



Murphy lab SLIF Service

Home

Search by caption

Search by resolution

Search by protein name

Search FMI figure/panel

● Fresh search

○ Search within results

Has FMI panel (s)	Figure	Caption
Yes	Click to view a larger image	MAP2 is absent from dendritic spines. (A and B) Like the embryonic low-molecular weight variant MAP2c, high-molecular weight MAP2b is confined to the somatodendritic domain of hippocampal neurons and is absent from the axon (arrow). Shown are both a phase (A) and a fluorescence (B) image of a live neuron transfected with GFP-tagged MAP2b and kept in culture for 14 days. (Bar = 25 #x03BCm.) (C and D) This neuron transfected with MAP2b-GFP was maintained in culture for 4 weeks, by which time cells carry many dendritic spines. In the enlarged image (D) of the area outlined in C, the restriction of MAP2b-GFP fluorescence to microtubule bundles in the dendritic shaft is obvious. No fluorescence is detected in spine protuberances from the dendrite. (Bars: C = 15 #x03BCm; D = 2 #x03BCm.) Click to view the paper
		Fluorescence video microscopy images of cells cultured on flexible polyacrylamide gels (A, B, D, and E) or a

Figure 2. Example SLIF web page showing results of sequential queries on the PNAS test set. The initial query was for figures whose captions contained “microtubule,” “mt,” or “tubulin.” The results of that query were then searched for figures classified as containing a fluorescence microscope image (FMI). Only the first of many figures matching both queries is shown. (Note that the figure is an example of a multi-panel figure in which the panels are not aligned in a regular pattern.)

- Inclusion of scoping markup in captions to identify which portions of the caption refer to which panels
- Inclusion of database links for named entities such as proteins

We believe that making an appropriate automated editing tool readily available to the scientific community would permit the additional XML-encoded annotations described above to be generated with minimal extra effort by authors.

6. Conclusion

We have described a system that extracts information on one particular aspect of biology from a combination of text and images in journal articles. The system includes methods for analyzing both images and text, and also for associating information extracted from images with that extracted from the accompanying caption text. The system can be used to find and display potentially relevant images on the basis of text and/or image content. It can also be used to create a structured database of image information, which allows integration of the information contained within structured biological databases with the information contained in article images–images that often provide the primary data presented in an article.

7. Acknowledgements

We thank Bruce Alberts, Ken Fulton and David Lipman for their help in obtaining access to the test set of articles from the Proceedings of the National Academy of Science. The work described here was supported in part by research grant 017396 from the Commonwealth of Pennsylvania Department of Health and by NIH K25 grant DA017357-01.

References

[1] M.V. Boland, M.K. Markey, & R.F. Murphy, Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images, Cytometry, *33 (3),* 1998, 366-375.
[2] M.V. Boland & R.F. Murphy, A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope

Images of HeLa Cells, Bioinformatics*, 17 (12),* 2001, 1213-1223.

[3] R.F. Murphy, M. Velliste, & G. Porreca, Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images, Journal of VLSI Signal Processing*, 35 (3),* 2003, 311-321.

[4] K. Huang & R.F. Murphy, Boosting Accuracy of Automated Classification of Fluorescence Microscope Images for Location Proteomics, BMC Bioinformatics*, 5,* 2004, 78.

[5] 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.

[6] W.W. Cohen, Infrastructure Components for Large-Scale Information Extraction Systems, Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico, 2003, 71-78.

[7] 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.

[8] K. Huang & R.F. Murphy, Automated classification of subcellular patterns in multicell images without segmentation into single cells, 2004 IEEE International Symposium on Biomedical Imaging (ISBI-2004), 2004, 1139-1142.

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