Structured literature image finder: Open source software for extracting and disseminating information from text and figures in biomedical literature

Abdul-Saboor Sheikh, Amr Ahmed, Andrew Arnold, Luis Pedro Coelho, Joshua Kangas, Eric P. Xing, William Cohen, and Robert F. Murphy

October 2009
CMU-CB-09-101

RAY AND STEPHANIE LANE
Center for Computational Biology

Carnegie Mellon

Structured literature image finder: Open source software for extracting and disseminating information from text and figures in biomedical literature

[1]**Abdul-Saboor Sheikh** [2,3]**Amr Ahmed** [2]**Andrew Arnold** [1,4,5]**Luis Pedro Coelho** [1,4,5]**Joshua Kangas** [1,2,3,4,5,6]**Eric P. Xing** [1,2,3,4,5]**William W. Cohen** [1,2,4,5,6,7]**Robert F. Murphy**

October, 2009
CMU-CB-09-101

Lane Center for Computational Biology
Carnegie Mellon University
Pittsburgh, PA 15213

[1]Center for Bioimage Informatics, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Language Technologies Institute, Carnegie Mellon University
[4]Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology
[5]Lane Center for Computational Biology, Carnegie Mellon University
[6]Department of Biological Sciences, Carnegie Mellon University
[7]Department of Biomedical Engineering, Carnegie Mellon University

Previous work on SLIF was supported by NIH grant CA83219, NSF Science and Technology Center grant MCB-8920118, research grant 017396 from the Commonwealth of Pennsylvania Department of Health, and NIH grant K25 DA017357-01. Facilities and infrastructure support have been provided by NSF grant EF-0331657, NIH grant U54 RR022241, and by NIH grant U54 DA021519. The work described in this report was primarily supported by NIH grant R01 GM078622.

Keywords: Automated Image Analysis, Biomedical Literature, Data and Image Mining, Figure and Caption Modeling, Information Retrieval, Machine Learning, Natural Language Processing

Abstract

The SLIF project combines text-mining and image processing to extract structured information from biomedical literature.

SLIF extracts images and their captions from published papers. The captions are automatically parsed for relevant biological entities (protein and cell type names), while the images are classified according to their type (e.g., micrograph or gel). Fluorescence microscopy images are further processed and classified according to the depicted subcellular localization. The results of this process can be queried online using either a user-friendly web-interface or an XML-based web-service. As an alternative to the targeted query paradigm, SLIF also supports browsing the collection based on latent topic models which are derived from both the annotated text and the image data.

In addition to a description of the SLIF system, this technical report describes the hand-labeled datasets used for training SLIF components. These datasets, and the SLIF web application, are publicly available at http://slif.cbi.cmu.edu.

1 Introduction

Biomedical research worldwide results in a very high volume of information in the form of publications. Biologists are faced with the daunting task of querying and searching these publications to keep up with recent developments and to answer specific questions.

In the biomedical literature, data is most often presented in the form of images. A fluorescence micrograph image (FMI) or a gel is sometimes the key to a whole paper. Compared to figures in other scientific disciplines, biomedical figures are frequently a stand alone source of information that summarizes the finding of the research under consideration. A random sampling of such figures in the publicly available PubMed Central database reveals that in some, if not most of the cases, a biomedical figure can provide as much information as a normal abstract. The information-rich, highly-evolving knowledge source of the biomedical literature calls for automated systems that would help biologists find information quickly and satisfactorily. These systems should provide biologists with a structured way of browsing the otherwise unstructured knowledge in a way that would inspire them to ask questions that they never thought of before, or reach a piece of information that they would have never considered pertinent to start with.

Relevant to this goal, our team developed the first system for automated information extraction from images in biological journal articles (the "Subcellular Location Image Finder," or SLIF, first described in 2001 [1]).

Since then, we have reported a number of improvements to the SLIF system [2, 3, 4, 5, 6, 7]. Moreover, we recently made major enhancements and additions to the system in response to the opportunity to participate in the Elsevier Grand Challenge. In part reflecting this, we rechristened SLIF as the "Structured Literature Image Finder." The new SLIF provides both a pipeline for extracting structured information from papers and a web-accessible searchable database of the processed information. Users can query the database for various information appearing in captions or images, including specific words, protein names, panel types, patterns in figures, or any combination of the above. We have also added a powerful tool for organizing figures by topics inferred from both image and text, and have provided a new interface that allows browsing through figures by their inferred topics and jumping to related figures from any currently viewed figure.

2 Overview

SLIF consists of a pipeline for extracting structured information from papers and a web application for accessing that information. The SLIF pipeline is broken into three main sections: caption processing, image processing and topic modeling, as illustrated as Figure 1.

The pipeline begins by finding all figure-captions pairs and creating database entries for each. Each caption is then processed to identify biological entities (e.g., names of proteins and cell lines) and these are linked to external databases (e.g., UniProt). Pointers from the caption to the image are identified, and the caption is broken into "scopes" so that terms can be linked to specific parts of the figure.

The image processing section begins by splitting each figure into its constituent panels, and then identifying the type of image contained in each panel. The original SLIF system was trained

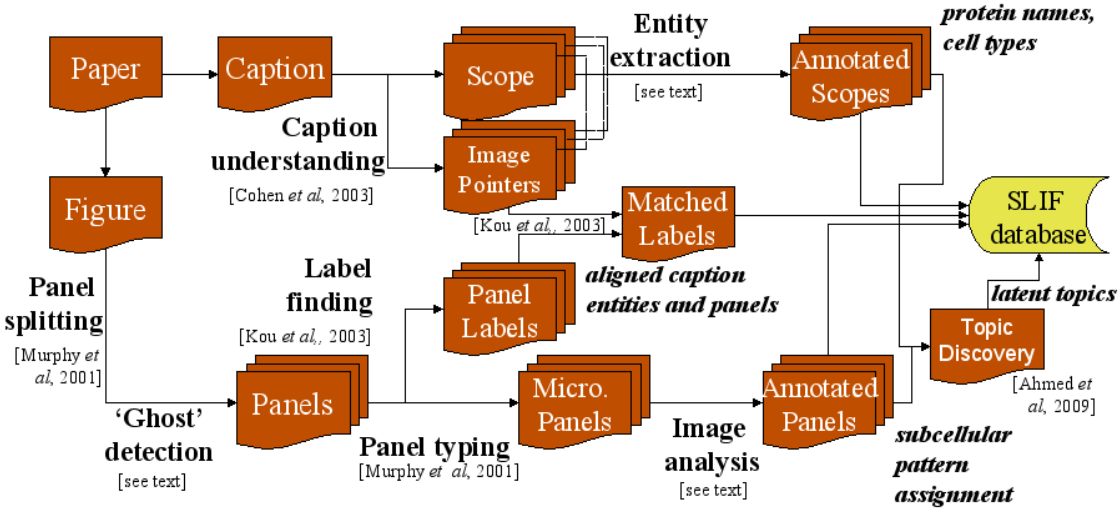


Figure 1: SLIF Pipeline. This figure shows the general pipeline through papers are processed.

to recognize only those panels containing fluorescence microscope images (FMIs), but as part of the work for the Elsevier Grand Challenge we have extended SLIF to recognize other types of panels. The patterns in FMIs are then described using a set of biologically relevant image features [1], and the subcellular location depicted in each image is recognized.

The first two sections result in panel-segmented, structurally and multi-modally annotated figures. The last step in the pipeline is to discover a set of latent themes that are present in the collection of papers. These themes are called topics and serve as the basis for visualization and semantic representation. Each topic consists of a triplet (possibly extended to higher order) of distributions over words, image features and proteins (possibly extended to include GO terms and subcellular locations as well). For instance, a topic about "tumorigenesis" is expected to give high probability to words like ("tumor", "positive", "hlb") in its distribution over words, and similarly to proteins like ("Caspase", "Actin") which are known to be related to the tumorigenesis processes. Each figure in turn is represented as a distribution over these topics, and this distribution reflects the themes addressed in the figure. This representation serves as the basis for various tasks like image-based retrieval, text-based retrieval and multimodal-based retrieval. Moreover, these discovered topics provide an overview of the information content of the collection, and structurally guide its exploration: for instance, the biologist might ask the system to retrieve articles that have figures in which the "tumorigenesis" topic is highly represented.

All results of processing are stored in a database, which is accessible via web interface or XML (SOAP) queries. The results of queries always include links back to the panel, figure, caption and the full paper.

3 Database Access

3.1 Web Interface

The results of processing papers are stored in a searchable database and are made available to the user through an interactive web-interface. The interface permits the user to query the database in a variety of ways. A user can query the database for:

- Text within captions
- Proteins extracted by protein name annotators
- Different properties of the image panels (depicted protein or cell type, panel type, subcellular location)
- Subcellular locations in images retrieved from GO (Gene Ontology) term database
- Images of specific resolution
- Viewing and browsing the latent topics discovered from figures and captions
- Any combination of the above

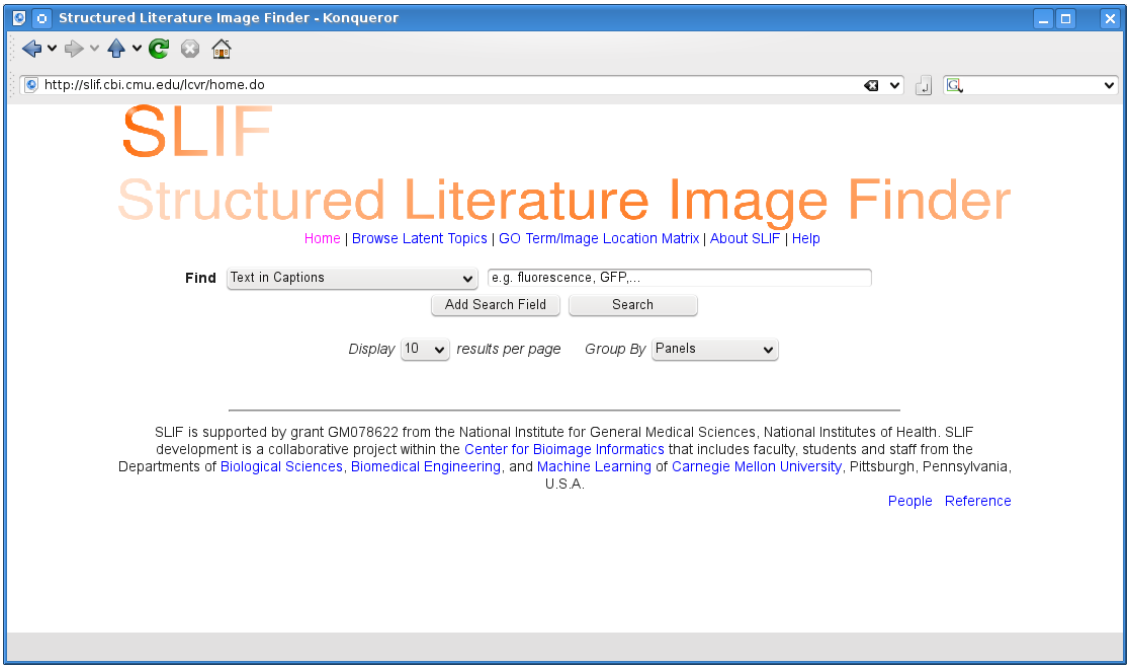


Figure 2: Screenshot of new SLIF home page.

From the results of a search, users can view the underlying papers or the UniProt record corresponding to an extracted protein name. They can also refine the search results by adding more conditions. The interface also incorporates relevance feedback, which further allows users to interactively filter search results by marking certain results as ”interesting” and asking the system to show results that are similar to them. Alternatively, users can also browse results clustered as latent topics (Section 6).

3.1.1 Interface Revision

The user interaction has been reworked from the ground up during the Elsevier Competition to provide both more functionality and a better user experience. The interaction with the database is now more similar to other search engines, which should be familiar to users. Figure 2 shows the new home page.

The new interface also makes it easier to ask more sophisticated questions. For example, a user can at once search for any exclusive combination of information (e.g., a particular panel type with a given depicted protein), while specifying multiple keys for each of the information fields (e.g. panel types FMI or Gel, locations Nuclear or Punctate etc.). This is achieved from the home page by adding more search terms.

Moreover, the new interface also allows refining the current search (this was the main mode in which complex queries were built in the old interface). A user can opt to perform any subsequent search in nested fashion by checking “within current results” box that appears next to “Search” button after every search (by default, this box is unchecked).

The input fields in the interface interactively assist the user by showing and auto-completing the possible options for the active search type (e.g., if the user is attempting to type in a protein name, she will see a list of proteins in the database as suggestions). This is implemented for all search fields excluding free text search. If the user has selected the "within current results" option, the suggestions are confined to those that are meaningful in these results (a protein that does not occur in these results is not displayed, for example).

For displaying results, the new interface has revised Paper and Panel views, as well as a new Caption-Figure view.

The Panels layout is the most detailed view in that it displays all the information pertaining to a panel (type, resolution, locations etc.). In this layout, user can also reorder the results by various fields. Captions and figures with which the panels are associated are by default not visible in 'Panels' view; however, for a specific record, corresponding caption and figure can be viewed by clicking on the record's 'Show Caption/Figure' link. Figure 3 shows an example of this interaction.

Caption-Figure and Paper views give summarized views by displaying the entire information extracted from caption-figure pairs and papers respectively. The 'Papers' layout also comprises 'Show Caption/Figure' links for viewing captions and figures in a paper.

The user can interactively switch between these views and the same set of results will be redisplayed in the requested view.

All the layouts let the user select interesting figures or panels and query for similar entities using Relevance Feedback. Furthermore, the query results always contain links to the original papers in addition to other links pointing to the latent topics associated with a figure, a panel or a protein, or links to UniProt for specific proteins.

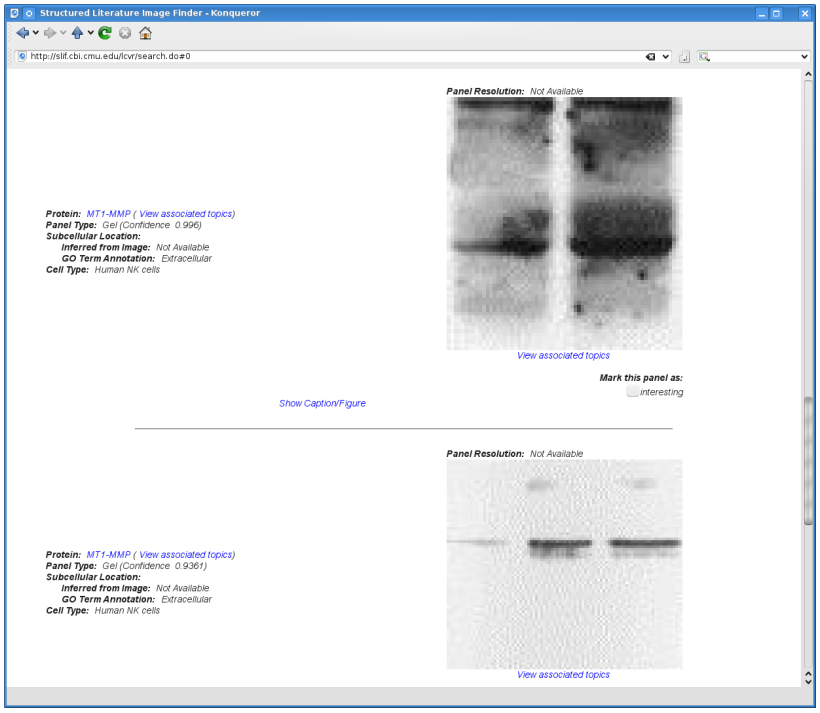
3.2 Machine-Accessible Interface

We also make the results available via web service architecture. This enables other machines to consume SLIF results in automated fashion. For a set of processed results, we publish a WSDL (Web Services Description Language) document on the SLIF server that declares the database query procedure for clients in standard XML based description language. We have defined this query procedure to take an XML structured SLIF query as its argument and perform the relevant database search. Hence a client can remotely invoke the procedure by defining its query parameters in XML format and embedding it in a SOAP (Simple Object Access Protocol) standard message, and then sending the message to the SLIF server. The server then returns to the client the XML format search results that are returned by the procedure.

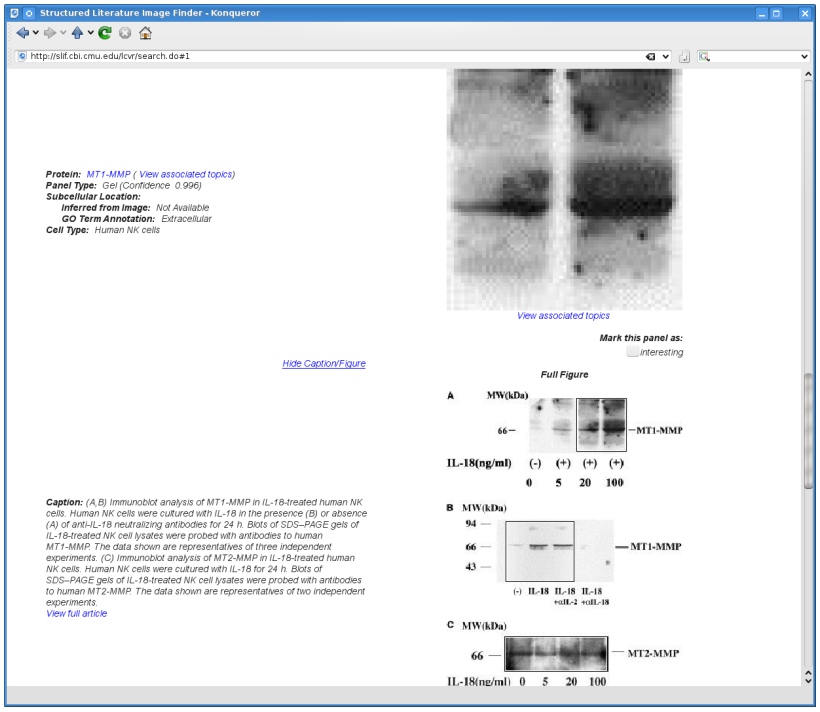
4 Caption Processing

4.1 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



(a)



(b)

Figure 3: Screenshot of Panel View. Top panel shows the Panel view as returned by a query for gels of "MT1-MMP." The bottom panel shows the result of expanding the caption and figure to show more detail on the first result.

(especially for multipanel figures). We therefore implemented a system for processing captions with three goals: identifying the image pointers (e.g., (A) or (red)) in the caption that refer to specific panel labels or panel colors in the figure [2], dividing the caption into fragments (or scopes) that refer to an individual panel, color, 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). This type of matching is implemented as a combination of dynamic programming and stacked learning using graphical models [5]. Using a set of labeled captions from PNAS, the precision of the final matching process was found to be 83% and the recall to be 74% [4].

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) [8]. The second approach uses exact matching to a dictionary of names extracted from UniProt protein databases to obtain 51% precision and 22% recall. The protein names found by this approach can be associated with a supporting protein database entry. Both approaches combined yield a precision of 40% with 44% recall[1].

Finally, the task of simply segmenting a paper and extracting the caption, even without named entity recognition or panel scoping, has proven very useful to our users, allowing easy search of free text which can be limited to the captions, and therefore the figures, of a paper.

4.1.1 Remote Tagging

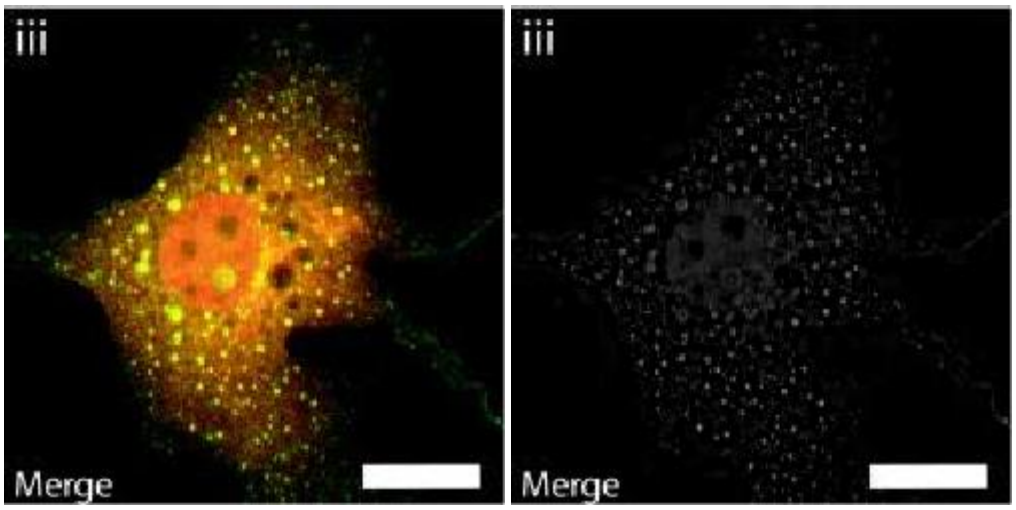
Recently we have also added an interface to Reflect [9], through which we annotate the captions for protein entities.

5 Image Processing

5.1 Figure Splitting

The first step in our image processing pipeline is to divide the extracted figures into their constituent components, since in majority of the cases (nearly in all the cases of our interest), the figures are comprised of multiple panels to depict similar conditions, corresponding analysis, etc. For this purpose, we employ a figure-splitting algorithm that recursively finds constant-intensity boundary regions in between panels. These projections are calculated by summing up the pixel values of

[1]The corpus of labeled captions used for classifier training and performance assessment of protein entity recognition is publicly available at http://slif.cbi.cmu.edu.



(a) Color image (b) Blue channel

Figure 4: Example of a ghost image. Although the color image is obviously a two-channel image (red and green), there is a strong bleed-through into the blue component.

regions of a figure along both horizontal and vertical directions. We have previously shown that the algorithm can effectively split figures with complex panel layouts [1]. The algorithm yields subimages which are stored as panels. The remaining small subimages often contain useful textual information such as panel labels, and are stored for later scoping between panels and caption text.

5.2 "Ghost" Detection

FMI panels are often false color images composed of related channels. However, due to treatment of the image for publication or compression artifacts, it is common that an image that contains one or two logical colors (and is so perceived by the human reader), will have signal in all 3 color channels. The extra channel, we call a "ghost" of the signal-carrying channels. Figure 4 exemplifies this phenomenon.

To detect ghosts, we first compute the white component of the image, i.e., the pixel-wise minimum of the 3 channels. We then subtract this component from each channel so that the regions with homogeneous intensities across all channels (e.g., annotations or pointers) get suppressed. Then, for each channel, we verify if the 95%-percentile pixel is at least 10% of the overall highest pixel value. These two values were found empirically to reject almost all ghosts, with a low rate of false negatives (a signal carrying channel that has less than 5% bright pixels will be falsely rejected, but we found the rate of false positives to be low enough to be acceptable). Algorithm 1 illustrates this process in pseudo-code.

Algorithm 1: Ghost Detection Algorithm

White := pixelwise-min(R,G,B)
M := max(R−White,G−White,B−White)
foreach *ch* ∈ *(R,G,B)* **do**
    Residual := ch−White
    sort pixels from Residual
    if $\frac{\textit{95\% highest pixel}}{M} < 10\%$ **then**
        ch is a ghost

5.3 Panel Type Classification

SLIF was originally designed to process only FMI panels. As part of our work for the Elsevier challenge, we expanded the classification to other panel types. This mirrors other systems which have appeared since the original SLIF to include more panel types [10, 11, 12].

We have manually labeled circa 700 panels into six panel classes: (1) FMI, (2) gel, (3) graph or illustration, (4) light microscopy, (5) X-ray, or (6) photograph. The labeled panels were selected through active learning, using the algorithm presented by Roy and McCallum [13] of empirical risk reduction. We used a libSVM-based classifier as the base algorithm. In order to speed up the process, at each round, we labeled the 10 highest ranked images plus 10 randomly selected images. The process was seeded by initially labeling 50 randomly selected images.

To illustrate that our system can handle other types of images, we decided to concentrate our efforts on creating a high-quality classifier for the *gel* class, given its importance to the working scientist. Towards this goal, we define a set of features based on whether certain marker words appeared in the caption that would signal gels[2] as well as a set of substrings for the inverse class[3]. A classifier based on these boolean features was obtained by the ID3 decision tree learning algorithm [14] with precision on the positive class as the target function. This technique was shown, through 10 fold cross-validation, to obtain very high precision (91%) at the cost of moderate recall (66%). Therefore, examples considered positive are labeled as such, but examples considered negative are passed on to a classifier based on image features. In addition to the features developed for FMI classification, we introduce a measure of how horizontal the image is, as the fraction of variance that remains in the image formed by the differences between horizontally adjacent pixels:

$$h(I) = \frac{\mathrm{var}(I_{i-1,j} - I_{i,j})}{\mathrm{var}(I_{i,j})}. \tag{1}$$

Gels, consisting of horizontal bars, score much lower on this measure than other types of images. Furthermore, we used 26 Haralick texture features [15]. Images were then classified into the six panel type classes using a support vector machine based classifier based on the libSVM system. On this system, we obtain an overall accuracy of 61%.

Therefore, the system proceeds through 3 classification levels: the first level, classifies the image into FMI or non-FMI using image based features; the second level, uses the textual features described above to identify gels with high-precision; finally, if neither classifier has fired, a general purpose support vector machine classifier, operating on image-based features does the final classification.

5.4 Subcellular Location Pattern Classification

Perhaps the most important task that SLIF supports is to extract information based on the subcellular localization depicted in FMI panels.

[2]The positive markers were: *Western*, *Northern*, *Southern*, *blot*, *lane*, *RT* (for "reverse transcriptase"), *RNA*, *PAGE*, *agarose*, *electrophoresis*, and *expression*.

[3]The negative markers were: *bar* (for bar charts), *patient*, *CT*, and *MRI*.

To provide training data for pattern classifiers, we hand-labeled a set of images into four different subcellular location classes: (1) *nuclear*, (2) *cytoplasmic*, (3) *punctate*, and (4) *other*, following the active learning methodology described above for labeling panel types. The active learning loop was seeded using images from a HeLa cell image collection that we have previously used to demonstrate the feasibility of automated subcellular pattern classification [16].

The dataset was filtered to remove images that, once thresholded using the methods we described previously [16], led to less than 80 above-threshold pixels, a value which was empirically determined. This led to the rejection of 4% of images. In classification, if an image meets the rejection criterion, it is assigned into a *don't know* class.

We computed previously described field-level features [17] to represent the image patterns (field-level features do not require segmentation of images into individual cell regions). We added a new feature for the size of the median object (which is a more robust statistic than the previously used mean object size). If the scale is inferred from the image, then we normalize this feature value to square microns. Otherwise, we assume a default scale of $1\mu m/$pixel.

We also adapted the threshold adjacency statistic features (TAS) from Hamilton et. al [18] to a parameter-free version in an analogous way. The original features depended on a manually controlled-two-step binarization of the image. For the first step, we use the Ridler–Calvard algorithm to identify a threshold instead of a fixed threshold [19]. The second binarization step involves finding those pixels that fall into a given interval $[\mu - M, \mu + M]$, where μ is the average pixel value of the above-threshold pixel and M is a margin (set to 30 in the original paper). We set our margin to the standard deviation of above threshold pixels. We call these parameter-free TAS. The other binarizations proposed by Hamilton et. al. were adapted to a parameter-free version.

Feature selection using stepwise discriminant analysis [20] was performed before classifier training. On the 3 main classes (Nuclear, Cytoplasmic, and Punctate), we obtained 75% accuracy (as before, reported accuracies are estimated using 10 fold cross-validation and the classifier used was libSVM based). On the four classes, we obtained 61% accuracy.

5.5 Panel and Scope Association

Panels were associated with their scopes based on the textual information found in the panel itself and the areas surrounding the panels. Each figure is composed of a set of panels and a set of subimages which are too small to be panels. All of these sections are analyzed using optical character recognition (OCR) to identify potential image pointers. The caption of the figure was previously analyzed to find the set of associated image pointers. In the most simple case, the number of panels matches the number of image pointers discovered in the caption. In this case, each panel is matched to the nearest unique image pointer found in the figure using OCR. This enables panels to be directly associated with the textual information found in the caption scope.

6 Topic Discovery

The goal of the topic discovery phase is to enable the user to structurally browse the otherwise unstructured collection. This problem is reminiscent of the actively evolving field of multimedia

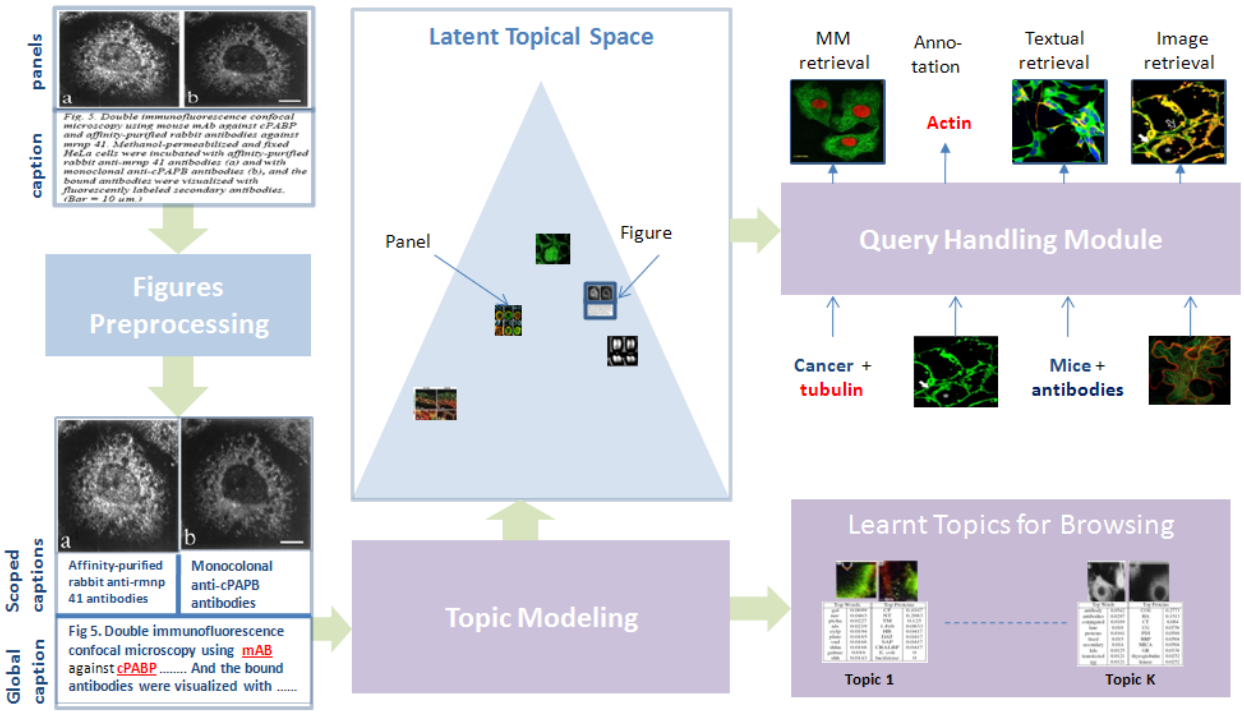


Figure 5: Overview of the topic discovery module, please refer to Section 6 for more details.

information management and retrieval. However, *structurally-annotated* biological figures pose a set of new challenges to mainstream multimedia information management systems that can be summarized as follows:

- **Structured Annotation** As shown in Figure 5, and as discussed in the preceding sections, biological figures are divided into a set of sub-figures called *panels*. This hierarchical organization results in a local and global annotation scheme in which portions of the caption are associated with a given panel via the panel pointer (e.g., "(a)" in Figure 5), while other portions of the caption are shared across all the panels and provide contextual information. We call the former a *scoped* caption, while we call the latter a *global* caption. How can this *annotation* scheme be modeled effectively?

- **Free-Form Text** unlike most associated text-image datasets, the text annotation associated with each figure is free-form text as opposed to high-quality, specific terms that are highly pertinent to the information content of the figure. How can the relevant words in the caption be discovered automatically?

- **Multimodal Annotation** although text is the main source of modality associated with biological figures, the figure's caption contains other entities like protein names, GO-term locations and other gene products. How can these entities be extracted and modeled effectively?

We addressed the problem of modeling structurally-annotated biological figures by developing a model that we call structured correspondence topic model, that addresses the aforementioned challenges. A full description of this model has been presented [21]. Figure 5 provides an overview of how the topic model module fits in the SLIF overall pipeline. The input to the topic modeling system is the panel-segmented, structurally and multimodally annotated biological figures. The goal of our approach is to discover a set of latent themes in a given paper collection. These themes are called topics and serve as the basis for visualization and semantic representation. Each biological figure, panel, and protein entity is then represented as a distribution over these latent topics. This representation serves as the basis for various tasks like image-based retrieval, text-based image retrieval, multimodal-based image retrieval and image annotation. We compared our model to various baselines over the aforementioned tasks with favorable results [21]. In the following two subsections, we *illustrate* the use of our system in structurally guiding biologists in browsing the otherwise unstructured paper collection.

6.1 Structured Browsing

Topic models endow the user with a bird's eye view over the paper collection as shown in Figure 6. In this figure, each topic represents a theme addressed in the collection. Each theme is summarized along three dimensions: top words, top proteins, and a set of representative panels. If a topic interests the biologist, she can click on the browse button to display all panels (figures) that are relevant to this topic. She can then further navigate to the papers containing these figures.

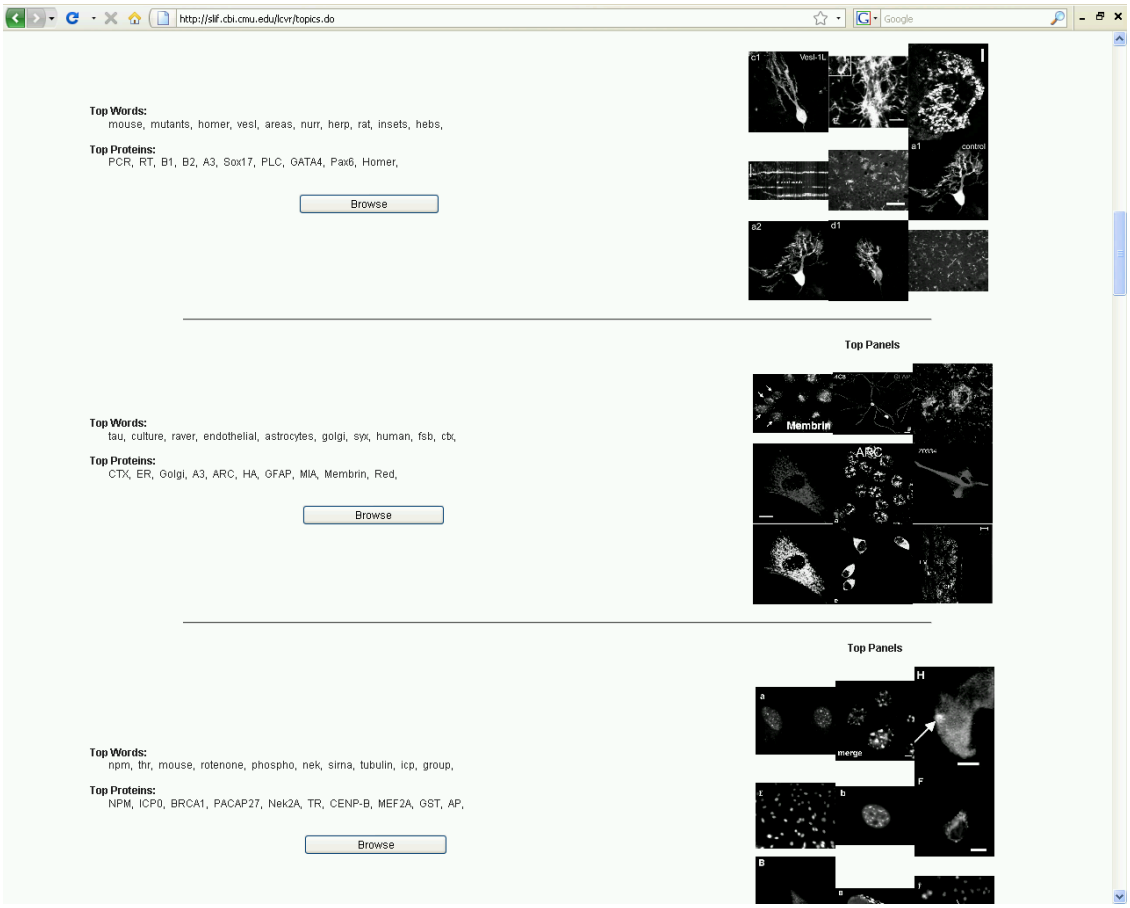


Figure 6: A sample three topics from the list of 25 topics discovered in a given paper collection. Each topic is represented by its top words, top proteins, and a set of representative panels. The user can explore each topic via the *browse* button.

Moreover, if the biologist has a focused search needs, the system can confine the displayed topics to those topics associated with panels (figures) that interest the biologist. For instance, lets assume that the biologist searched for *high-resolution*, FMI panels that contain the protein *MT1-MMP* as depicted in Figure 7. The biologist can then click the *"view associated topics"* link below the displayed panel. The system will display only the topics addressed in this panel as shown in Figure 8. One of these topics shows a set of proteins (MT1, ACAT, ACAT-1) that seems interesting to the biologist. The biologist can then browse for more panels that show the pattern(s) captured by this topic by clicking on the browse button. The result of this action is displayed in Figure 9.

6.2 Relevance Feedback: Topic-based Similarity Browsing

In addition to providing the biologist with a breakdown of the thematic (topical) decomposition of each figure or panel in the collection, the system can directly give the user the ability to find similar

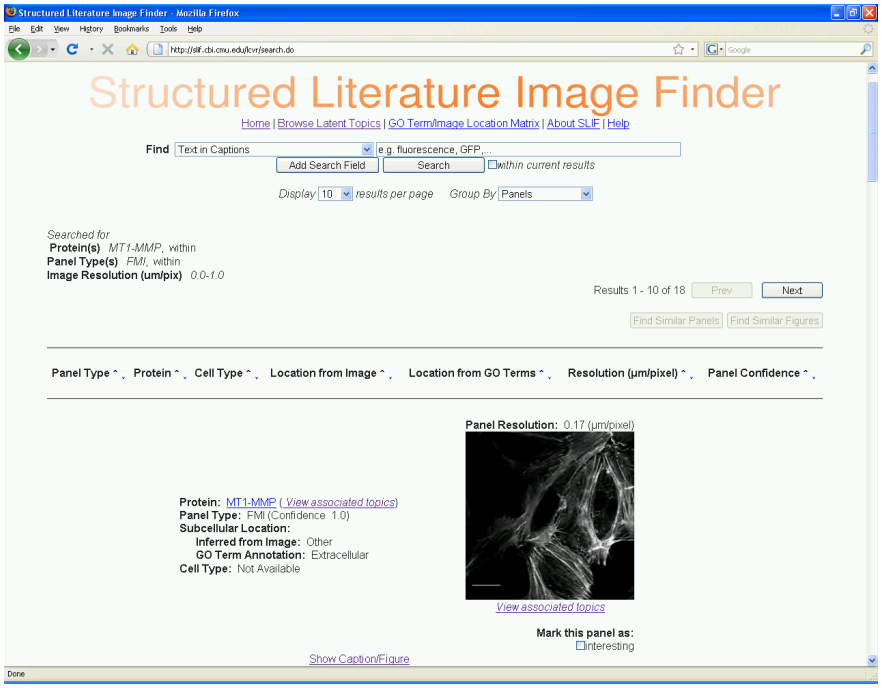


Figure 7: The top panel that results from searching for high-resolution, FMI images that show localization of the *MT1-MMP* protein.

panels (figures) based on their collective latent representation over topics. For instance, consider the panel in Figure 7. Instead of viewing its latent topic decomposition as shown in Figure 8 and then deciding which topic is more interesting to pursue, the biologist might want to avoid this route and instead asks a more direct question: *show me more panels like this*, or *find similar panels to this one.* To answer this query, recall that each panel is represented as a point in the topical space. In other words, each panel is represented as a vector where each component corresponds to a topic, and the value of each component indicates how strongly the corresponding topic is expressed in the panel. Therefore, the biologist's query can be answered by ranking the panels in the database based on the similarity of their latent representations to the latent representation of the query panel. In the current release of SLIF, we used the Euclidean distance as a measure of similarity.

As shown in Figure 7, the user can select the "*interesting*" check-box under the displayed panel, which will enable the button "*Find similar panels*" at the top of Figure 7. Once clicked, the system will retrieve and display a sorted list of panels similar to the query panel as shown in Figure 10. Moreover, the biologist can select multiple panels as interesting and the system will find similar panels based on the collective latent representation of these selected panels. This process can be repeated recursively to refine the search result until a satisfactory result is reached.

Finally, if the biologist switched to the "*figure*" view, she can use the same approach to and search for similar figures using the "*Find similar figures*" button. In this case the system will rank all figures in the database based on similarity of their latent representations to the latent representation of the selected figure(s). For instance, Figure 11 shows the top most similar figure (second

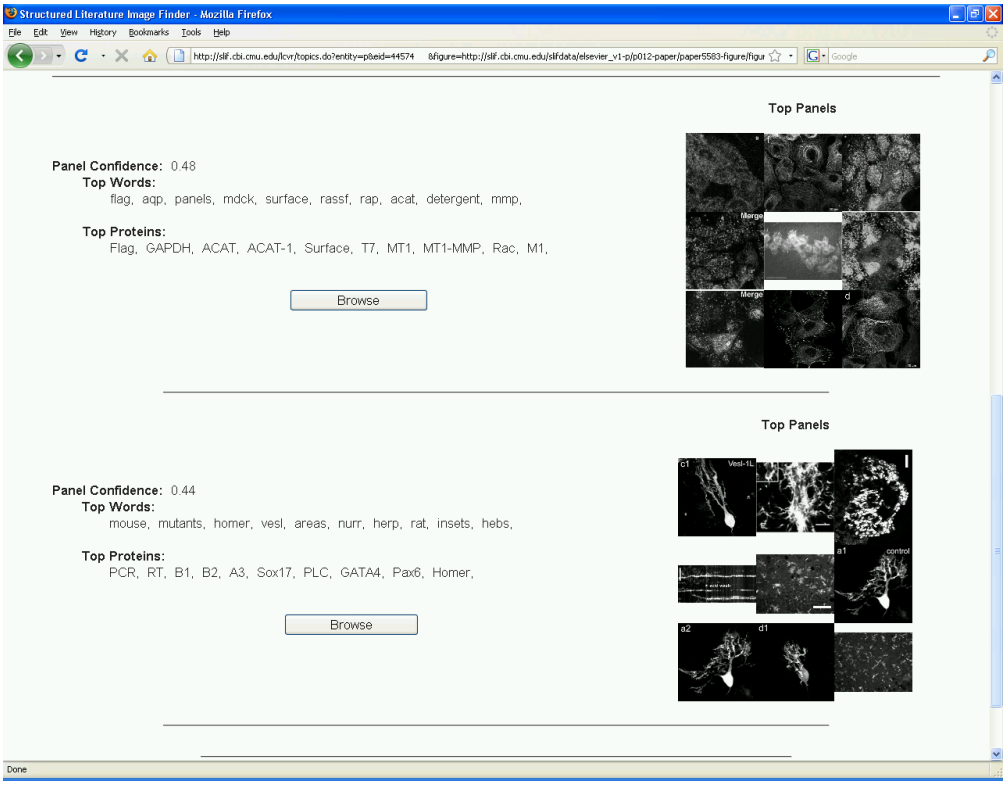


Figure 8: The top two topics addressed in the panel shown in Figure 7. Note the interesting set of proteins in the first topic.

figure) to the *query figure* that the biologist finds interesting (first figure). The utility of this figure view is demonstrated clearly in Figure 11. The first figure is from a study of the time-varying expression level of the PANX1 protein in the mouse retina at different stages of the animal's life. The most similar figure returned by the system is from a study of cone-photoreceptor development in the retina of living zebrafish. Thus these two figure shares a common theme: studying development of the retina albeit in two different species. Is this association interesting? This depends primarily on the biologist's goal, however we would like to emphasize that this approach enables the biologist to find a piece of potentially useful information she might not consider searching for. Moreover, in Figure 12, we show another similar figure returned by the system for the same query figure (the top figure) in Figure 11. In this case, the figure shows *RanBPM* localization in the mammalian retina. Other top figures returned by the system show different aspects of development in the retina or in the nervous system. The system links the query to figures of the nervous system because "PANX1" is also expressed there.

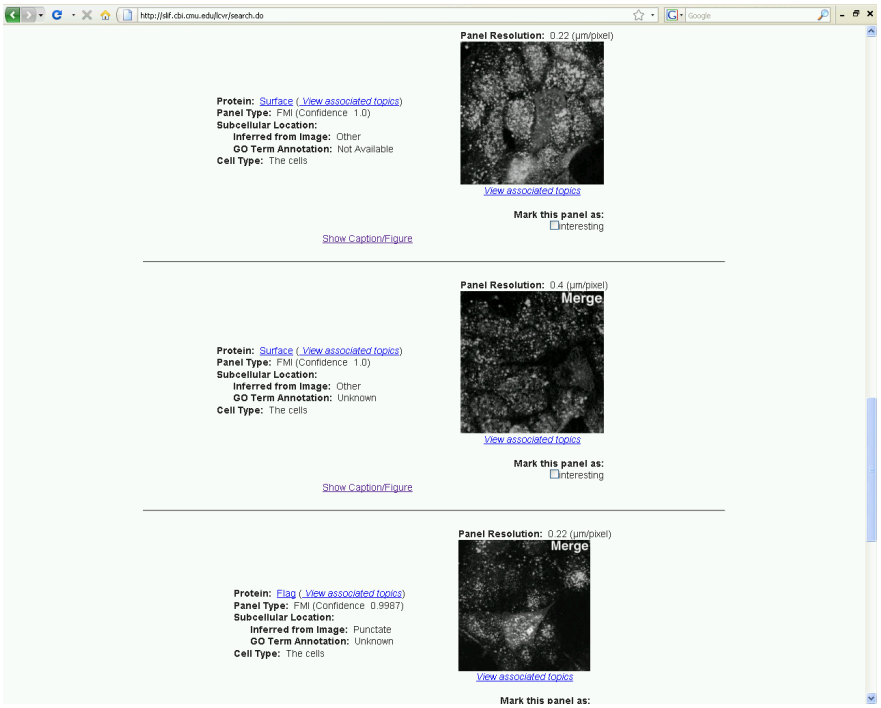


Figure 9: A set of panels that exhibit the same pattern captured by the first topic in Figure 8. This result is displayed after clicking the browse button under the first topic in Figure 8.

7 User Study

As part of the further development of SLIF during the Elsevier Challenge competition, we also conducted a user study to validate the usability and usefulness of our technology. Appendix B contains full documentation of the study, including the documents that were handed out to study participants, and all of their individual responses. Our target users were graduate students in the fields of biology, computational biology, and biomedical engineering. We focused the user study along two dimensions:

- ***Usability***. The goal here is to get reliable answers to the following questions (Appendix B.3):

  1. Is SLIF easy to use? Is the interface intuitive and user-friendly?
  2. Are the results displayed in a clear way that allows the user to view the retrieved information at multiple levels of details?
  3. Can the user pave their way through SLIF with no external help?

- ***Effectiveness***. Aside from the ease of use, we were interested in understanding the effectiveness of our technology. Specifically, were were interested in getting reliable answers to the following questions:

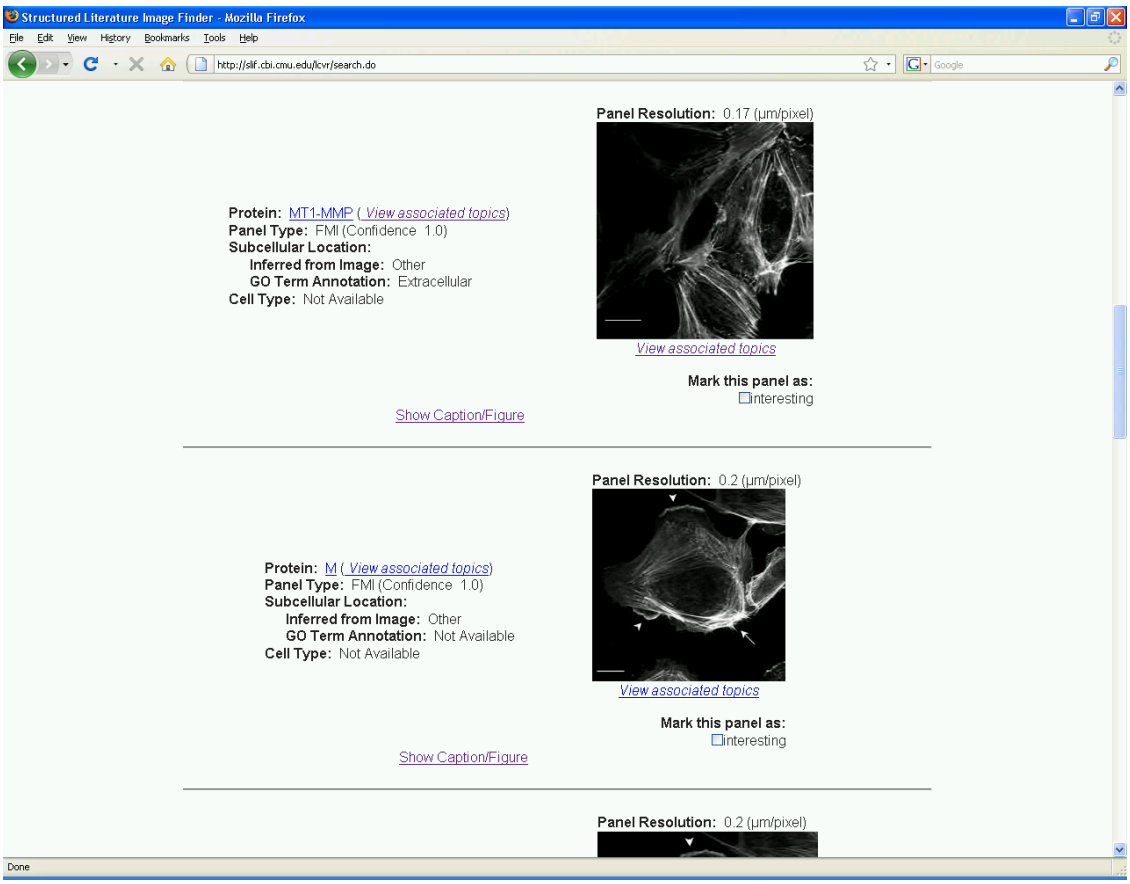


Figure 10: Top similar panels to the panel shown in Figure 7. Note that these panels are different from those shown in Figure 9 — see text for more details.

1. Does SLIF help user finding relevant papers in general as compared to using the user's favorite search engine?
2. Under which search scenarios is SLIF most useful?

Our goal in obtaining answers to the above questions is in fact two-fold. On one hand, we were motivated in building SLIF by what the team, as a group of researchers, thinks is most useful and beneficial. However, our team consists of researchers across the fields of machine learning, text mining, and bio-imaging — indeed, our expertise might not reflect the opinion of all the users targeted by SLIF. On the other hand, we were eager to enhance our system based on the user feedback to make SLIF more usable and effective.

The design of the user study proceeds as follows. Each user was given an instruction sheet Appendix B.3 that lists a set of tasks to be performed using both SLIF and the user's favorite search engine. Moreover, no instructions on how to use SLIF were given to the users; we asked the users to find their own path through the system and discover its capabilities along the way. Each user was given half an hour to finish executing the list of *four* search scenarios using both SLIF and another

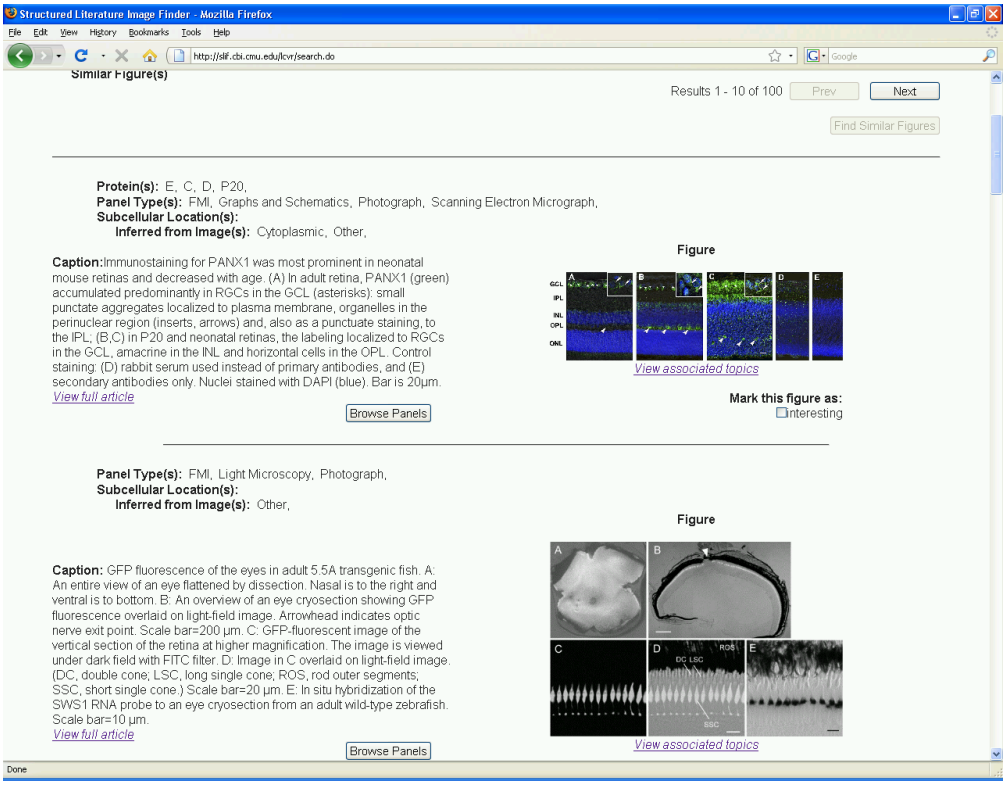


Figure 11: The first biological figure at the top is the query figure. The second biological figure is the most similar figure returned by the system, where similarity is measured based on the latent representation of each figure.

search engine. In addition, the user was given *no separate* or extra time for learning how to use SLIF as we mentioned above — indeed we subjected the SLIF's user interface design to a hard test: the user is asked to compare SLIF against her favorite search engine used on a daily basis, while not even given enough time or instructions on how to use SLIF!

The list of tasks (Appendix B.2) that the user were asked to perform are *summarized* below:

- Search for papers that contain fluorescence images that detail where a given protein (ACAT-1 in our study) locates.
- Search for papers with gel images for any protein you study in your research.
- Search of *high-resolution* flourescence microscopy images of a given protein (Actin in our study)
- Search for papers that have images related to "nuclear translocation."
- Use the "browse by topic", and the "view associated topics" features over the returned result. Did you find it useful?

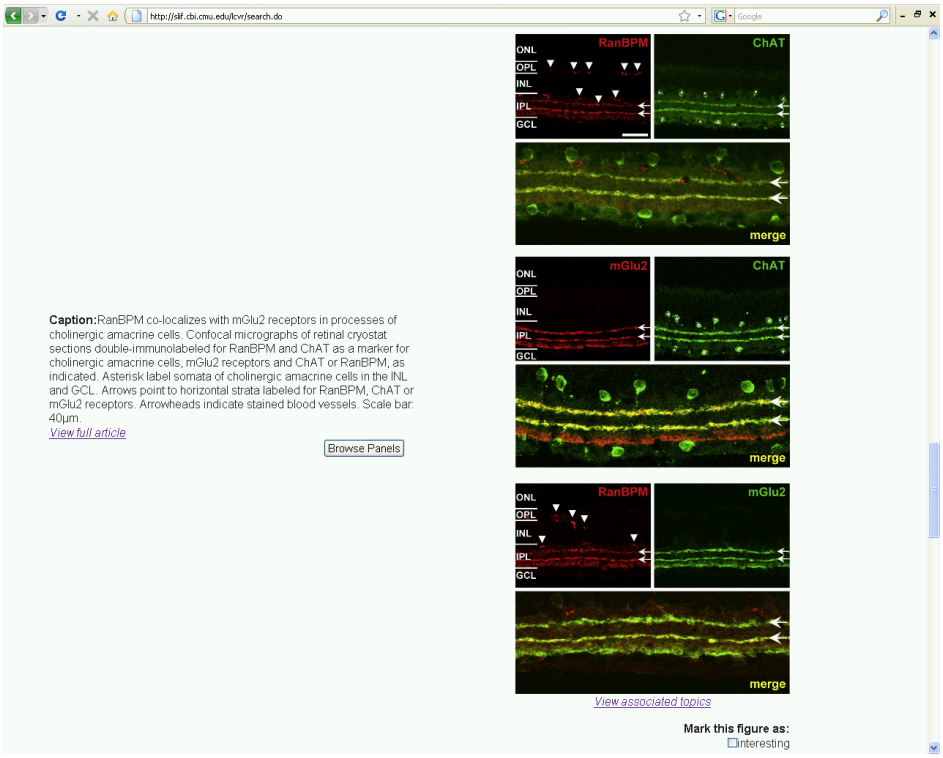


Figure 12: Another top similar biological figure to the query figure shown in Figure 11.

Indeed the above questions are addressed by biologists on a daily basis. For instance, when a biologist become interested in a given protein, she starts to search for papers that summarize research findings about this protein using various experimental technology like FMI, Gel; to name a few. Most of the users in our study used either Google scholar, regular Google (to search for images), or PubMed.

7.1 Results of User Study

Six out of *eight* users considered SLIF useful and a seventh stated that the system had "great potential" (Appendix B.4). To some extent, this mimics the results of Hearst et al. [22] who performed a user study on the viability of using caption searching to find relevant papers in the bioscience literature and found that "7 out of 8 [users] said they would use a search system with this kind of feature'.

Only one user found that the alternative search engine returned better results. Half found SLIF better and more relevant, and the other three thought the results were not directly comparable.

Moreover, *six* out of the *eight* users said that using topic-models in organizing the information was very useful or at least interesting (a sample comment states that it was "useful in terms of depicting 'intuitive' relationships between various queries"). On the other hand, *two* of them said that it does not help or that they did not understand how to interpret the displayed topic. In fact, topic-models provide a clustered view of the corpus and summarize these clusters or themes using

salient entities (protein, key words, panels) inside these clusters. Thus if the user does not recognize any of these salient entities, they can not evaluate the usefulness of this approach. To remedy this, we enhanced SLIF by adding relevance feedback, or the ability to search the collection based on what the user finds interesting to refine the search result until a satisfactory result is reached (see Section 6.2). This abstracts the role of the discovered topics. In other words, the user need not worry about the underlying technology used to carry out this search. However, since the majority of the users found topic-decomposition of panels and figures to be useful, we kept it also in the final release of SLIF. Our philosophy in doing so was as follows: SLIF should enable users to query and browse the paper collection regardless of their technological background, but at the same time it should endow advanced-users with extra capabilities that match their expertise.

Finally, we asked the users to suggest what features they would like to see in SLIF. The users requested that the "paper view" displays more information about the papers. Therefore, we enhanced it by including a list of the proteins and cell types discussed in the paper. We plan to also include the abstract, but this has not yet been implemented as it requires more changes to the pipeline.

Negative remarks centered either on the presence of some user-interface bugs (which we later fixed) or on the fact that a normal search engine returns more results than does SLIF, which is operating with a smaller collection of papers (when compared to Google, for example).

8 Discussion

SLIF demonstrates how text-mining and image processing can inter-mingle to extract information from scientific figures. Figures are broken down into their constituent panels, which are handled separately. Panels are classified into different types, with the current focus on FMI and gel images, but this could be extended to other types. FMIs are further processed by classifying them into their depicted subcellular location pattern. The results of this pipeline are made available through a either a web-interface or programmatically using SOAP technology.

A new addition to our system is latent topic discovery which is performed using both text and image. This enables users to browse through a collection of papers by looking for related topics. This includes the possibility of interactively marking certain images as relevant to one's particular interests, which the system uses to update its estimate of the users' interests and present them with more targeted results.

Our most recent human-labeling efforts (of panel types and sub-cellular location) were performed using active learning to extract the most out of the human effort. We plan to replicate this approach in the future for any other labeling effort (e.g., adding a new collection of papers). Our current labeling efforts were necessary to collect a dataset that mimicked the characteristics of the task at hand (images from published literature) and improve on our previous use of datasets that did not show all the variations present in real published datasets.

Although it is crucial that individual components achieve good results (and we have shown good results in our sub-tasks), good component performance is not sufficient for a working system. SLIF is a production system which working scientists in biomedical related fields have described as "very useful."

8.1 Future Work

There are many possible extensions to the current system which would add value for users.

Extending the system to interpret information from more panel types in addition to FMI is a natural growth area.

In the case of FMI subcellular pattern classification, we currently treat each channel as a separate entity. This procedure is correct, but does not fully exploit the relationship between channels presented together. In particular, we could extract information about co-localization experiments automatically. Often the caption will make a mention of what different colors represent and we extract that, but when no information is directly present, we could disambiguate using the image itself. For example, if a caption mentions both a protein and a well-known DNA stain such as DAPI and one of the channels shows a nuclear pattern and another a punctate pattern, the system should be able to reason that the nuclear pattern is associated with DAPI staining. As an additional advantage, we could use nuclear channels (when present) to aid in pattern classification as they allow us to compute protein/DNA features. In a similar vein, we could explore the relationship between panels presented together in the same figure. A typical display format consists of two different channels side by side and a merged version. This could be automatically detected and exploited.

An alternative path to take is to make different uses of the information extracted by the system. Automated assertion extraction as performed by SLIF can be used to automatically detect disagreements between different published results and provide new hypotheses for research. If the same protein is found to display differing patterns in different publications, this might lead to new interesting ideas.

Finally, we could explore other modes of interaction with the database in addition to the current searching modality. For example, we could present "related images" as a direct link inside the electronic versions of papers (the PDF format certainly provides such capabilities). Our database could also be used to enhance the *preview* capabilities of current collection browsers, which generally allow viewing of only the abstract and citations. SLIF could be used to summarize the information in the figures of the paper or allow direct linking to papers discussing similar topics. In addition to SLIF providing important new capabilities to users logged in to proprietary collections, the structured information that SLIF can make available could provide a new mechanism to drive users to individual papers in such collections.

To facilitate further research in the area of automated analysis of images and text in biomedical literature, we have made the hand-labeled datasets used to train SLIF components publicly available (http://slif.cbi.cmu.edu/training_data/). A list of these datasets is provided in Appendix A.

References

[1] R. F. Murphy, M. Velliste, J. Yao, G. Porreca, Searching online journals for fluorescence microscope images depicting protein subcellular location patterns, in: BIBE '01: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, IEEE Computer Society, Washington, DC, USA, 2001, pp. 119–128.

[2] W. W. Cohen, R. Wang, R. F. Murphy, Understanding captions in biomedical publications, in: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2003, pp. 499–504. doi:http://doi.acm.org/10.1145/956750.956809.

[3] J. Hua, O. N. Ayasli, W. W. Cohen, R. F. Murphy, Identifying fluorescence microscope images in online journal articles using both image and text features, Proceedings of the 2007 IEEE International Symposium on Biomedical Imaging (2007) 1224–1227.

[4] Z. Kou, W. W. Cohen, R. F. Murphy, Extracting information from text and images for location proteomics, in: Proceedings of the Third ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD), 2003, pp. 2–9.

[5] Z. Kou, W. W. Cohen, R. F. Murphy, A stacked graphical model for associating sub-images with sub-captions, Pacific Symposium on Biocomputing 12 (2007) 257–268.

[6] R. F. Murphy, Z. Kou, J. Hua, M. Joffe, W. W. Cohen, Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder, in: Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering, 2004, pp. 109–114.

[7] Y. Qian, R. F. Murphy, Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models, Bioinformatics 24 (2008) 569–576.

[8] Z. Kou, W. W. Cohen, R. F. Murphy, High-recall protein entity recognition using a dictionary, Bioinformatics 21 (2005) i266–i273.

[9] E. Pafilis, S. I. O'Donoghue, L. Jensen, H. Horn, M. Kuhn, N. Brown, R. Schneider, Reflect: Augmented browsing for the life scientist, Nature Biotechnology 27 (2009) 508–510.

[10] J.-M. Geusebroek, M. A. Hoang, J. van Gernert, M. Worring, Genre-based search through biomedical images, in: 16th International Conference on Pattern Recognition (ICPR), Vol. 1, 2002, pp. 271–274 vol.1. doi:10.1109/ICPR.2002.1044683.

[11] B. Rafkind, M. Lee, S. Chang, H. Yu, 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, Vol. 6, 2006, pp. 73–80.

[12] H. Shatkay, N. Chen, D. Blostein, Integrating image data into biomedical text categorization, Bioinformatics 22 (14) (2006) e446–e453.

[13] N. Roy, A. Mccallum, Toward optimal active learning through sampling estimation of error reduction, in: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, 2001, pp. 441–448.

[14] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[15] R. M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (1979) 786–804.

[16] 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. arXiv:http://bioinformatics.oxfordjournals.org/cgi/reprint/17/12/1213.pdf, doi:10.1093/bioinformatics/17.12.1213.

[17] K. Huang, R. F. Murphy, Automated classification of subcellular patterns in multicell images without segmentation into single cells, in: ISBI, IEEE, 2004, pp. 1139–1142.

[18] N. Hamilton, R. Pantelic, K. Hanson, R. Teasdale, Fast automated cell phenotype image classification, BMC Bioinformatics 8 (1) (2007) 110. doi:10.1186/1471-2105-8-110.
URL http://www.biomedcentral.com/1471-2105/8/110

[19] T. Ridler, S. Calvard, Picture thresholding using an iterative selection method, IEEE Trans. Systems, Man and Cybernetics 8 (8) (1978) 629–632.

[20] R. Jennrich, Stepwise Regression & Stepwise Discriminant Analysis, John Wiley & Sons, Inc, New York, 1977, Ch. 2 and 3, pp. 58–95.

[21] A. Ahmed, E. P. Xing, W. W. Cohen, R. F. Murphy, Structured correspondence topic models for mining captioned figures in biological literature., in: KDD '09: Proceedings of the fifteenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 39–48.

[22] M. A. Hearst, A. Divoli, J. Ye, Exploring the efficacy of caption search for bioscience journal search interfaces, in: ACL 2007 Workshop on BioNLP, 2007, pp. 73–80.

A Training Data for Text Annotators & Image Classifiers Used in SLIF

All datasets are publicly available at http://slif.cbi.cmu.edu/training_data

A.1 Text Annotation

A.1.1 Protein Entities

Archive (http://slif.cbi.cmu.edu/training_data/text/protein_annotation.tgz) comprises a set of abstracts as plain text files which are annotated for protein entities using <prot> < /prot> tags.

The abstracts were aggregated from three different datasets. In the dataset:

- 738 files with names 'abstract*-*.txt' came from the University of Texas, Austin dataset (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/proteins.tar.gz).
- 559 files with names 'geniadatafile_*.txt' came from the GENIA dataset (http://www-tsujii.is.s.u-tokyo.ac.jp/ genia/topics/Corpus/posintro.html).
- 200 files with names 'yapexdatafile_*.txt' came from the YAPEX dataset (http://www.sics.se/humle/projects

For more information, view Kou et. al. [8].

A.1.2 Cell Types

Archive (http://slif.cbi.cmu.edu/training_data/text/cell_annotation.tgz) contains captions as plain text files which are marked for positions of cell type occurrences in the text. These positions are stored in a separate file 'cell.labels'. The annotated cell types can be extracted by executing an included java program 'PrintCellLabels'.

The captions files 'p<PMID> $-$fig_ $*$ _$*$' were aggregated from PNAS articles, where PMID is a unique PNAS assigned id of the corresponding article. The fig_$*$ part in the file names corresponds to the figure number, to which the caption belongs. The articles are accessable online. For example, to access the article corresponding to the file 'p$9405699-$fig_4_1 ' in the dataset, goto: http://www.pnas.org/cgi/pmidlookup?view=full&pmid=9405699

A.2 Image Analysis

A.2.1 Recognition of FMI

Archive (http://slif.cbi.cmu.edu/training_data/image/fmi/fmi_all.tgz) contains two datasets of image panels segmented from figures in PNAS articles and hand-labeled as FMI or non-FMI. The archive also contains datasets' meta-information in XML format (http://slif.cbi.cmu.edu/training_data/image/fmi/fmi_datasetA_meta_info.xml, http://slif.cbi.cmu.edu/training_data/image/fmi/fmi_datasetB_meta_info.xml).

For more information, view Qian et. al. [7].

A.2.2 Recognition of Other Panel Types

Archive (http://slif.cbi.cmu.edu/training_data/image/panel_types.tgz) contains image panels segmented from figures in Pubmed Central articles and hand-labeled as 'FMI', '*Gel*', '*Light Microscopy*', '*Photograph*', '*Scanning Electron Micro-graph*' or '*X-Ray*'. The archive also contains data meta-information in XML format (http://slif.cbi.cmu.edu/training_data/image/panel_types_meta_info.xml).

A.2.3 Subcellular Locations

Archive (http://slif.cbi.cmu.edu/training_data/image/subcellular_classification.tgz) contains image panels segmented from figures in Pubmed Central articles, recognized as FMI and hand-labeled for subcellular location. The labels are '*Cytoplasmic*', '*Nucleary*', '*Punctate*' and '*Other*'. Data meta-information is also included in the archive as an XML file (http://slif.cbi.cmu.edu/training_data/image/subcellular_classification_meta_info.xml).

B SLIF User Study

B.1 Invitation

Help improve access to scientific literature and you could win $100 in 30 minutes!!

You can participate if:

- **You access biological or bio-medical articles for your research.**

What you will do:

- **We will give you access to our SLIF system that goes beyond simple keyword search capabilities.**
- **You will be asked to use our system to perform simple tasks.**

Why you should participate:

- **It is fun and there will be coffee and cookies**
- **There will be a $100 prize for a lucky winner!**

When:

- **It will take about 30 minutes, but you can come at any time between 4.30 and 6.30pm on Wednesday or Thursday (March 25&26).**
- **For questions contact Luis at lpc@cmu.edu or 412.330.8306**

Where:

- **CBI conference room (C119 Hammerschlag Hall, which is at the parking lot level near CMU's scaiffe hall, see http://www.cbi.cmu.edu/ for a map [click on the directions link]).**

B.2 Introduction & Study Tasks

Structured Literature Image Finder (SLIF)

Introduction

The structured literature image finder is a different type of search engine, one focused on images and their captions. *Why images and captions?* Because primary data in bio-medical fields is images. SLIF also allows smarter browsing. You can hop from one paper to a similar one based on its content, or browse by topic.

Notes

We are not very concerned with performance at this point, so please don't consider it in your evaluation. Also, some small bugs might pop up here and there. If you are using your own laptop, we ask that you use firefox as a web-browser.

The release of SLIF you are testing is only indexing a sub-set of Elsevier journals. Therefore, it is possible that your favorite paper isn't present in our collection.

If you get stuck, try to figure it out by yourself for one or two minutes, but do ask for help if it takes you longer than that to make progress.

Tasks

(0) Go to http://slif.cbi.cmu.edu/lcvr to start the study.

(1) You have become interested in the protein *Acyl-CoA cholesteryl acyl transferase-1 (a.ka. ACAT-1)*. In particular, you think it would be nice to have some fluorescence image data for it to see where it locates. Before spending time in the lab, you perform a literature search. Find some papers which have imaged *ACAT-1* using fluorescence.

(2) Try the same task on google scholar (http://scholar.google.com) or regular google (or any search engine you'd regularly use).

(3) Pick any protein you like and search for papers with gels involving that protein. How many papers are there in the collection? Do these results make sense?

(4) Can you try a similar thing with a regular search engine?

(5) Click on the *browse by latent topic* link on the top of the page. Take a moment to look at the topics, in particular the third topic. Browse through the *third topic*.

(6) Go back to the *browse by topic* page and look through any topic that mentions proteins or keywords that are familiar to you. Browse through it and see if it makes some intuitive sense. Look at the images and the text.

(7) Look for high-resolution flourescence microscopy images of *Actin* (you wanted something nice for a poster).

(8) Pick any of the high-resolution flourescence images you found and see its associated topics. Browse through these.

(9) Search for papers that have images related to "nuclear translocation." See both the papers and the images that are returned. Now search for papers that analysed translocation using fluorescence microscopy.

(10) Search for "nuclear translocation" using a regular search engine. This is as above.

(11) Take a couple of minutes to browse the site freely. Perform any search or browsing that you want.

B.3 Questionnaire

Please elaborate your answers.

Major/Program:_____

How often would you say you search for papers?

[ ] daily [ ] weekly [ ] every once in a while [ ] seldomly

Overall, how useful did you find SLIF? _____

_____

_____

_____

_____

_____

Was there anything that you particularly liked?_____

_____

_____

_____

_____

Was there any thing you particularly disliked?

_____

_____

_____

_____

Task 1&2

Which alternative search engine did you use?_____

Did you think that SLIF returned better results, worse results, or not comparable?_____

_____

_____

_____

_____

Under which system was it easier to filter/parse the results?_____

Would you consider SLIF for a similar search in the future? Why? _____

_____

_____

_____

_____

_____

_____

Task 3 & 4

Which protein did you pick?_____

Did you easily find a list of the papers using SLIF?_____

_____

Did you think those papers were relevant?_____

_____

_____

_____

Tasks 5 & 6

Did the *browse by topic* feature make sense to you?

[ ] very much [ ] somewhat [ ] not really [ ] no

Do you think it's useful?

[ ] very much [ ] somewhat [ ] not really [ ] no

Do you have any suggestions for improvement of the presentation of this feature?

_____

_____

_____

_____

Do you have any suggestions for improvement of this feature in general?

_____

_____

_____

_____

_____

Tasks 8

Did the *associated topics* feature make sense to you?

[ ] very much [ ] somewhat [ ] not really [ ] no

Do you think it's useful?

[ ] very much [ ] somewhat [ ] not really [ ] no

Do you have any suggestions for improvement of the presentation of this feature?

_____

_____

_____

_____

Do you have any suggestions for improvement of this feature in general?

_____

_____

_____

_____

_____

_____

Tasks 9 & 10

Did the looking only at captions return meaningful results?

[ ] very much [ ] somewhat [ ] not really [ ] no

Do you think it's useful?

[ ] very much [ ] somewhat [ ] not really [ ] no

Which alternative search engine did you use? _____

How do you compare these with the results returned by the search engine?_____

_____

_____

_____

_____

_____

_____

General Comments

Do you have any suggestions for improvements to SLIF? This can be either based on the user-interface or functionality.

Please write down any other comments you might have

B.4 Responses

User	Major/Program	Paper Search
1	Computational Biology	daily
2	Computational Biology	weekly
3	Computational Biology	weekly
4	Biomedical Engineering / Biochemistry	daily
5	Biomedical/Electrical Engineering	once in a while
6	Information Security	once in a while
7	Information Security	daily
8	Mechanical Engineering	once in a while

Overall, how useful did you find SLIF?

User 1 I found SLIF to be quite useful. Unlike Scholar or PubMed which give mostly "literature" based on some acronyms, SLIF was fairly specific. Query was quite specific which was helpful.

User 2 At this stage, not very useful. However, it has great potential!

User 3 I like it. Useful. It's a different paradigm for searching lit – could prevent a lot of repeat expts. If a protein is not the primary focus of a study it may get left out of a paper's title, but the paper may have data useful to studying that protein. etc.

User 4 No response.

User 5 It was a little hard to search for things. I think it would be better arranged if when you need to add more fields for search, instead of adding them have them all displayed. Since there are only 5 options it won't look wierd.

User 6 It was fine, but some features are not working. But overall it's OK.

User 7 It appears to be very helpful. The present image search in Google does not present expected results for a particular technical topic. In such cases, SLIF would really help.

User 8 Very helpful for finding images and performing searches with a set of criterion, even if they are unrealated.

Was there anything that you particularly liked?

User 1 I like the way the results were presented. The stream-lined interface is useful. The interface is simple and easy to navigate.

User 2 Be able to see figures and captions

User 3 Ability to tell SLIF:
- interesting / not interesting
- subcellular localization info.
- cell line info
all easy to find.

User 4 I like that you could search by assay (panel) & that you could search by location in a cell.

User 5 Having the "Show Caption/Figure" option. It's hard to completely tell if it is related unless there is a small snippet from it.

User 6 The association of information, like by topic, top words, top proteins etc.

User 7 The layout of the web-page is quite user-friendly. The number of choices provided in the search is also very nice.

User 8 I preferred the table views to the image views mostly because the results were grouped by paper making them redundant. I get how you can get more details in the caption, then hide it and continue with your results without navigating away from the page.

Was there anything that you particularly disliked?

User 1 Nothing in particular. It would be better to see at least part of the caption when the figure is showing because in some figures, the axes are not marked.

User 2 Many bugs. Difficult to understand the query form. For example how to specify image resolution or what are the allowed panel types.

User 3 Would be good to see paper ref. even when searching by figures. Also, I'd like to modify my search terms without restarting. At the top of results my terms are listed. It would be cool to choose a subset of them and search again.

User 4 Some parts were too slow to load (like "Browse" in topics). The graphs didn't include axes labels of stats or the associated figure text... so they were meaningless.

User 5 It was slow but that might be because of the internet.

User 6 Mostly GUI issues, also taking a lot of time.

User 7 The different choices should be rearranged a bit.

User 8 The associated topics seemed to be less "associated" than I would expect, and I was unclear as to the term 'Panel Confidence'. It would also be useful to have a 'nothing found' page as some "associated topics" link lead back to the home page. Oh, and a "return to search results" link, I was having trouble with the back button in Firefox.

Task 1 & 2

Which alternative search engine did you use?

User 1 PubMed
User 2 Google Scholar
User 3 Google Scholar
User 4 Google Scholar& image, & PubMed
User 5 Google Scholar
User 6 Google
User 7 Google Scholar & image search
User 8 Google Scholar

Did you think that SLIF returned better results, worse results, or not comparable?

User 1 I think in some fairness, the results are not comparable. SLIF images provided better matches, but PubMed search was a broader one.
User 2 Worse results. I could not find meaningful results. Wait! I found some images but it was after 20 mins!
User 3 Google Scholar returned results whose main focus was fluorescence studies of ACAT–1, whereas SLIF returned studies that employed fluoresscence of ACAT–1. If I'm studying ACAT–1, SLIF is probably better.
User 4 It didn't have as many results, some were duplicates (which is obnoxious), and many are not compatible
User 5 Not comparable. Both have problems. Google is too much text while SLIF is not enough. Each results from SLIF looks too generic from the next. I have to click on the caption to get more info.
User 6 Better and relevant results.
User 7 It returned more relevant results as compared to Google Scholar.
User 8 Better. The images were immediately available. Show Caption/Figure didn't always work but I like that pulled down option.

Under which system was it easier to filter/parse the results?

User 1 PubMed. It showed authors, abstracts too.
User 2 Google Scholar, but I liked to be able to see the figures and captions at SLIF.
User 3 Google
User 4 Google
User 5 Google
User 6 SLIF
User 7 SLIF
User 8 Couldn't narrow SLIF by fluorescence after entering 'ACAT–1', so I think Google Scholar.

Would you consider SLIF for a similar search in future? Why?

User 1 I would consider SLIF in future if it were purely based on images.

User 2 At this stage, No! because it has so many bugs. I tried to query protein: ACAT–1 and text in caption: fluorescence but it did not work.

User 3 Sure. I feel like it's a more thorough search of the lit.

User 4 Probably not – it just didn't include enough results – if you filtered results, you are left with no hits.

User 5 Yes especially if I am looking for images. If I already know what I'm looking for specifically then SLIF would work best.

User 6 Yes because of relevancy of information and also good association of information.

User 7 Yes, surely.

User 8 if pictures/figures were important – yes, but the initial text quote by Google Scholar helped to narrow things down about the actual topics rather than just the picture, protein and cell info.

Task 3 & 4

User	Which protein did you pick?	Did you easily find papers using SLIF?	Were papers relevant?
1	dihydrofolate reductase	Yes	Somewhat of the first 10 results, I found the top two relating more to FTPase than DHFR. But, the results were relevant.
2	mdm2	7 papers	Yes, but I could not find gel image in a paper about mdm2 The browse figures links were broken.
3	Ubiquitin	Yes	I couldn't browse figures! Paper no longer in DB. View paper link went to wrong paper!
4	IL–18	Sort of; I only found 93 whereas Google found $\sim$ 30 million after filtering.	I couldn't tell – none of the text about image, including axes labels for graphs, was included.
5	Actin	Yes	Somewhat. Once again need to find a balance too little/ too much text.
6	TLR 2	Yes	Yes
7	Actin	Yes	Yes, quite relevant
8	GAPDH	Just took that protein from ACAT–1 results, found $>$ 14,000 papers.	Most results on the first and 1/2 second page were DJ–1, then the GAPDH started to come up.

Task 5 & 6

User	Did the browse by topic feature make sense?	Do you think it's useful?
1	somewhat	very much
2	no	not really
3	somewhat	very much
4	no	no
5	very much	somewhat
6	somewhat	very much
7	somewhat	somewhat
8	very much	somewhat

User	Any suggestions for improvement of the presentation of this feature?	Any suggestions for improvement of the this feature in general?
1	While quite useful in terms of depicting "intuitive" relationships b/w various queries, the latent topics must also be presented say along with the query to show "context".	I think it would be better to have the queries relate back to the user (such that the main search itself shows context) and one can easily refine them. May be a graphical representation (see vivisimo's interface) also helpful to visualize.
2	It looked random. The user cannot choose/input his/her topic.	No response
3	No response	I need to think on this. There's something you can do to make more obvious what the utility, but I'm not sure what titles for each set of related topics?
4	Is there any reason behind how these are clustered? Everything is just in random sets.	If you could select images of cells that are all the same type (all hMSC cells) and then cluster them by the morphology (shape) of the cell... and then have links to the papers that all have images of the cells in that particular shape (like – a list of all papers that have hMSC's in spidle shapes), that would be pretty sexy.
5	Once you set into associated topics the table is all over the place. It would be better if each topic in the table was same width/height – constant image sizes.	Make "Browse" link more apparent, it's kind of hidden.
6	Please make it more user-friendly and input got erased each time after a search	You can show references or from where the paper is retrieved.
7	The option can be located below the search text-box because a normal user feels that the links above it are not a part of search option	No response
8	If the "words" or "proteins" are common in the database, make them links? The "Searched for" does not reflect the chosen image from the previous page, and I don't know enough to see if it's the correct images.	It was very slow after clicking browse.

Task 8

User	Did the associated topics feature make sense?	Do you think it's useful?
1	very much	very much
2	no	no
3	somewhat	somewhat
4	not really	not really
5	very much	somewhat
6	very much	very much
7	very much	very much
8	somewhat	very much

User	Any suggestions for improvement of the presentation of this feature?	Any suggestions for improvement of the this feature in general?
1	It would be easier if the images were somewhat "panelled"/grouped. That way a user can decide what group he wants to browse.	It would be cool to have some feature like "Cool Iris" where images are moving along on a stream. The user can control what to look at & what not.
2	I could not find it!	No response
3	No response	No response
4	Add information for all types & environments (plates, gels, scaffold) & 'Western blot analysis' has nothing to do with cell type.	Maybe switch to having thumbnails of images with protein labels above & have 'sort by' protein function working.
5	Same opinion as before about table.	Same as before.
6	It is fine.	No response
7	No particular suggestion.	No particular suggestion
8	What does "Panel Confidence" mean?	Words and Proteins had very little in common among the "associated topics", only 1 were common if any between the 3 results returned. Therefore, I don't know how realistically useful it is but the concept is very useful.

Task 9 & 10

User	Did the looking only at captions return meaningful results?	Do you think it's useful?
1	very much	very much
2	somewhat	somewhat
3	No response	No response
4	not really	not really
5	very much	somewhat
6	very much	somewhat
7	very much	somewhat
8	somewhat	somewhat

User	Which alternative search engine did you use?	How do you compare with the results returned by the search engine?
1	Google Scholar	I think SLIF's image and content based search is far superior to that of Scholar. While Scholar returns relevant papers, it has to go through multiple steps to refine back to the similar results.
2	Google Scholar	Google Scholar returns the paper and a fragment of text containing the keyword. SLIF returns the image and caption. However, none of them free the user from the work of reading the paper to confirm the desired information.
3	Google Scholar	Google searches on topic. SLIF searches on content – you get more robust examples e.g. of nucleus translocation, I think. Though fewer.
4	Google	It only returned one paper?
5	Google Scholar	They were about the same.
6	Google	It is somewhat comparable.
7	Google image search	The results for the SLIF seemed more relevant.
8	Google Scholar	The 'paper' setup on SLIF was much better then the setup for Google $\rightarrow$it was much easier to compare, look for dates or authors. Also, you get and idea of the # of figures you can access. I didn't see a 'look only at captions' option but I did find that the caption/images returned the same result for papers a few times each with a different figure. This was a bit redundant and hard to look through, but I suppose if you can identify the picture without that it would be useful. Overall, I liked the 'paper' view better.

General Comments

User	Any suggestions for improvements to SLIF?	Any other comments?
1	a) I find the UI intuitive & friendly/simple. b) I would like the UI to be somewhat streamlined – where as the searches are "columnless" the Latent topics part is "tabular" which is a bit annoying. c) I feel people have a tendency to click on images more than links on the left/bottom.	Overall, It's a great effort.
2	There were many bugs. I think that the amount of bugs prevented me from focusing my evaluation on the quality/meaningfulness of the hits.	I liked to see the figure and the caption. It seems that the engine is biased towards biomedical images. I would suggest to expand this feature to other types of images & fix the bugs.
3	No response	No response
4	It needs a larger database and more complete info. & can load faster.	No response
5	I think the functionality is find but the layout needs work.	No response
6	Very slow, and please make it user-friendly	No response
7	There are many options in SLIF. An improvement would be presenting options in more user-friendly manner.	Very nice concept. It will help searching for research papers.
8	How does interesting/not interesting feature work? I like the idea of it, especially if it's session based while you're researching a specific topic, and can be cleared at another time. The table format of both "paper view" and GO term/Image locator were very useful for comparison. Also, a legend discussing what some of the key terms mean or where the links lead (either a more thorough description on SLIF or the protein links to external sites).	The "add search field" could also benefit from a 'not' option but I like its configuration and it makes boolean-type searches more streamlined. I don't know much about the type of the database, but it is very aesthetically appealing and pretty user-friendly. It also ran very slow at times, but I'm sure that's expected.