Robust Segmentation of Biomedical Figures for Image-based Document Retrieval

Luis D. Lopez, Jingyi Yu, Catalina O. Tudor, Cecilia N. Arighi, Hongzhan Huang, K. Vijay-Shanker, Cathy H. Wu
Department of Computer and Information Sciences
University of Delaware. Newark, USA
ldlopez@udel.edu, yu@eecis.udel.edu, oanat@udel.edu, arighi@dbi.udel.edu,
huang@dbi.udel.edu, vijay@udel.edu, wuc@dbi.udel.edu

***Abstract*—Figures play an important role in illustrating concepts, methodology and results in biomedical literature. However, figures in biomedical literature are often composed of multiple subfigures (panels), which may illustrate diverse methodologies or results. Robust and accurate panel partitioning is crucial to support article categorization based on methods or experimental results and to provide the evidence source for derived assertions. But, it is a challenging task. In this paper, we present a comprehensive framework for harvesting multimodal panels in biomedical literature, and demonstrate its application to protein-protein interaction (PPI)-related literature as a use case. A unique feature of our solution is that we combine pixel-level representations of images with figure captions. Our approach first analyzes figure captions to identify the label style used to mark panels. We then use pixel-level representations to partition a figure into a set of bounding boxes of connected components. We also perform a lexical analysis on the text within the figure to locate panel labels that match the caption analysis results. Finally, we estimate the optimal panel layout and use the layout to partition the figure. We tested our system on a dataset provided by the Molecular INTeraction database (MINT), and show that our approach surpasses pure caption-based and pure image-based approaches, achieving a 96.64% precision.**

***Keywords*-Literature mining; biomedical image analysis; image processing; database curation; protein-protein interaction; figure panels; image segmentation**

I. Introduction

Figures play an important role in illustrating concepts, methodology and results in biomedical literature. Recent studies have shown that combining information from figures and captions can help to improve classification and retrieval tasks. For example, Murphy *et al.* [1] developed a method to retrieve biomedical documents based on their graphical content. Similarly, Shatkay *et al.* [2] proposed an approach to integrate information from text and figures to automatically classify biomedical documents, while Demner-Fushman *et al.* [3] developed a comprehensive system for identifying images relevant to medical diagnosis.

At the core of figure-assisted document classification algorithms is the task of robustly extracting figures from documents. However, figures in biomedical literature are often composed of multiple subfigures (panels) which cannot be identified by parsing the document. Since each panel may correspond to a completely different methodology, panel partitioning is a crucial task for content-based document retrieval and triage. The problem, however, is particularly challenging for biomedical articles since different types of panels can still share similar image characteristics, e.g., bar charts vs. line charts, multiple gel images, and DNA sequences vs. protein sequences image. As a result, classical image segmentation techniques based on low-level image features such as edges or color can be fragile.

In this work we focus on the literature about protein-protein interactions (PPIs) as a use case, since detection of the methodology is a critical aspect contributing to the confidence of PPI assertions [4]. In this case we use the literature corpus and annotated data provided by the Molecular INTeraction database (MINT) [5], a manually annotated PPI database, which consists of $2,848$ articles with experimental evidence for PPIs. On average, each article in this data set contains 6 figures and each figure contains 3 panels. Figure 5A) shows an example multimodal figure relevant for PPI curation extracted from PMID:10871282.

In this paper, we present a comprehensive framework for harvesting unimodal panels in biomedical literature related to protein-protein interactions (PPI). The core of our approach is to integrate in-image text with pixel-level representation of images and figure captions to robustly estimate the panels in the image. Specifically, we present a technique that can simultaneously identify the numbering style and layout of figures. Our approach first analyzes figure captions to identify the label style (e.g., a, b, c..., 1, 2, 3, etc) used to mark panels. We then apply an image processing module that uses pixel-level representations to partition a figure into a set of bounding boxes of connected components. We also perform a lexical analysis on the text within the figure to locate the panel labels consistent with the caption analysis results. Finally, we estimate the panel layout and use the layout to optimally partition the figure. On the MINT dataset, we show that our approach greatly surpasses pure caption-based and pure image-based approaches and achieves a 96.64% precision. To allow for efficient retrieval of information as well as tasks such as classification, we further developed a web application which allows users to retrieve the partitioned panels, the text embedded within each panel, and their corresponding subcaption.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

II. Methodology

Figure 1 shows our processing pipeline, which consists of five major components. The first component is a robust technique for the automatic **extraction of figure-caption pairs** from biomedical articles in PDF format [6]. Given these figure-caption pairs, we look at the caption and image separately in an effort to identify the exact number of panels. In particular, we first developed a **Caption Segmentation** module that simultaneously identifies the number of panels in the figure and separates the caption into its corresponding subcaptions. Next, the **Image Preprocessing** module uses pixel-level information to partition each figure into a set of bounding boxes of connected components. The **In-Image Text Processing** module then performs a lexical analysis on the text within the figures to identify the corresponding panel labels. The **Panel Segmentation** module utilizes the results of the previous modules to estimated the correct number of panels in the figure, and then partition the figure using the image-level and label-level details.

A. Extraction of Figure-Caption Pairs

Similar to the SLIF system proposed by Cohen *et al.* [7], we start by automatically extracting the figure-caption pairs from each biomedical publication in our database. While [7] uses a modified version of the publicly available PDF2HTML tool, we have developed a more robust framework. Our system is able to effectively remove irrelevant figures such as journal logos and combine fragmented figures, which is largely missing in previous figure-caption extraction solutions.

Our approach relies on the idea that the document layout can be used to identify encoded figures and figure boundaries within the PDF, and enforce constraints among figure-regions. This allows us to harvest fragments of figures, from the PDF, correctly merge subfigures that belong to the same figure, and identify the captions associated with each figure. Our method simultaneously recovers figures and captions and applies an additional filtering process to remove irrelevant figures such as logos, and to eliminate text passages that were incorrectly identified as captions. More details of this method can be found in [8].

Once we finish extracting the figure-caption pairs from the PDF files, we use the ABBYY OCR software (www.abbyy.com) to recognize the text within each of the figures. We identify and store both the text found in the figure and its corresponding bounding box.

B. Caption Segmentation

The caption of a figure plays a key role in explaining the figure contents [9]. However, they are largely neglected in previous figure extraction and figure-based document classification approaches. Captions reveal important information, such as the number of panels in the figure and the type of labels used to delimit these panels. However, parsing the captions can be particularly difficult for biomedical figures: a figure can contain different types of subfigures, while its corresponding caption can be written using a complex structure [7], [10]. Further, captions from biomedical publications typically contain abbreviations, special characters, and ambiguous words that can generate a relatively high number of false positive labels [7].

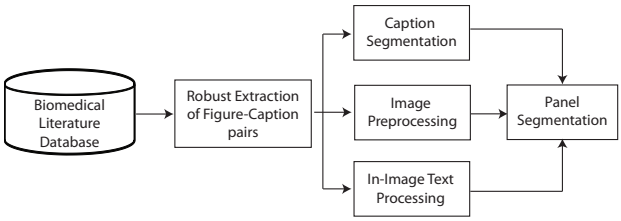

Figure 1. Overview of our system to automatically segment unimodal panels from multimodal biomedical figures.

We process the captions to identify the number of panels in the corresponding image, and a textual descriptor specific to each image panel in the figure. To do so, we developed a three-phase algorithm. In the first phase, we perform a simple lexical analysis to identify the potential panel labels in the captions. In particular, we use a set of hand coded rules similar to the ones proposed by Cohen *et al* [7] to identify: 1) simple parenthesized expressions of the form "(X)", "(x)"; 2) complex parenthesized expressions indicating range of labels "(X - Y)", "(x - y)", "(x to y)", "(X, Y, and Z)", "(x and y)", etc; and 3) open expressions of the form "X.", "x)", "X,", "x;". The difference between the two approaches is that we identify all possible combination of labels, and in a subsequent phase perform a simple but effective pruning technique to filter out false positive labels. The patterns recognize characters from the English and Roman alphabets, as well as numbers. Thus, from the sample caption in Figure 5 extracted from *PMID:10871282*, we expect to identify the labels "*(A)*", "*(B)*", '*(C)*", and '*(D)*" indicating that the corresponding image contains four panels.

In the second phase, we analyze the list of potential panel labels to identify and remove false positives. Our approach first creates separate clusters for sequences containing English letters, Roman numbers, and Arabic numbers. The labels that do not form a sequence are then removed from the clusters. If we encounter multiple sequences of the same type we simply keep the first occurrence and remove the rest. In the rare case when a single caption contains expressions from more than one cluster we select the group with the maximum cardinality and discard the rest. Finally, from the sequences that are left, we give priority to letters, then Roman numbers, then Arabic numbers.

Finally, in the third phase, we segment the captions according to the set of panels that were identified. We use the position of the tokens in the sequence to split the caption and generate a list of subcaptions. For all the cases with a range of labels (e.g., (A-D)) we simply repeat the subcaption as many times as the labels in the range (four in this example).

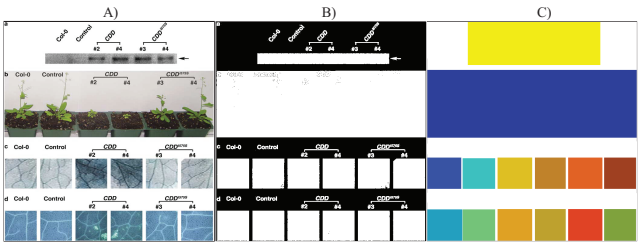

Figure 2. Example of our image processing algorithm. (A) Original image extracted from PMID:16003391 [18]; (B) Binarization result using the threshold value calculated by the triangle method [16]; (C) Bounding box of connected components after removing regions with text.

C. Image Preprocessing

We process the figures at the pixel level to separate the foreground area from the background area, as well as to generate a bounding box for each connected component in the figure. Automatically segmenting biomedical figures is challenging, since they can contain dense heterogeneous objects with distinct intensity scale parameters, and a single threshold value may not be able to recover the complete region of all objects in the figure [11]–[13]. Previous solutions attempted to resolve this problem by segmenting the image using heuristics to identify discontinuities in the intensity value [1], or by calculating the threshold that minimizes the intraclass variance [11], [14] using the Otsu filter [15]. Here, we utilized the triangle method proposed by Zack *et al.* [16] to compute the optimum threshold value for segmenting the figure. This method constructs a line between the first non-zero bin to the highest histogram value. For each intermediate point in the histogram we compute the distance normal to the line and between the line and the histogram. Finally, the intensity value in the histogram with the maximum distance defines the threshold value. This technique has previously shown to be particularly effective in noisy images and images with low contrast [17].

To improve robustness, we further remove the regions identified as text by the OCR software and then perform a connected components (CC) [19] analysis on the resulting image to generate a set of uniquely labeled regions using a 4-neighbor connectivity. We then compute a bounding box for each foreground region and merge connected regions by identifying overlapping bounding boxes. Figure 2 shows the regions (Fig. 2C) extracted from a sample image (Fig. 2A) after removing the text.

D. In-Image Text Processing

We also analyze the text embedded in the figures to detect the set of potential panel labels. Different from caption analysis module, directly applying an OCR tool for identifying panel label from in-image text can cause many additional problems. For example, most OCR tools generate noisy characters in regions with complex textures. Thus, the results could be incorrectly identified as panel labels or protein names. In addition, OCR tools typically fail to recognize characters in regions with low contrast, which can generate gaps in the sequence of panel labels [20]. To resolve these problems, our solution first performs a lexical analysis to recognize the words forming a potential panel label (e.g., “A”, “B.” “(C)”). Our method can identify three types of panel labels: regular labels consisting of one alphabetic character (e.g., “a” ,“A”), right-closed labels consisting of an alphabetic character followed by a right delimiter (e.g., “a)” ,“A.”), and closed labels consisting of a left delimiter following by a right-closed label (e.g., “(a)” , “[A]”).

Once we extract all the potential panel label sets, we cluster them using their type and letter case words (e.g. {“(A)”,“(B)”}, {“a.”,“b.”}). We then sort the labels using their position (e.g. position(’a’)=1, position(’C’)=3). In theory, our solution should generate only one set containing the panel labels. In reality, it may also mistakenly generate additional sets from incorrectly recognized words. Therefore, in our solution we compute a confidence value for each non-empty set i using the following equation ($confidence_i = \frac{1}{1+Gaps} * Panels$), where $Panels$ is the total number of panels identified in the figure, and $Gaps$ is the total number of gaps in the sequence; we then select the set with the highest confidence and discard the rest.

E. Panel Segmentation

Once we recover the set of subcaptions, panel labels, and connected components from the previous modules, we estimate the number of panels in the figure, and partition it into a set of panel-subcaption pairs. Recall that the set of subcaptions and the set of panel labels are recovered independently using the **Caption Segmentation** module (Section II-B) and the **In-Image Text Processing** module (II-D). Thus, the two reported numbers may be different if any or both approaches fail to identify the correct number of panels. In our approach, if the number of subcaptions and panel labels are consistent, we then simply use this value as the estimated number of panels. Otherwise, we take into consideration the number of connected components to determine which set contains the correct number of panels.

Notice that some panels in a figure may contain several connected components. Therefore, when neither of the three recovered sets are consistent (e.g., the recovered sets have different cardinalities), we use the decision tree algorithm C4.5 [21] to estimate the correct number of panels in the figure. To build our decision tree model, we randomly selected two hundred figures from our set of manually annotated figure-caption pairs. We used the number of subcaptions, panel labels and connected components in the image, as well as the number of gaps in the sequence of panel labels as features for training the model. The value generated by the decision tree model is then taken as the correct number of subfigures. Figure 3 shows the flowchart of our proposed algorithm to estimate the number of panels in a figure.

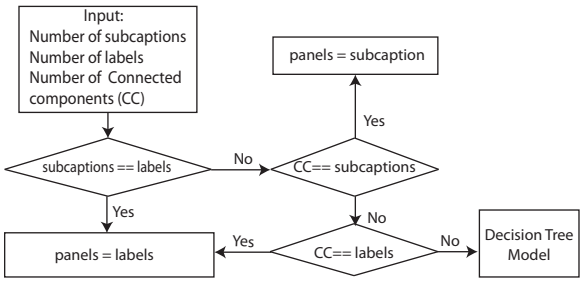

Figure 3. Algorithm flow for estimating the number of panels in a multimodal biomedical figure.

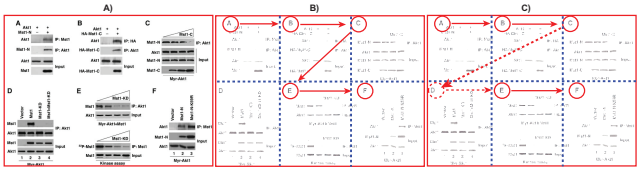

Figure 4. Estimating the position of unrecognized pointers. (A) Sample input image extracted from pmid:17932490 [22]; (B) Layout graph containing a gap in the first column; (C) Our final recovered layout graph including the estimated node "D".

To partition the figure, we first create a node for each recovered panel label and then create an edge connecting each node to its closest horizontal and vertical neighbors (Fig. 4B). Notice that OCR engines typically have poor performance when recognizing characters under low contrast conditions, and in regions with complex textures. For this reason, our sequence of panel labels can contain gaps. We fix the gaps by introducing missing labels in the sequence and estimating their position in the figure. To do so, we use symmetry to estimate the spatial position of the node. For a missing pointer between two nodes in the same column or row, we create a new node as the midpoint between the previous node and the next node in the sequence. For more complex cases, we start by grouping the nodes by columns and rows, and then use the structure of the nodes in parallel groups to estimate the position of the missing pointers. For example, to recover the position of node "D" from the layout graph shown in figure 4, we use the edge $\overline{AB}$ to estimate the horizontal distance to node "E" and the edge $\overline{CE}$ to compute the vertical distance to node "A". Once we finish constructing the *panel label graph*, we start partitioning the figure by computing a vertical cut for each horizontal edge in the *panel labels graph* (Fig. 4C). Similarly, each vertical edge generates a horizontal cut in the input figure. Figure 4D) shows our computed layout of a sample figure.

III. Results and Discussion

A. Data and Gold Standard

To evaluate the performance of our proposed solution, we have randomly selected $2,256$ full-text biomedical documents in PDF format from the annotated corpus provided by the MINT database. We chose to use this dataset because we are particularly interested in extracting figures that illustrate various types of PPI methods (e.g., yeast two-hybrid, chromatography technology, and fluorescence microscopy). Many biological databases are implementing evidence and methodology codes to provide a measure of confidence of the assertions made (e.g. experimental vs. predicted, and large-scale vs. small scale experiment). Therefore, it is relevant to identify not only the figure related to the assertion, but also the accompanying methodology, usually found at the panel level. The MINT database, however, does not directly provide the articles in PDF format. Therefore, we manually downloaded them from PubMed Central (PMC) and the publisher websites. Since we use a large number of documents, manually extracting panels is impractical. Therefore, we first used our system to automatically extract the panel-subcaption pairs, and then asked a group of annotators to check the figures and the captions, and mark through a web-interface the number of panels that they can distinguish in the image. We then manually fixed errors. This has significantly reduced the effort of generating the ground truth dataset. Our dataset consists of $13,147$ figure-caption pairs containing $41,341$ panels with associated subcaptions. Three curators helped in marking the figure-caption pairs and completed the task in 180 hours.

B. Identification of Panel Labels

For the subtask of detecting panel labels within the figure, our approach was able to correctly find the complete set of panel labels in $9,688$ figures (73.69%). Of these, 532 (5.49 percentage) figures contain label gaps that are successfully fixed using our algorithm described in Section II-D.

We further analyzed the remaining $3,459$ figures that had an incorrect number of panel labels to determine the different sources of errors. We found that $1,479$ figures contained more labels than the actual number of panels, and that $1,980$ figures contained an incomplete set of labels. The main causes for finding additional panel labels embedded in the figures are: the presence of noisy characters randomly generated by the OCR engine in zones with complex textures; and the presence of letters in the bin of the barcharts. For the set of figures with an incomplete number of panel labels, our approach selects the incorrect sequence of panel labels in in 597 figures. The remaining incorrect ones are caused by the failure the OCR engine (e.g., failing to recognize the labels). For example, some labels are in zones with low contrast which makes challenging to correctly recognize the text. In addition, the OCR engine make random changes to the labels by either adding an space between the words and the delimiters or changing the letters with low recognition confidence (e.g., changing character "D" to "I").

C. Caption Segmentation

Table I shows the performance of our system on the subtask of segmenting captions. In our evaluation, a caption is viewed as successfully segmented when our system

Description		Images	Rate
Successfully Segmented		11,984	91.15%
Cause of error	Incomplete Captions	673	5.12%
	Inconsistent Captions	292	2.22%
	Missing Identifiers	198	1.51%

Table I
RESULTS OF CAPTION SEGMENTATION ON THE MINT DOCUMENT SET

identifies the correct number of panels in the figure and the caption is correctly fragmented. Our caption parsing scheme is able to achieve an accuracy of 91.15%. We have further analyzed the set of $1,163$ failure cases to identify the main cause of error. We attribute 58% of the errors to the module that extracts figure-caption pairs (as described in Section II-A). This problem occurs since biomedical figures tend to be highly complex, and consequently, they require long captions to describe them. As a result, it is a common practice to split them on different pages or to use complex document layouts to optimize the space. This causes our solution to fail. In 25.10% of the failure cases we found figures with inconsistent captions. For example, captions containing multiple non-overlapping sequences of subcaptions, or captions with protein names and abbreviations using words that are commonly used to describe a subcaption (e.g., using (C) for carbon). Finally, 17.02% of failure cases were caused by captions describing a figure with multiple panels that does not contain a label to trigger our caption segmentation approach.

D. Extraction of Unimodal Panels

As mentioned in Section II-E, an important feature of our system is its ability to accurately estimate the number of panels in figures. In our experiments, our panel segmentation and in-image text processing algorithms have identified a similar number of labels and subcaptions in $9,485$ cases, above 72.14% of the figure-caption pairs in the dataset. Among the correctly estimated ones, $9,294$ figures 96% are consistent in caption-based and figure-based panel estimation. For the set of $3,662$ failure cases, our decision tree model successfully estimates the number of panels in 73.15% of the cases. Since our approach does not attempt to obtain the position of the missing panels in the figures. When the estimated number of panels is bigger than the set of labels recovered from the figure, we rely on the connected components to estimate the layout of the missing panels.

Finally, to evaluate the performance of our panel segmentation algorithm we use our system to obtain the set of panel-subcaption pairs and verify the results manually. Specifically, we deem that the result is correct if the panel contains the precise set of objects and text from the input figure and is linked with the correct subcaption. Out of $41,341$ panels, our system correctly produced $39,951$ or 96.64% in the dataset. It is important to note that even when our approach estimates an incorrect number of panels in the figure, it is still able to extract a subset of correct panels. This is because we compute the graph used to estimate the figure layout using only local information. Therefore, our solution can still be useful even if the panel number estimation is incorrect. Figure 5B) shows the set of panel-subcaption pairs extracted from two sample figures 5A) using our approach.

IV. CONCLUSION AND FUTURE WORK

We have presented a comprehensive and robust system for automatically harvesting panels (subfigures) from PDF biomedical articles. Our approach integrates information derived from figures, texts embedded within figures, and figure captions to first estimate the number of panels, then determine the panel layout, and finally partition each figure into panels. A unique feature of our system is a lexical analyzer coupled with a set of image-processing filters to identify the number of panels. We have validated our system on a large publicly available biomedical corpus provided by MINT and demonstrated our solution achieves high robustness and accuracy.

There are a number of areas we plan to explore in the future. For the MINT databaset, we plan to extract additional knowledge from panels and their captions such as protein entities, gene names, interaction between proteins, and the experimental method used to generate the figures. This would allow us to apply traditional data mining techniques and information retrieval techniques to facilitate content-based archiving and retrieval of relevant biomedical articles for protein-protein interaction. We also seek to further improve our solution. Our analysis shows that noisy characters generated by the OCR tool are a main cause of errors. In our current implementation, we assume that texts identified by the OCR tool contains the complete and correct sequence of panel labels in the figure, which can be violated for figures with complex textures and low-contrast labels. To resolve this issue, we plan to integrate more robust techniques in Natural Language Processing (NLP), such as statistical language modeling, to improve noisy OCR data.

ACKNOWLEDGMENT

We would like to thank Gianni Cesareni and Livia Perfetto from the MINT database for providing the annotated dataset at the figure level.

REFERENCES

[1] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca, "Searching online journals for fluorescence microscope images depicting protein subcellular location patterns," in *IEEE Int. Conf. on Bioinformatics and Bioengineering*, 2001, pp. 119–128.

[2] H. Shatkay, N. Chen, and D. Blostein, "Integrating image data into biomedical text categorization," in *ISMB (Supplement of Bioinformatics)*, 2006, pp. 446–453.

[3] H. Kilicoglu, D. Demner-Fushman, T. C. Rindflesch, N. L. Wilczynski, and R. B. Haynes, "Viewpoint paper: Towards automatic recognition of scientifically rigorous clinical research evidence," *JAMIA*, vol. 16, no. 1, pp. 25–31, 2009.

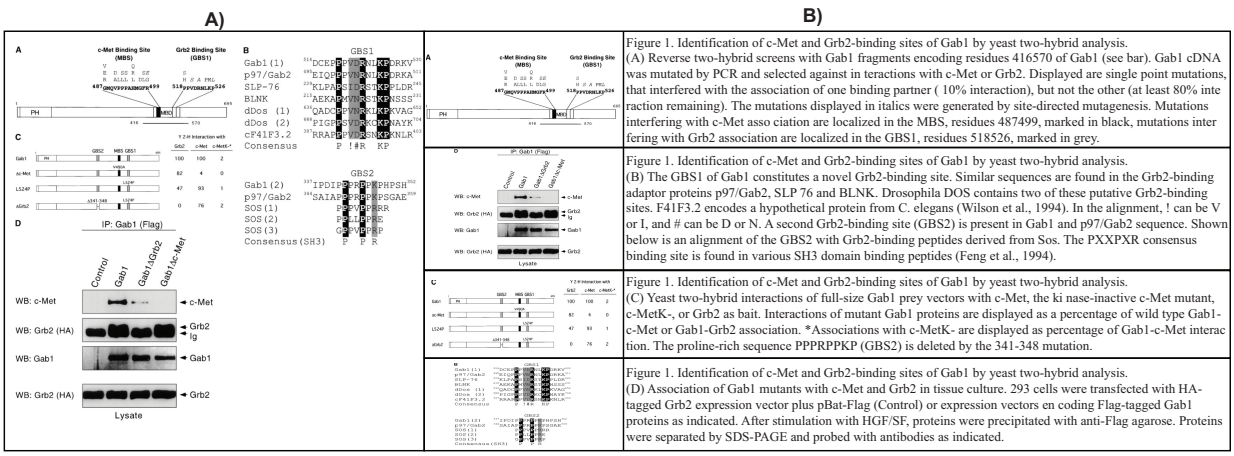

Figure 5. A gallery of panel-subcaption pairs extracted by our framework. (A) The input image; (B) The segmented image panels with their corresponding subcaptions.

[4] S. Orchard, S. Kerrien *et al.*, "Protein interaction data curation: the International Molecular Exchange (IMEx) consortium," *Nature Methods*, vol. 9, no. 4, pp. 345–350, Mar. 2012.

[5] A. Ceol, A. Chatr-aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, "Mint, the molecular interaction database: 2009 update," *Nucleic Acids Research*, vol. 38, no. Database-Issue, pp. 532–539, 2010.

[6] C. Adobe Systems Inc, *PDF Reference with Cdrom*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2000.

[7] W. W. Cohen, R. Wang, and R. F. Murphy, "Understanding captions in biomedical publications," in *Procs of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining 03*. ACM, 2003, pp. 499–504.

[8] L. D. Lopez, J. Yu, C. N. Arighi, H. Huang, H. Shatkay, and C. H. Wu, "An automatic system for extracting figures and captions in biomedical pdf documents," in *BIBM*, 2011, pp. 578–581.

[9] X. Chen, C. Lu, Y. An, and P. Achananuparp, "Probabilistic models for topic learning from images and captions in online biomedical literatures," in *Proc. of the 18th ACM CIKM '09*. ACM, 2009, pp. 495–504.

[10] R. F. Murphy, Z. Kou, J. Hua, M. Joffe, and W. W. Cohen, "Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder," in *Proceedings of KSCE-04*, 2004, pp. 109–114.

[11] S. Antani, D. Demner-Fushman, J. Li, B. V. Srinivasan, and G. R. Thoma, "Exploring use of images in clinical articles for decision support in evidence-based medicine," in *DRR*, 2008, p. 68150.

[12] L. P. P. Coelho, A. Ahmed, A. Arnold, J. Kangas, A.-S. S. Sheikh, E. P. Xing, W. W. Cohen, and R. F. Murphy, "Structured Literature Image Finder: Extracting Information from Text and Images in Biomedical Literature." *Lecture notes in computer science*, vol. 6004, pp. 23–32, 2010.

[13] D. You, S. Antani, D. Demner-Fushman, V. Govindaraju, and G. R. Thoma, "Detecting figure-panel labels in medical journal articles using mrf," in *ICDAR*, 2011, pp. 967–971.

[14] B. Cheng, S. Antani, R. J. Stanley, and G. R. Thoma, "Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval," in *DRR*, 2011, pp. 1–10.

[15] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, Mar. 1979.

[16] G. W. Zack, W. E. Rogers, and S. A. Latt, "Automatic measurement of sister chromatid exchange frequency." *J Histochem Cytochem*, vol. 25, no. 7, pp. 741–753, Jul. 1977.

[17] F. Sadeghian, Z. Seman, A. R. Ramli, B. H. Abdul Kahar, and M.-I. Saripan, "A framework for white blood cell segmentation in microscopic blood images using digital image processing," *Biological procedures online*, vol. 11, no. 1, pp. 196–206, 2009.

[18] C. H. Kang, W. Y. Jung, Y. H. Kang, J. Y. Kim, D. G. Kim, J. C. Jeong, D. W. Baek, J. B. Jin, J. Y. Lee, M. O. Kim, and et al., "Atbag6, a novel calmodulin-binding protein, induces programmed cell death in yeast and plants." *Cell Death and Differentiation*, vol. 13, no. 1, pp. 84–95, 2006.

[19] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.

[20] D. Kim and H. Yu, "Figure text extraction in biomedical literature," *PLoS ONE*, vol. 6, no. 1, p. e15338, 01 2011.

[21] J. R. Quinlan, *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*, 1st ed. Morgan Kaufmann, Oct. 1992.

[22] B. Cinar, P. Fang, M. Lutchman, D. Di Vizio, R. Adam, N. Pavlova, M. Rubin, P. Yelick, and M. Freeman, "The pro-apoptotic kinase mst1 and its caspase cleavage products are direct inhibitors of akt1." *EMBO J*, 2007.