Accessing bioscience images from abstract sentences
Hong Yu1,* and Minsuk Lee2
1Department of Health Sciences and 2CEAS, University of Wisconsin-Milwaukee, Wisconsin

ABSTRACT
Images (e.g., figures) are important experimental results that are typically reported in bioscience full-text articles. Biologists need to access images to validate research facts and to formulate or to test novel research hypotheses. On the other hand, biologists live in an age of information explosion. As thousands of biomedical articles are published every day, systems that help biologists efficiently access images in literature would greatly facilitate biomedical research. We hypothesize that much of image content reported in a full-text article can be summarized by the sentences in the abstract of the article. In our study, more than one hundred biologists had tested this hypothesis and more than 40 biologists had evaluated a novel user-interface BioEx that allows biologists to access images directly from abstract sentences. Our results show that 87.8% biologists were in favor of BioEx over two other baseline user-interfaces. We further developed systems that explored hierarchical clustering algorithms to automatically identify abstract sentences that summarize the images. One of the systems achieves a precision of 100% that corresponds to a recall of 4.6%.

Contact: hongyu@uw.edu or hy52@columbia.edu

1 INTRODUCTION
The rapid growth of electronic publications in bioscience has made it necessary to create information systems that allow biologists to navigate and search efficiently among them. PubMed is an information retrieval system that returns a list of documents in response to users’ queries. Arrowsmith helps biologists uncover biologically significant relations between two previously disparate fields of inquiry (Smallheiser and Swanson 1998). BioText is an information retrieval system that allows biologists to refine the retrieved MEDLINE articles and to categorize the retrieved articles based on the MeSH terms that were assigned to the articles (Hearst 2003). GeneWays is an information extraction and visualization system that extracts from literature molecular interactions related to pathways (Rzhetsky et al. 2004). iHOP is an online service that identifies sentences that relate two genes (Hoffmann and Valencia 2005). BioMedQA is a question answering system that provides short text in response to questions posed by biomedical researchers and physicians (Yu et al. 2006). See the review article (Jensen et al. 2006) for other information systems. Additionally, there are numerous annotated databases, e.g., SWISSPORT, OMIM (Hamosh et al. 2005), and BIND (Alfarano et al. 2005), that provide different levels of annotated literature information about genes and molecular interactions.

Most of the information systems, however, target text information only and ignore other important data such as images. Images (e.g., figures) are usually the “evidence” of biological experiments. An image is worth a thousand words. Biologists need to access image data to validate research facts and to formulate or to test novel research hypotheses. For example, a biologist may want to see the image (Figure 1) that supports the fact that “a stem cell can generate sebaceous glands.” Additionally, full-text articles are frequently long and typically incorporate multiple images. For example, we have found an average of 5.2 images per biological article in the journal Proceedings of the National Academy of Sciences (PNAS). Biologists need to spend significant amount of time to read the full-text articles in order to access specific images.

In order to facilitate biologists’ access to images, certain online journal publishers (e.g., Science direct) introduce a service called SummaryPlus (as shown in Figure 2) which lists images and their captions that appear in the full-text article. Such presentation has the promise of improvement over the traditional single-document-per-article format that has dominated bioscience publications since the first scientific article appeared in 1665 (Gross 2002).

We hypothesize that we can further enhance the SummaryPlus user-interface design. For example, the current SummaryPlus user-interface does not show any connections between images; this is contradictory to the fact that images reported in a full-text article are not disjointed. On the contrary, images are related to each other and typically, as a whole, leads to the conclusion of the full-text paper. Additionally, the associated text other than an image caption is frequently useful to illustrate the image content.

By working with hundreds of biologists, we conclude that much of the image data that appear in a full-text article can be summarized by the sentences in the abstract of the full-text article. Because biologists must read the abstract in order to understand a full-text article; linking abstract sentences to images will be the most effectively and convenient way for biologists to access images. This study reports our design and evaluation of BioEx (as shown in Figure 4), a user-interface that links abstract sentences to images. We further explored natural language processing approaches, in particular, hierarchical clustering to automatically link abstract sentences to images.

2 DO ABSTRACT SENTENCES CORRESPOND TO IMAGES?
We hypothesize that images reported in a full-text article can be summarized by sentences in the abstract. To test this hypothesis, we randomly selected a total of 329 biological articles that are recently published in four journals Cell (104), EMBO (72), Journal of Biological Chemistry (92), and Proceedings of the National Academy of Sciences (PNAS) (61). For each article, we emailed the corresponding author and invited him or her to identify abstract sentences that summarize image content in that article. In order to
eliminate the errors that may be introduced by sentence boundary ambiguity, we manually split abstract sentences and sent the sentences as the email attachments. A total of 119 biologists from 19 countries participated voluntarily in the annotation to identify abstract sentences that summarize figures or tables in their publications, resulting in a total of 114 annotated articles (39 Cells, 29 EMBO, 30 Journal of Biological Chemistry, and 16 PNAS), a collection that is 34.7% of the total articles we requested. The responding biologists included the corresponding authors to whom we had sent emails, as well as the first authors of the articles to whom the corresponding authors had forwarded our emails. None of the biologists were compensated.

This collection of 114 full-text articles incorporates 742 figures, 75 tables, and 826 abstract sentences. The average number of figure or table per document is 7.2±1.7 and the average number of sentences per abstract is 7.2±1.7. Our data show that 87.9% figures and 85.3% tables correspond to abstract sentences and 66.5% abstract sentences correspond to images; those statistics have empirically validated our hypothesis that image content can be summarized by abstract sentences. Since an abstract is a summary of a full-text article, our results also empirically validate that images are important content in full-text articles.

Table 1. The numbers of links between abstract sentences to images

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>151</td>
</tr>
<tr>
<td>1:2</td>
<td>145</td>
</tr>
<tr>
<td>1:3</td>
<td>53</td>
</tr>
<tr>
<td>1:4</td>
<td>26</td>
</tr>
<tr>
<td>1:5</td>
<td>9</td>
</tr>
<tr>
<td>1:6</td>
<td>4</td>
</tr>
<tr>
<td>1:7</td>
<td>1</td>
</tr>
<tr>
<td>2:1</td>
<td>173</td>
</tr>
<tr>
<td>3:1</td>
<td>36</td>
</tr>
<tr>
<td>4:1</td>
<td>14</td>
</tr>
<tr>
<td>5:1</td>
<td>2</td>
</tr>
</tbody>
</table>

1:1: An abstract sentence is linked to only one image and the image is only linked to the abstract sentence.
1:N: An abstract sentence is linked to N images, N > 1.
N:1: N abstract sentences are linked to one image, N > 1.

Note that the total number of tables is a small fraction (10.1%) of the total number of figures. Furthermore, out of the four journals, only EMBO includes Table as images. The total number of table images in our data collection is 15, which represents only 2% of the total image files. We have therefore focus only on the 742 figure images in this study.

We identified three types of links between abstract sentences and images. One-to-one is defined as an abstract sentence that is linked to only one image and the image is only linked to the abstract sentence. One-to-many is defined as an abstract sentence that is linked to two or more images. Many-to-one is defined as an image that is linked to two or more abstract sentences. Table 1 shows the numbers of the three categories in our 114 annotated full-text articles.

After manually examining the annotated articles, we found that we could approximately group full-text articles into four link patterns (examples are shown in Figure 3) based on the positions in which abstract sentences or images orderly appear in the abstract or the full-text articles. In Figure 3a, the abstract sentences are aligned with images in the order they appear in the full-text articles. Figure 3b shows that abstract sentences do not correspond to images in the order they appear in the full-text articles. Figure 3c shows that images are linked to only a few abstract sentences. Figure 3d shows that some images are aligned with images in the order they appear in the full-text articles and some do not. We speculate that the link patterns may be useful as additional features for inference authorship. Previously, word frequency has been explored for this task (Mosteller and Wallace 1963). On the other hand, the irregular alignment has made the task of automatically mapping abstract sentences to images more challenging, which will be discussed in Section 4.

3 BIOEX USER-INTERFACE DESIGNS AND EVALUATION

We have shown in Section 2 that biologists have judged that 87.9% images in the total of 114 full-text publications can be summarized by abstract sentences. We hypothesize that accessing images by
abstract sentences is an improvement over the SummaryPlus user-interface because the former will overcome the disadvantages of disjoint image content and may be the most efficient way to access images.

In order to evaluate whether biologists would prefer to accessing images from abstract sentence links, we have designed BioEx (Figure 4) and two other baseline user-interfaces (Figure 5). All three user-interfaces can be accessed at http://dmbi.columbia.edu/~yuh9001/BioEx.html. BioEx is built upon the PubMed user-interface except that images can be accessed by the abstract sentences. We have chosen the PubMed user-interface design because it has more than 70 million hits a month and represents the most familiar user-interface to biologists. Other information systems have also adapted the PubMed user-interface for similar reasons (Smalheiser and Swanson 1998; Hearst 2003). The two other baseline user-interfaces (as shown in Figure 5) were the original PubMed user-interface (Figure 5A) and a modified version of the SummaryPlus user-interface (Figure 5B), in which the images are listed as the disjointed thumbnails, rather than the links by abstract sentences.

We asked the 119 biologists who had linked sentences to images in their publications to assign a label to each of the three user-interfaces as “My favorite”, “My second favorite”, or “My least favorite”. We requested completed the evaluation. Table 2 shows their choices. Table 2 shows that 36 or 87.8% of the total 41 biologists judged that BioEx is “My favorite”. One biologist judged all three user-interfaces to be “My favorite”. Five other biologists considered SummaryPlus as “My favorite”, two of whom (or 4.9% of the total 41 biologists) judged BioEx to be “My least favorite”. The SummaryPlus user-interface was the second choice by a majority of biologists (63.4%).

A total of eighteen biologists not only evaluated the three user-interfaces, but also provided us with additional text comments. Of those 18 biologists, 17 of them made positive comments regarding to BioEx and 3 of 17 additionally made suggestions to enhance the BioEx user-interface design. Table 3 shows selected original comments made by the biologist evaluators: two are positive (C1 ~ C2); C3 is negative; and C4 shows a suggestion to enhance the BioEx interface design.

4 STRATEGIES FOR LINKING ABSTRACT SENTENCES TO IMAGES

One way to implement BioEx is to ask the authors of a paper to link abstract sentences to images. However, currently, PubMed has more than 15 million citations. It is not feasible to ask the authors to perform such a large scale of annotation, although it may be feasible for the publishers to request such a task when a new manuscript is accepted. In order to implement BioEx, we need to explore approaches to automatically identify abstract sentences that summarize images.

We may simplify the task of linking abstract sentences to images as a task of aligning abstract sentences to other associated text (i.e., captions and other embedded text) that correspond to the same images. Such simplification is based on two assumptions. The first is that image content consistently corresponds to its

---

1We assume BioEx is a useful improvement over the PubMed user-interface, a baseline that has already been favored by biologists.

2Our initial BioEx user-interface applied thumbnails at the end of abstract sentences. Three biologists had made suggestions to replace the thumbnails with “Fig X”. Our current BioEx therefore is implemented based on such recommendations.
A Rac1-containing Rho Guanine-nucleotide exchange factor is required for mitotic spindle assembly.

Rafkind MD, Kucharczuk JH, Wei K.
Department of Radiation and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA.

Controversy independent microtubule polymerization normal chromosomes have been shown to require a local gradient of RacGTP, which discharges mitotic spindles from the nuclear import of microtubule motors. Here, we have used an activity-based assay of Xenopus egg extracts to study the microtubule length of the spindle assembly. Factor regulated by this pathway, RacGTP, is a mitotic spindle, a factor that binds directly to microtubule motors. Regulation of RacGTP activity is coregulated in a mitotic spindle assembly, a factor that binds directly to microtubule motors. Regulation of RacGTP activity is coregulated in a mitotic spindle assembly, a factor that binds directly to microtubule motors. Regulation of RacGTP activity is coregulated in a mitotic spindle assembly, a factor that binds directly to microtubule motors. Regulation of RacGTP activity is coregulated in a mitotic spindle assembly, a factor that binds directly to microtubule motors.

Figure 6 shows that both TF*IDF weighted cosine similarity and Dice score can separate linked pairs from unlinked pairs and the TF*IDF weighted cosine similarity shows an advance over the Dice's score for separating the two. The results empirically validate our assumption that there are word similarities between abstract sentences to their corresponding image captions.

We may explore different models to map abstract sentences to images. For example, linking abstract sentences to image captions and other associated text can be treated as a task of sentence alignment in machine translation. However, we consider that the former is a more challenging task than the latter. In machine translation, most of the sentences are aligned and typically a majority of sentences are aligned one to one (i.e., one sentence is translated to only one sentence in the second language). For example, in (Gale and Church 1993), 89% sentences belonged to this category. However, in our data collection, many abstract sentences and images do not have any corresponding images or sentences and many abstract sentences and images correspond to two or more images and abstract sentences (details in Table 1).

Furthermore, techniques that were successful in machine translation might not apply to our task of linking abstract sentences to images. For example, sentence length (i.e., a long sentence must be translated to a long sentence in another language) was found to be powerful in sentence alignment. However, we do not find direct correspondence between the length of an abstract sentence and the length of the corresponding image caption. Additionally, in machine translation, most of the sentences were aligned in the order they appear. However, orderly alignment does not apply to many cases in our data collection (examples shown in Figure 3b-d). We therefore explored a model that applies hierarchical clustering algorithms to cluster abstract sentences and images based on word similarities which has shown in Figure 6 to be able to separate linked abstract image pairs from unlinked ones. In our application, if abstract sentences belong to the same cluster that includes images, the clustering model holds advantages over other models in that the clustering methods flexibly allow “one-to-many” and “many-to-one” mapping. Furthermore, we will show later (Section 5.3) that it is a relatively a simple task to incorporate positional information.

5 APPLYING HIERARCHICAL CLUSTERING ALGORITHM FOR AUTOMATICALLY LINKING ABSTRACT SENTENCES WITH IMAGES

Hierarchical clustering algorithms are well-established algorithms that are widely used in many other research areas including
neighboring text, synonyms, or combined. We have explored bag-of-words and n-grams as features for the clustering tasks. Additionally, we have explored different feature combinations that include features in caption, other associated text, clustering strategies. We have previously shown that the TF*IDF weighted cosine similarity because we had identified this “other associated text” by surface cues: we extract paragraphs incorporating “Figure X” from the full-text article, then merge these paragraphs with the corresponding image captions and subject the merged text to the clustering procedure. Our approach stems from the fact that biologists frequently devote an entire paragraph or more to describing the results of one experiment.

(3) Neighbouring Text Abstract sentences are coherent and the neighbouring sentences (the preceding and the following sentences) may be content-related. Furthermore, we found that 135 out of the total 746 images or 18% images in our data collection correspond to consecutive abstract sentences. For example, Figure 3 shows that the two abstract sentences “a purified Rae1 complex stabilizes microtubules in egg extracts in a RanGTP/importin beta-regulated manner” and “interestingly, Rae1 exists in a large ribonucleoprotein complex, which requires RNA for its activity to control microtubule dynamics in vitro” point to the same image “Fig 6”. We therefore explored “neighbouring text” as additional features: we merged the features of the neighbouring abstract sentences, namely, the previous and the following sentences, with the abstract sentence to be examined and applied the merged features to identify images that are associated with the abstract sentence.

Table 2. Preferences made by 41 biologists who evaluated the three user-interfaces

<table>
<thead>
<tr>
<th></th>
<th>Favorite</th>
<th>Second Favorite</th>
<th>Least Favorite</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>1</td>
<td>11</td>
<td>29</td>
</tr>
<tr>
<td>SummaryPlus</td>
<td>6</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td>BioEx</td>
<td>36</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3. Comments made by biologists who evaluated BioEx and two other baseline user-interfaces

C1. C (i.e., BioEx) would be useful, because one can easily confirm the strength/validity of a sentence in the abstract. Sometimes I search abstracts looking for information on a specific question; this would be helpful to evaluate the abstracts. B (i.e., SummaryPlus) is not very useful, because random images are difficult to interpret.

C2. Adding links to the figures significantly facilitates more in depth skimming of the literature. Case B (i.e., SummaryPlus) is a significant improvement over case A (i.e., PubMed). Case C (i.e., BioEx) simplifies accessing the appropriate figures to evaluate the approaches used and is a useful improvement over case B.

C3. The second (i.e., SummaryPlus) permits accessing figures while retaining continuity of the abstract and remaining an economical extension and improvement to the existing PubMed system.

C4. Instead with thumbnails the links could be labeled with Fig X. This would be more informative.

5.1 Word features
We have explored bag-of-words and n-grams as features for the clustering tasks. Additionally, we have explored different feature combinations that include features in caption, other associated text, neighboring text, synonyms, or combined.

(1) Caption An image caption usually incorporates multiple sentences or phrases. The heading usually provides an abstraction of the entire image content and the first sentence of each subheading provides a summary of each sub-experiment. We have explored the combinations of the heading and the first sentences of the subheading. Specifically, we explored 1) all words in the caption, 2) heading plus the first sentence of each sub-experiment in the image caption, and 3) the first sentence of each sub-experiment.

(2) Other Associated Text The image caption is not the only content that describes the experiment. There is other associated text in the full-text document that may provide additional discriminating features for clustering. We have identified this “other associated text” by surface cues: we extract paragraphs incorporating “Figure X” from the full-text article, then merge these paragraphs with the corresponding image captions and subject the merged text to the clustering procedure.

(3) Neighbouring Text Abstract sentences are coherent and the neighbouring sentences (the preceding and the following sentences) may be content-related. Furthermore, we found that 135 out of the total 746 images or 18% images in our data collection correspond to consecutive abstract sentences. For example, Figure 3 shows that the two abstract sentences “a purified Rae1 complex stabilizes microtubules in egg extracts in a RanGTP/importin beta-regulated manner” and “interestingly, Rae1 exists in a large ribonucleoprotein complex, which requires RNA for its activity to control microtubule dynamics in vitro” point to the same image “Fig 6”. We therefore explored “neighbouring text” as additional features: we merged the features of the neighbouring abstract sentences, namely, the previous and the following sentences, with the abstract sentence to be examined and applied the merged features to identify images that are associated with the abstract sentence.

(4) Synonym Expansion Abstract sentences and image captions do not always use the exact same words. Synonym expansion
might enhance the clustering performance. We applied the large biomedical knowledge resource the Unified Medical Language System (UMLS) (Humphreys et al. 1998) to expand synonyms. The UMLS incorporates more than one million biomedical concepts with synonyms. We applied a simple string matching to capture the terms and to map terms to the UMLS concepts and synonyms.

5.2 Word weight
For document clustering, we applied the TF*IDF weighted cosine similarity, which was shown in the previous section 4 to have a better discrimination than the Dice’s score. We treat each sentence or image caption as a ‘‘document’’ and the features are bag-of-words. We explored three different methods to obtain the TF*IDF value for each word feature:

(1) IDF(abstract+caption): the IDF values were calculated from the pool of abstract sentences and image captions;
(2) IDF(full-text): the IDF values were calculated from all sentences in the full-text article;
(3) IDF(abstract):IDF(caption): we obtained two sets of IDF values. For words that appear in abstracts, the IDF values were calculated from the abstract sentences; for words that appear in image captions, the IDF values were calculated from the image captions.

5.3 Position
Although we show that in many of the annotated full-text articles, the abstract sentences do not correspond to images in the order they appear in the full-text articles (examples shown in Figure 3b–d), we found that the chance that two abstract sentences or images link to an image or an abstract sentence decreases when the distance between two abstract sentences or images increases. For example, two consecutive abstract sentences have a higher probability to link to one image than two abstract sentences that are far apart. Such ‘‘positional distance’’ also applies to images: two consecutive images have a higher chance to link to the same abstract sentence than two images that are separated by many other images. To integrate such positional information into our existing hierarchical clustering algorithms, we modified the TF*IDF weighted cosine similarity with positional distance. Assuming that we consider an abstract sentence or an image caption as a document, the TF*IDF weighted cosine similarity for a pair of document $i$ and $j$ is:

$$SIM(i, j) = \text{sim}(i, j) * \left(1 - \frac{P_i}{T_i} \frac{P_j}{T_j}\right)$$

(4)

(1) If $i$ and $j$ are both abstract sentences, $T_i=T_j=\text{total number of abstract sentences}$; and $P_i$ and $P_j$ represents the positions of sentences $i$ and $j$ in the abstract.

(2) If $i$ and $j$ are both image captions, $T_i=T_j=\text{total number of images}$ that appear in a full-text article; and $P_i$ and $P_j$ represent the positions of image(s) $i$ and $j$ in the full-text article.

(3) If $i$ and $j$ are an abstract sentence and an image caption, respectively, $T_i=\text{total number of abstract sentences}$ and $T_j=\text{total number of images}$ that appear in a full-text article; and $P_i$ and $P_j$ represent the positions of abstract sentence $i$ and image $j$.

5.4 Clustering strategy
Although there are a great deal of word similarities between abstract sentences and their corresponding image captions, there are also significant differences between the two texts. In general, image captions tend to be long and incorporate content-lean experimental details. For example, the image caption (Fig 1) in Figure 3 is

‘‘(A) Schematic of the assay used to identify Rae1. Sequential affinity chromatography steps were used to deplete metaphase-arrested CSF Xenopus egg extracts: first, a RanGTP matrix was used to remove RanGTP binding proteins including importin $\beta$ (ARanBP Extract), freeing cargoes that caused spontaneous microtubule aster formation…’’

which is in contrast to its succinct abstract sentence:

‘‘Here, we have used an activity-based assay in Xenopus egg extracts to purify the mRNA export protein Rae1 as a spindle assembly factor regulated by this pathway’’.

To best capture the differences between abstract sentences and image captions, we explored three clustering strategies; namely, per-image, per-abstract sentence, and mix.

(1) Per-image clusters each image caption with all abstract sentences. The image is assigned to (an) abstract sentence(s) if they belong to the same cluster. This method values features in abstract sentences more than image captions because the decision that an image belongs to (a) sentence(s) depends upon the features from all abstract sentences and the examined image caption. The features from other image captions will not play a role for the clustering.

(2) Per-abstract-sentence takes each abstract sentence and clusters it with all image captions that appear in a full-text article. Images are assigned to the sentence if they belong to the same cluster. This method values features in image captions higher than the features in abstract sentences because the decision that an abstract sentence belongs to image(s) depends upon the features from the image captions and the examined abstract sentence. The features from other abstract sentences will not play a role for the clustering.

(3) Mix clusters all image captions with all abstract sentences. This method treats features in abstract sentences and image captions equally.

In addition, because the clusters generated by the hierarchical clustering algorithms are typically mutually exclusive, Mix will never achieve 100% accuracy for detecting the following links:

<table>
<thead>
<tr>
<th>Abstract sentence 1</th>
<th>Image 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract sentence 2</td>
<td>Image 2</td>
</tr>
</tbody>
</table>

If grouping into four clusters, Mix will create three false negatives; if grouping into three clusters, Mix will create at least two false negatives; if grouping into two clusters, Mix will create one false negative; if grouping into one cluster, Mix will create one false positive.
6 DATA AND EVALUATION METRICS

The 114 bioscience articles we described previously (Section 2) were used to evaluate the mapping between abstract sentences and images. We report recall and precision as the evaluation metrics for linking sentences to images. Recall is the total number of correctly predicted links divided by the total number of annotated links. Precision is the total number of correctly predicted links divided by the total number of predicted links.

7 RESULTS AND DISCUSSION

Figures 7–13 show the results in which we explored different combinations of features and algorithms. The default parameters for all these experiments were “per image”, “without UMLS synonyms”, “bag-of-words”, and “IDF(abstract_caption), “without neighboring sentences” and “without position”.

Figure 7 shows the results in which we explored image captions, the combined heading with the first sentence from each sub-experiment, and the first sentence from each sub-experiment. The results show that incorporating all image captions as features leads to a slightly better performance over the other features.

Figure 8 shows that the clustering performance increases when features include other associated text. The results directly support our assumptions that other associated text represents images content and that there are lexical similarities between abstract sentence and other associated text that correspond to an image. Because the feature spaces have been expanded, the overall recall and precision have increased. On the other hand, the high-end precision has dropped from 100% to 80%. This can be explained by the fact that although other associated text may incorporate useful word features that do not appear in captions, they may also include words that never appear in the corresponding abstract sentences, and those words introduce “noise” at the clustering. Additionally, we currently implemented a simple approach for identifying other associated text: we identified the entire paragraph as the “other associated text” if the paragraph contains the surface cue.
Figure X. The approach will introduce significant “noise” because frequently, a paragraph may describe more than one experiment.

Figure 9 shows that “without neighboring sentences” greatly outperformed “with neighboring”. Recall that neighboring sentences are adjacent sentences (or proceeding and following sentences) of the examined sentence. The results conclude that the “useful” information introduced by the neighboring sentences is overshadowed by the “noise”. The results are not entirely surprising. Although 18% images in our data collection correspond to consecutive abstract sentences, we found that a majority of images do not. Specifically, 424 (57.1%) images correspond to single abstract sentences, 91 (12.3%) images correspond to non-consecutive abstract sentences, and 92 (12.4%) images do not link to any of abstract sentences.

Figure 10 shows that synonym expansion has a disappointing performance. The results may contribute to several factors, including how robust was the mapping between a string to the UMLS concepts and the problems of homonyms. We will describe in the next section (Section 8 Future Work) how we will explore different approaches to enhance synonymous term identification.

Our results also show little performance differences between unigram and n-grams (data not shown). The results are not surprising because of the problem of data sparseness. Many other natural language processing systems have found little gain of n-gram in either topic detection (Lee et al. 2006) or document and sentence classification (Yu and Hatzivassiloglou 2003).

Figure 11 shows the performance of three different methods for calculating the IDF values. The results show that the “global” IDFs, or the IDFs obtained from the full-text article, has a much lower performance than “local” IDFs, or IDFs calculated from the abstract sentences and image captions. The results suggest that abstract sentences and image captions alone are more accurate than the whole full-text article for estimating the importance of features in our task of linking abstract sentences to image captions. In addition, IDFs that were separately calculated from the abstract sentences and image captions performs slightly better than the combined IDFs. The results suggest that the distributions of features are different between abstract sentences and image captions.

Recall that we have explored three strategies for linking abstract sentences to images; namely, Per-image that takes each image caption and clusters it with abstract sentences, Per-abstract-sentence that takes each abstract sentence and clusters it with image captions, and Mix that clusters all image captions with all abstract sentences. As we have predicted, Figure 12 shows that both Per-image and Per-abstract-sentence out-performs Mix. Furthermore, Per-image significantly out-performs Per-abstract-sentence. The results suggest that features in abstract sentences are more useful than features in caption for the task of clustering.

Figure 13 shows that combining word features with position has significantly enhanced the performance. When the recall is 33%, the precision of combining TF*IDF with positional information increases to 72% from the original 38%, which corresponds to a 34% absolute increase. The results strongly indicate the importance of positional information. When the precision is 100%, the recall is 4.6%. We believe that a high precision is the key to success for
this application. Many previous successful and applicable natural language processing systems have also achieved high precisions (e.g., (Friedman et al. 2001)). However, the low recall will render our current system’s application for the real application. We have implemented BioEx (with a recall of 33% and a precision of 72%) that can be accessed at http://dbmi.columbia.edu/~yuh9001/BioEx.html, from which a user can query 17,000 downloaded full-text articles.

Recall that our evaluation data consists of three types of mapping between abstract sentences and images. They are one to one, one to many and many to one. Previous dynamic programming methods in machine translation had shown significant decreases in performance when a sentence was aligned to multiple sentences (Gale and Church 1993). We therefore examined the performance of our algorithms for each type. Since we could not measure the precision for this task because we miss the false positives for each type, we compared the recall for different type (results shown in Table 4). We chose the system with the overall f-score=44.4% (F-score=2∗recall∗precision/(recall+precision)). Our results, in contrast, do not show significant differences in recall among three types of mapping. Our results may support the robustness and advantages of the hierarchical clustering methods over the dynamic programming method for this application.

### 8 FUTURE WORK

Our current evaluation data were annotated by biologists who are the authors of their publications. We have observed inconsistency in annotation. For example, by manually examining 114 annotated full-text articles, we found that many biologists assigned images to conclusions and speculations, while others did not. For example, the last sentence in the abstract (pmid=15933717) is:

```
“Taken together, our observations suggest that PIASy is a critical regulator of mitotic SUMO-2 conjugation for Topoisomerase-II and other chromosomal substrates, and that its activity may have particular relevance for centromeric functions required for proper chromosome segregation.”
```

which was assigned to all five images that appeared in the full-text article. On the other hand, in two other articles (Figure 3a and d), the authors did not assign any images to their corresponding conclusions and speculations. We believe that such inconsistency may be fixed in the future if with a carefully designed annotation instruction. Additionally, future work one may need to measure the inter-rater reliability for linking abstract sentences to images; this requires that one article must be annotated by two or more people. One way is to ask the co-authors to annotate their articles independently and then measure the agreement among them.

We believe that there are rooms to enhance the performance for linking abstract sentences to images. In this study, we applied word-level similarity to measure the link between abstract sentences to images. However, exact word matching may not be the best solution in this application. One may need to capture synonymous terms. For example, if we could capture the abbreviation “NMR” and map it to the corresponding full form “nuclear magnetic resonance”, the clustering algorithm would be able to link the two texts:

```
“Here we report nuclear magnetic resonance and X-ray protein structures of the N-terminal substrate recognition domain of FimD (FimDN) before and after binding of a chaperone C subunit complex” (an abstract sentence)
```

and

```
“NMR studies on FimD N…” (an image caption; pmid=15920478).
```

One may explore the work of (Aronson 2001) that applied the large biomedical knowledge resource the Unified Medical Language System (UMLS) for synonym identification and the work of (Yu et al. 2002) that explored rule-based approach for capturing abbreviations and full forms from literature.

Additionally, it may also be important to capture semantic similar terms. For example, if we link “Death” to “toxicity,” we could recover the link between the following two statements:

```
“Acute and chronic exposure to kainate caused extensive oligodendrocyte death in culture” (an abstract sentence)
```

and

```
“Kainate toxicity in oligodendrocytes derived from P7 rat optic nerves” (an image caption; pmid=9238063).
```

For identifying semantically related terms, one may explore the work of (Lin 1998a; Lin 1998b; Yu and Agichtein 2003)

This work has mainly explored word similarity between abstract sentences and image captions. Future work one may explore document rhetorical structure to assist on the task. For example, biological full-text articles typically include the Result and Discussion section, which further incorporate multiple paragraphs. Each paragraph may be thought of as a semantic unit. The first and the last sentence of each paragraph have higher chance to summarize the content of the paragraph. One may explore matching last sentence of Result and Discussion paragraph to abstract sentences.

In addition, the semantic relations among images may be captured by the distance of image descriptions. For example, the following paragraph indicates that the two figures (Figure 1 and 2) are linked:

```
“Plant RalGAPs do not contain this C-terminal domain, but contain the N-terminal MAF1-like WPP domain instead (ref. 20
```

<table>
<thead>
<tr>
<th>Type</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>32.5%</td>
</tr>
<tr>
<td>1:2</td>
<td>49.3%</td>
</tr>
<tr>
<td>1:3</td>
<td>36.5%</td>
</tr>
<tr>
<td>1:4</td>
<td>30.8%</td>
</tr>
<tr>
<td>1:5</td>
<td>24.4%</td>
</tr>
<tr>
<td>1:6</td>
<td>29.2%</td>
</tr>
<tr>
<td>1:7</td>
<td>57.1%</td>
</tr>
<tr>
<td>2:1</td>
<td>44.2%</td>
</tr>
<tr>
<td>3:1</td>
<td>39.8%</td>
</tr>
<tr>
<td>4:1</td>
<td>35.7%</td>
</tr>
<tr>
<td>5:1</td>
<td>50.0%</td>
</tr>
</tbody>
</table>

### Table 4. The recall values for different types of links.
and Fig. 1B). All plant sequences have a domain with similarity to the leucine-rich repeat (LRR) domain, followed by an acidic domain (Fig. 1A and Fig. 2A).  

Although we explored hierarchical clustering methods in this study and had shown the advantages of these methods. Future work one may explore dynamic programming that has been successful for many other tasks including sequence alignment (Lawrence et al. 1993), gene or protein name recognition (Krauthammer et al. 2000), paraphrasing (Barzilay and Lee 2003), and sentence alignment (Chen 1993). The current algorithm does not consider the “ordering effect”, which is that the position order of image pairs (i.e., one image appears ahead of the other image) reflects the position order of abstract sentences or the abstract sentence(s) appear with the same order of corresponding images. Although such alignment does not apply to every full-text article, we found that out of the total of 1649 image pairs in our 114 annotated full-text articles, 1207 or 73.6% image pairs appear with the same order of their corresponding sentences. Dynamic programming methods had shown to be powerful for detecting such alignment.

9 CONCLUSION

As described in this paper, we have designed and evaluated a novel user-interface BioEx that allows biologists to directly access images by abstract sentences. Current, more than 40 biologists evaluated the BioEx user-interface and 87.8% of them were in favor of BioEx over two other baseline systems. Additionally, we have also explored natural language processing approaches, specifically, the hierarchical clustering algorithms, to automatically link abstract sentences to images. We have explored different features and algorithms. One of the best systems shows a performance of 100% precision with 4.6% recall. We believe a high precision is a key to success for this application, although BioEx may not be applicable to real use at the current stage. We have implemented BioEx (with a recall of 33% and a precision of 72%) that can be accessed from the link at http://dbmi.columbia.edu/~yuh9001/BioEx.html, from which biologists can query 17,000 downloaded Proceedings of the National Academy of Sciences (PNAS) full-text articles.

ACKNOWLEDGEMENTS

The authors thank 119 biologists who had annotated data for us and 41 biologists who had evaluated the BioEx user-interface. The list of biologists is available upon request. The authors in particular thank Dr. Weiqing Wang for her contribution to this work. Dr. Wang helped us identify the importance of this project; collected some of the annotations; and evaluated the BioEx user-interface. This work was supported by JDRF 6-2005-835. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect JDRF’s views.

REFERENCES


