AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

OCR-Based Image Features for Biomedical Image and Article Classification: Identifying Documents Relevant to Genomic Cis-Regulatory Elements

Hagit Shatkay[1,2,3], Ramya Narayanaswamy[1], Santosh S Nagaral[1], Na Harrington[3], Rohith MV[1], Gowri Somanath[1], Ryan Tarpine[4], Kyle Schutter[4], Tim Johnstone[4], Dorothea Blostein[3], Sorin Istrail[4], Chandra Kambhamettu[1]

[1]Dept. of Computer and Inf. Sciences, University of Delaware, Newark, DE
[2]Center for Bioinformatics & Computational Biology, University of Delaware, Newark, DE
[3]School of Computing, Queen's University, Kingston, Ontario, CA
[4]Center for Computational Molecular Biology, Dept. of Computer Science, Brown University, Providence, RI
Corresponding author: *shatkay@cis.udel.edu*

Abstract

Images form a significant information source in published biomedical articles, which is under-utilized in biomedical document classification and retrieval. Much current work on biomedical image retrieval and classification uses simple, standard image representation employing features such as edge direction or gray scale histograms. In our earlier work (Shatkay Chen, and Blostein, 2006) we have used such features as well to classify images, where image-class-tags have been used to represent and classify complete articles.

Here we focus on a different literature classification task: identifying articles discussing cis-regulatory elements and modules, motivated by the need to understand complex gene-networks. Curators attempting to identify such articles use as a major cue a certain type of image in which the conserved cis-regulatory region on the DNA is shown. Our experiments show that automatically identifying such images using common image features (such as gray scale) is highly error prone. However, using Optical Character Recognition (OCR) to extract alphabet characters from images, calculating character distribution and using the distribution parameters as image features, forms a novel image representation, which allows us to identify DNA-content in images with high precision and recall (over 0.9). Utilizing the occurrence of DNA-rich images within articles, we train a classifier to identify articles pertaining to cis-regulatory elements with a similarly high precision and recall. Using OCR-based image features has much potential beyond the current task, to identify other types of biomedical sequence-based images showing DNA, RNA and proteins. Moreover, automatically identifying such images is applicable beyond the current use-case, in other important biomedical document classification tasks.

Introduction

Classifying biomedical documents according to their relevance to a given topic is a basic step toward biomedical database curation. Classification also forms a major component in biomedical text mining applications. For instance, consider the process used by the Mouse Genome Informatics (MGI) resource at the Jackson labs (Eppig et al, 2005), in which curators need to identify published literature containing information about gene expression in the mouse (Smith et al, 2007; Hersh et al, 2005). A first step in this process requires obtaining all and only articles describing experiments relevant to this topic. The articles are then read, and the most significant information is extracted and curated. Another example involves finding articles containing experimental evidence for protein-protein interaction. The latter task was part of the challenge posed in BioCreative III (Krallinger et al, 2011).

Images within articles provide significant cues to curators for deciding the relevance of an article with respect to a given biological task. We are interested in using a combination of information from images and text to classify biomedical articles, as we have already shown in an earlier work (Shatkay, Chen and Blostein 2006).

During the past decade, much research has been dedicated to content-based retrieval of images and to image classification, both within and outside the biomedical domain. Most of the work is concerned with content-based categorization and retrieval of images (i.e. not of documents). To do so, a corpus of training/test images is identified, certain features are extracted from the images, the images are represented as feature-vectors, and a classifier is trained to identify certain types of images within the corpus, under the specified feature-vector representation. Features that are often used for image representation include, among others, statistics based on gray-level histograms (Gonzalez and Woods, 2002), Haralick's texture-features (Haralick, Shanmugam and Dinstein, 1973), and values from edge direction histograms (Jain and Vailaya, 1998). In our early work (Shatkay, Chen

and Blstein, 2006) we have used such features as well for image classification, where image-class-tags were used to represent and classify documents.

Here we discuss a different, specific document classification task, namely that of identifying articles discussing genomic cis-regulatory elements and modules, motivated by the need to understand complex gene-networks. The group working on the *CYRENE cis-regulatory browser project* at Brown University (Istrail et al, 2010) noted that to identify such articles in the vast literature, one can use as a significant cue a certain type of image showing the DNA and denoting the conserved cis-regulatory elements. An example of such an image is shown in Figure 1. We refer to images that show DNA content as *DNA-rich* images.

Based on our experiments, automatically identifying such images using common image features (like those mentioned above) proves highly error prone. However, using Optical Character Recognition (OCR) to extract alphabet characters from images, calculating character distribution and using the distribution parameters as image features, allows us to form a novel representation of images, and identify DNA-content in images with high accuracy. Using such DNA-rich images, we then train a classifier that identifies documents pertaining to cis-regulatory modules with high precision and recall.

While this paper focuses on the specific task of identifying cis-regulatory-related publications, the idea of using OCR as image features is applicable well beyond the current task, and can be utilized to identify other types of biomedical sequence-based images. Automatically identifying such images has much potential to be widely applicable in computational biomedicine. Throughout the rest of the paper we describe our approach, experiments and results. The next section briefly surveys image analysis in biomedical documents, highlighting the difference between previous work and the research presented here. We then discuss in more detail the specific problem we are addressing in the context of the CYRENE project, the datasets, and the methods we use to process and to represent images and articles. We follow with a section about experiments and results, followed by conclusions and an outline of future work.

Related Work

Research by Murphy et al. (e.g. Murphy et al, 2001; Cohen, Kou and Murphy, 2003; Quian and Murphy, 2008)[1] is among the earliest on using images within biomedical articles. Their work focused primarily on image categorization for identifying images and articles discussing protein subcellular localization. It constitutes an in-depth study of a specific task, namely, identifying and interpreting a certain type of microscopy images associated with protein localization experiments. The image processing employs standard image-features like the ones mentioned earlier. Notably, the tools used in that research aim at the protein-subcellular-localization task, and do not target biomedical text/image retrieval as a whole. Work by Rafkind et al. (2006) explored retrieval of biomedical images from the literature in a more general context, while work by Shatkay et al. (2006) started to examine the integration of text and image data for bioemedical document retrieval. Both used similar, standard image features such as gray-scale and edge-direction statistics.

Another related area, which focuses on image processing in the biomedical domain, is content-based retrieval of medical images and medical documents. One may look, for instance, for *x-ray images* of a certain limb or for documents containing such images. During the past few years, shared tasks that included challenges of this nature were introduced in ImageClef[2] leading to the development of systems addressing such challenges (e.g. Demner-Fushman et al, 2009). Typically, standard image features like those mentioned earlier (texture features, gray-scale-based features etc.) are used to represent the images.

Taking advantage of text associated with images for document- or for image- retrieval typically relied on using text from figure captions (an idea introduced by Regev et al, 2002), or possibly also text referencing images from within the article's body (Yu, Liu and Ramesh, 2010). Last, as a way to improve indexing and retrieval of biomedical images, Xu, Krauthammer et al. (2008) proposed to use *optical character recognition (OCR)* to extract text from within biomedical images, using the extracted words/terms to index images. In contrast to the work presented here, Xu et al's research was not concerned with image processing, representation or classification. It viewed OCR as means to obtain text for identifying images, rather than as a source of useful image-features. This latter idea, which to the best of our knowledge was not pursued before, is the focus of this work.

CYRENE and the Article Classification Task

The CYRENE project (Istrail et al, 2010) is concerned with obtaining, providing and displaying highly reliable information about cis-regulatory genomics and gene regulatory networks (GRN). Two of its components include the cisGRN-Lexicon and the cisGRN-Browser. The lexicon is a database containing high-quality information about the sites, function, operation mechanism and other aspects of cis-regulatory elements, currently including 200 transcription factors encoding genes and 100

[1] See also SLIF: Subcellular Localization Image Finder. Carnegie Mellon University. *http://slif.cbi.cmu.edu*.

[2] ImageCLEF Medical (since 2007): Cross-Language Image Retrieval Evaluation, *http://www.imageclef.org/* .

Figure 1. *An example of a DNA-rich diagram of the type that is over-represented in articles discussing cis-regulatory elements. Taken from PMID 15115437, Figure 7. (Ulvila et al, 2004. Image obtained from PubMed Central).*

other regulatory genes. (Primarily in human, mouse, fruit fly, sea urchin and nematode, with some information pertaining to other organisms). To be included in the lexicon, a regulatory mechanism must adhere to stringent criteria of experimental validation, in-vivo. Obtaining such highly reliable information that can be placed in the database requires scanning carefully through the literature, identifying the articles that describe the cis-regulatory mechanisms and the experiments validating them, annotating the relevant information within them, and depositing the information in the database. Here we focus the first step, namely, that of identifying articles that are likely to contain the high-quality information that can be curated into the CYRENE database.

As noted by the group working on creating and curating CYRENE (of which RT, KS, TJ and SI are a part), the most relevant publications in which pertinent information can be found often contain diagrams and graphs of a particular type (referred to by the team as the *quintessential diagrams* and the *quintessential graphs*). We focus here on the diagrams, which typically display short sequences of DNA, marking conserved regions, motifs or sites that are involved in the regulatory module described in the paper. Figure 1 shows an example of such an image, taken from one of the papers used to curate information into Cyrene (Ulvila et al, 2004).

The document classification task is thus to identify, among a set of candidate publications already containing basic terms such as "regulation" or published in the relevant journals (such as *Molecular and Cellular Biology*), those that are most likely to contain experimentally validated information about cis-regulatory elements and modules. We address this task using both a text-based classifier (briefly mentioned here), and an image-based document classifier, where we focus here mainly on the latter. In the next section we discuss the data and the methods that we use for training and testing such a classifier.

Data and Methods

The Dataset: CYRENE-related Articles

The CYRENE team of curators has initially identified a set of 271 publications containing experimentally-validated information about cis-regulatory modules. To obtain this set, they read through a subset of publications in a selected set of about 60 journals (primarily drawing on the main journals that publish in the area, including: *The Journal of Biological Chemistry, Molecular and Cellular Biology, Development, Gene & Development, Developmental Biology, The EMBO Journal, Gene, Biochemical and biophysical research communications, PNAS, Nucleic Acids Research*), published after 1985. A keyword search – based on keywords such as *regulatory, transcription, DNA element, DNA motif* – was applied to the many thousands of resulting articles, to reduce the set to those articles likely to discuss gene regulation. The resulting set of a few thousands articles, was examined by the curators to identify the ones showing experimentally validated cis-regulatory modules, thus forming the set of 271 articles. The latter is the *positive set*, i.e. the set of *Relevant* articles for the classification training/testing process.

Many of the remaining published articles were rejected from the CYRENE-relevant dataset. A small subset of those irrelevant publications, consisting of 78 articles, were identified and kept by the curators, and were the basis of our *negative* set of *Irrelevant* articles. As the resulting overall set is highly unbalanced for classification purposes, (271 positive examples and only 78 negative), we selected an additional set of 143 negative examples from the Journal of Molecular Cellular Biology – which is a journal from which about 20% of the 271 relevant articles originate. The negative documents were selected by going through the same volumes from which relevant articles were obtained, and identifying 10-20 articles that were not judged to be relevant by the curators – in each of these volumes. Selecting irrelevant articles from the same volumes in which relevant articles were found ensures that the overall style and discourse remain similar across the relevant and the irrelevant articles. That is, there is no shift in time and in the overall discussion areas between the subset of relevant articles and the subset of irrelevant articles. If such a shift existed, it could over-simplify the learning task of separating the relevant from the irrelevant articles, as separation could have then relied on differences in language and style, as opposed to on the difference in actual contents. The resulting dataset thus consists of 271 positive examples (CYRENE-relevant articles) and 221 negative examples (articles that are irrelevant for CYRENE). The PDF of the complete articles was obtained for 264 of the relevant articles and for 220 of the irrelevant ones. We further describe the training and the testing of an image-based document-classifier later in this section.

Representation and Classification of Image Panels

Multiple groups have already noted that figures in biomedical publications often consist of multiple subfigures or *panels,* (Murphy et al, 2001; Shatkay et al, 2006; Yu et al, 2010), as demonstrated in Figure 2. Each

panel typically consists of an individual image, and as such, when considering images, we separate figures into individual panels. To obtain images and image panels from the PDF file we use a tool that we have developed for this purpose, based on the Xerox Rossinante utility (*https://pdf2epub.services.open.xerox.com/*). A full description of this tool is to be published elsewhere.

As noted earlier, articles that discuss cis-regulatory modules are typically characterized by an over-representation of image panels containing DNA information, such as the one shown in Figure 1. As such, we hypothesized that being able to identify such images automatically – and identifying articles that show an over-abundance of such images, would help in identifying relevant documents for the CYRENE database. Again, we refer to image panels that show DNA regions, as *DNA-rich image panels*. To automatically identify such panels, we aim to train a classifier to distinguish between DNA-rich images and all other images. To attain this goal we need:
a) A set of positive image panels that contain DNA sequences, and a set of negative images panels, which do not contain DNA sequences; and
b) An image representation using features that expose the DNA-content. Once such features are identified, all the images in the positive and in the negative set can be represented as a weighted vector of these features, and a classifier that aims to distinguish between the two types of images can be trained and tested.

To achieve goal (a) above, we identified a set of 88 DNA-rich image panels, and 100 image panels that do not show DNA sequences. We use this set of 188 panels to train and test a classifier that distinguishes between DNA-rich and non-DNA-rich images.

To represent images as feature-vectors, so that the panel-classification task could be attempted, we introduce a novel *OCR-based representation* (aim **b** above). We apply an *optical character recognition* (OCR) tool, ABBYY Finereader *(http://finereader.abbyy.com/)* to all the panels, and obtain all the characters that occur in each panel. We count the number of times each character (A-Z, 0-9, Other) occurs, and represent each panel as a *37-dimentional* feature vector $<w_1 \ldots w_{37}>$, where w_i denotes the frequency of the i^{th} character in the panel. An example of the character frequency distribution for two different image panels is illustrated in Figure 3 (in which we only show the first 26 characters A-Z). The top-left panel in the figure is a DNA-rich panel, and as such its character frequency distribution shows four distinct peaks at A, C, G and T. In contrast, the top-right does not display a DNA sequence, and as such its associated character distribution assigns relatively low, similar values to quite a few characters including A, E, F, and I, and low values to C and G. Notably, the overall character-distribution is quite robust to OCR errors, as mis-reading some characters has only a small, local impact on the overall magnitude of character counts and on the distribution as a whole.

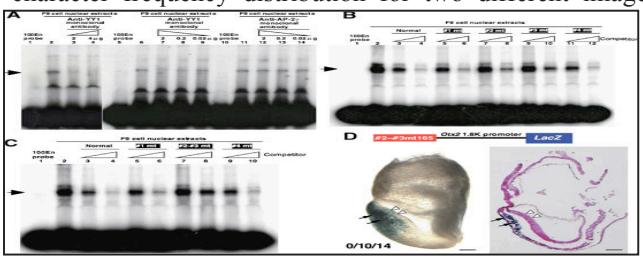

Figure 2 *An example of a composite figure, consisting of multiple image panels. Taken from PMID 17332747, Figure 2. (Takasaki et al, 2007. Image obtained from PubMed Central).*

We have also experimented with a similar, but more compact representation using a 5-dimensional vector, collapsing all characters except for A, C, G and T, into "Other", while registering the frequencies of A, C, G, and T. As shown later, the two representations perform at about the same level in our experiments. For comparison, we have also used a simple gray-scale histogram representation of all images and trained a classifier under this representation, as further discussed in the Experiments and Results section. Each of the 188 image panels is represented as a feature vector (under each of the feature types). To train and test classifiers using these representations, we use the standard WEKA tools (Witten and Frank, 2005) to train and test a decision-tree classifier, using the J48 algorithm. Further details regarding these experiments are provided in the Experiments and Results Section, as well as in a more extensive report soon to appear (Shatkay et al, 2012).

Representation and Classification of Articles

So far we described the representation and classification of image panels. However, our goal is to classify complete *articles* based on their relevance (or there lack-of) to CYRENE. The total dataset of identified articles consists of 271 positive (relevant) examples, and 221 negative (irrelevant) examples; from those we have obtained the full PDF text files for 264 positive and 220 negative articles .

Given an article *d* in the dataset, we create an image-based representation for it, by examining each image panel within the article and tagging it as DNA-rich or non-DNA-rich. While ultimately this step will be done automatically using the classifier trained on image data as described at the end of the previous section, in the experiments described here we used manual tagging of the images, to ensure independence between the results reported here for the image-classification step and those reported for the document-classification step. This issue is revisited in the Conclusions section. We then count how many panels in the article are DNA-rich and how many are not. For an article *d,* let A_d denote the number of DNA-rich panels in it, and N_d denote the number of non-DNA-rich panels. The article *d* is then represented as a simple 2-dimensional vector of the form:

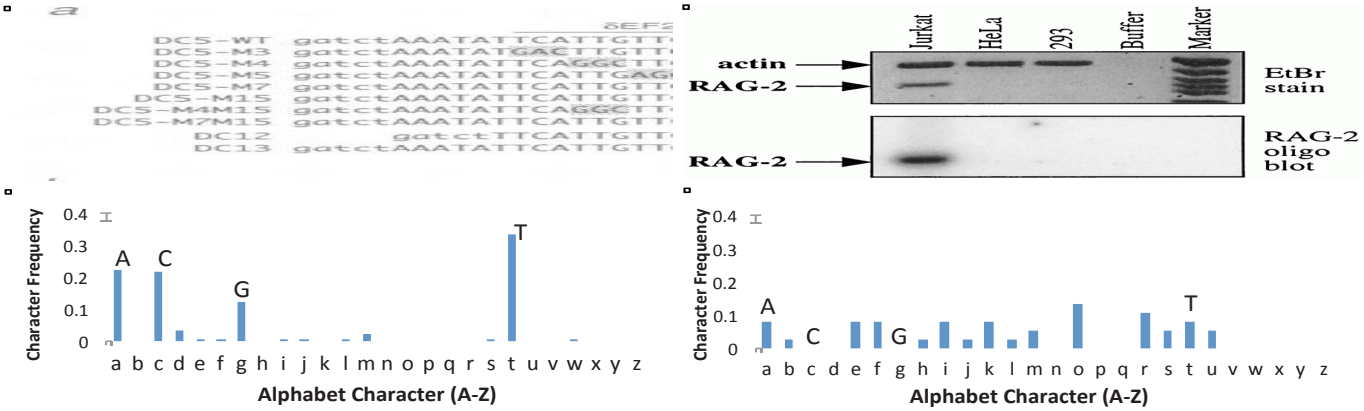

Figure 3. *An example of two distinct panels (top two panels). The respective character frequency distribution (shown only for the letters a-z) is provided below each of the panels. The top left panel (from Kamachi et al, 1995. Image btained from PubMed Central) shows a DNA-rich image, which translates to peaks on A, C, G and T in the character distribution, while the top right panel (from Wang et al, 2000. Obtained from PubMedCentral) does not display a DNA sequence, and shows a close-to-uniform distribution of all letters, all with low frequency.*

$$< A_d/(N_d + A_d)\ ;\ N_d/(N_d+A_d) >, \qquad \textbf{(Eq. 1)}$$

that is, the article is represented based on the relative frequency of its DNA-rich panels, and its relative frequency of non-DNA-rich panels. Using this simple representation of all 484 articles for which we have access to the full PDF, we again test and train a decision-tree classifier using the standard WEKA tools.

Finally, to compare the image-based classification to a text-based classification, we obtain the title and abstract of each article as they appear in PubMed and represent each article using a set of unigrams and bi-grams derived directly from the resulting corpus of text. Stop-words are excluded, and rare and frequent terms are removed. Moreover, as we have done before (Brady and Shatkay, 2008) terms that are uninformative for distinguishing between relevant and irrelevant documents (as measured within the *training* set, in each iteration of the cross-validation runs) are removed from the vocabulary. The vector representation for each article d is a simple binary vector of the form $<dt_1,\ldots,dt_n>$, where $dt_i = 1$ if the i^{th} term in the corpus-vocabulary is present in article d, and 0 otherwise. Given the relatively large number of features involved in such a representation (about 550 terms per vector), we use WEKA's naïve Bayes classifier (rather than decision tree), to train/test a classifier from the text representation of articles.

Experimental Setting

The experiments we carry out aim to examine two hypotheses: First, whether the OCR-based representation of image panels, as described earlier, is indeed effective for distinguishing DNA-rich image panels from non-DNA-rich ones, within biomedical publications; Second, whether the relative abundance of DNA-rich panels in a published article provides an effective means for assessing the article's relevance to the CYRENE dataset.

Accordingly two sets of experiments are described below: The first is concerned with image panel classification using OCR-based representation of image panels. The second set of experiments is focused on article classification, where the image-based representation of articles is used.

Classification of Image-Panels using OCR-based Representation: To evaluate the effectiveness of the OCR-based representation for distinguishing between DNA-rich and non-DNA-rich image panels, we use 188 image panels that were manually annotated for this purpose (as discussed in the previous section). For each of these image panels we construct three different representations, as follows:

1) A *37-dimentional* feature vector $< w_1^p \ \ldots\ w_{37}^p >$, where the weight in each of the first 36 positions corresponds to the relative abundance of each of the 36 characters (A-Z[3], 0-9) in the panel, while the 37th position corresponds to the relative abundance of *all other characters* combined. Thus w_i^p denotes the frequency of the i^{th} character among (A-Z,0-9,*Other*) in the image panel, that is: $w_i^p = \dfrac{\textbf{\# of times character } c_i \textbf{ occurs in panel } p}{\textit{Total \# of } \textbf{character occurrences in panel } p}$.
An example of such a representation is shown in Figure 3.

2) A *5-dimentional* feature vector $< w_1^p \ \ldots\ w_5^p >$, where the weight in each of the first 4 positions, w_1^p - w_4^p is the respective frequency of the characters A, C, G and T in the panel p, while w_5^p is the frequency of all other characters combined.

[3] While we use the upper case notation *A-Z* here, any capital letter *X* denotes here an occurrence of either the small (*x*) or the capital (*X*) letter within the image; the counts of small and capital occurrences are combined for each letter.

3) A simple gray-scale histogram representation. That is, a 256-dimensional vector $< w_1^p \ldots w_{256}^p >$, where the weight w_i^p is the number of pixels in panel *p* whose intensity level is *i*.

Under each of the representations we use WEKA's standard tools to train and test a decision tree classifier, using stratified 5-fold cross validation. That is, both the 100 positive examples and the 88 negative examples are partitioned into 5 subsets; 4/5 of both the positive and the negative examples are used for training while 1/5 are used for testing. The process is iterated 5 times, where a different 1/5 is left out at each iteration. To ensure stability of the results, we use five complete runs of 5-fold-cross-validation for each of the representations (a total of 25 runs per representation).

Classification of Articles using Image-Based Representation: To evaluate the utility of our image-based representation for the actual article-classification task, namely, separating CYRENE-relevant from non-CYRENE-relevant publications, the 484 pre-classified articles (264 CYRENE-related, 220 non-CYRENE-related, as discussed earlier) are represented using a simple 2-dimensional representation as described by Eq. 1 above. Again, we use WEKA's tools for training/testing a decision tree, but this time the classification is of articles rather than of images, and the classes are CYRENE-related vs. non-CYRENE-related. We use here five separate runs of 5-fold cross validation to ensure stability of the results.

We also tested a text-based representation of the articles, employing a simple bag-of-words model, as a point of comparison. We used text taken only from the article's title and abstract, rather than the full-text PDF (see Shatkay et al, 2012, for a discussion of this choice). The titles and the abstracts of the 484 articles – both positive and negative examples – were tokenized to obtain a dictionary of terms consisting of *single words (unigrams)* and pairs of *consecutive words (bigrams),* where words were stemmed using the Porter stemmer (Porter, 1997), and standard stop-words removed. Rare terms (appearing only in a single article) as well as very frequent ones (occurring in more than 60% of the articles) were also removed. The remaining set of terms was further reduced by selecting only distinguishing terms. These are terms whose probability to occur in positive (CYRENE-relevant) articles is statistically significantly different from their probability to occur in negative (non-CYRENE-relevant) articles. This is done following a procedure we have used in an earlier work (Brady and Shatkay, 2008). The resulting vocabulary of about 550 terms is used to represent each article *d* as a 550-dimensional vector of binary values, $< w_1^d \ldots w_{551}^d >$, where w_i^d =1 if the i^{th} term, t_i, occurs in document *d*, i.e. $t_i \in d$, and w_i^d =0 otherwise.

We use the naïve Bayes classifier in the WEKA tools, (as opposed to a decision tree, because of the higher-dimensionality of the representation), and employ 10-fold cross validation to train and test the classifier (for a more complete version of the work, in which 5-fold cross validation was used, see Shatkay et al, 2012).

The performance of all the classifiers described above is evaluated using the standard measures of *Precision, Recall*, *F-measure,* and overall accuracy (*Acc*) formally defined as:

$$Recall = \frac{TP}{TP+FN} \; ; \; Precision = \frac{TP}{TP+FP};$$

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \; ; \; Acc = \frac{TP+TN}{TP+FN+TN+FP} \; '$$

where *TP, FP, TN*, and *FN* denote the number of true positives, false positives, true negatives and false negatives, respectively. Notably a "positive" instance is a DNA-rich panel in the context of panel-classification, while it is a CYRENE-relevant article in the context of the article classification task.

Table 1. *Image-panel classification performance, averaged over 5 independent runs of 5-fold cross validation. The top two rows show results (Precision, Recall, Accuracy and F-measure) when the panel is represented using OCR-based features, while the bottom row shows results obtained using a gray-scale histogram representation. Standard deviation is shown in parentheses.*

Panel Representation	Avg Prec. (STD)	Avg Recall (STD)	Avg Acc. (STD)	Avg F
OCR: A-Z,0-9; Other	0.92 (.015)	0.89 (.015)	0.91 (.012)	0.90
OCR: ACGT; Other	0.93 (.006)	0.90 (.014)	0.92 (.007)	0.92
Gray-scale Hist.	0.64 (.009)	0.66 (0.00)	0.67 (.008)	0.65

Results

Image-Panel Classification: Table 1 summarizes the average image-panel classification results obtained from running five runs of stratified 5-fold cross validation, under each of the three image-panel representations we have used, as described above. The top two rows show the precision, recall, accuracy and F-measure when the OCR-based features are used to represent each image panel. The topmost results are of using a 37-dimensional vector, where the counts for each of the 26 alphabet letters and each digit (0-9) form separate feature values, and the counts for all other non-alphanumeric characters are grouped together into the 37th feature value. The middle-row shows the results for a more condensed 5-dimensional representation, where separate counts are calculated *only* for the letters *A,C,G,T,* and all other characters are grouped together into a fifth feature.

The top two rows show an average precision above 0.9 while the average recall is about 0.9. The second row shows slightly higher values than the first, but these differences are not statistically significant (p>>0.1).

In contrast, the third row, where image panels are represented based on their gray-scale histogram, shows a significantly lower performance on all measures. The

Table 2. *Article classification results, averaged over multiple cross-validation runs. The top row shows the results from using an image-panel based representation of each article, i.e. as a 2-dimensional vector representing the proportion of DNA-rich panels and of non-DNA-rich panels. The second row shows the results when using a standard binary term-vector representation, over a set of 551 distinguishing terms.*

Article Representation	Avg Prec. (STD)	Avg Recall (STD)	Avg Acc. (STD)	Avg F
Img-panel distribution (2-dimensional vector)	0.87 (.000)	0.89 (.000)	0.89 (.000)	0.88
Text (551-dimensional vector)	0.86 (.003)	0.86 (.007)	0.85 (.004)	0.84

difference in performance with respect to the top two rows is highly statistically significant ($p<0.0001$).

Article Classification: Table 2 shows average results, obtained from running five separate article-classification runs of stratified cross validation, using the image-panel-based representation and the text-based representation of articles. Recall that the image-based representation of an article is simply a 2-dimensional vector containing the proportion of DNA-rich panels and of non-DNA-rich panels in the article. The text-based representation is a 551-dimesnional vector of 0/1 denoting the absence/presence of each of the 551 distinguishing terms in the article. Results in the first row are based on 5-fold cross validation, while results in the second row used 10-fold cross validation runs. Results comparing 5-fold cross validation on both representations can be found in an upcoming and more complete report of this work (Shatkay et al, 2012).

These results suggest that the image-based classifier outperforms the text-based classifier according to all measures. The differences in Recall, F-score and Accuracy are visible, as well as highly statistically significant ($p<0.0001$). The average precision is only slightly higher for the image-based classifier, although this difference is still statistically significant as well ($p<0.002$).

While the image-based classifier does show here a better performance than the text-based classifier, we note that this is not the main message we aim to convey. The results show that despite its simplicity, the image-based classifier performs at a level that is at least comparable to the one demonstrated by a text-based classifier. This relatively high level of performance suggests that our approach to image-based classification can be effective, and can aid in improving current biomedical document classification and retrieval efforts. We further discuss the results and their implications below.

Discussion and Conclusions

To summarize, there are two main aspects to the work we have presented. First, we introduced new representation of biomedical images as distributions of characters, which is based on employing OCR. Second, we have demonstrated that through the identification and the use of certain image types, (in this case DNA-rich images vs non-DNA-rich images), one can represent scientific articles both simply and effectively, in a way that supports biomedical document classification.

In terms of image-representation, the results shown in the first part of the Results section strongly support the hypothesis that OCR-based character distribution provides a simple but useful representation of images. This type of representation is particularly applicable and important in the context of biomedical publications, because so much biomedical data, including RNA, DNA and Proteins, come in the form of character sequences. Moreover, many of the images in this domain contain text in various forms for a variety of reasons – ranging from sequence data, through tags on graphs and on diagrams, to cell-labels or region- or organ-labels in fluorescence images.

We also note that unlike the typical application of OCR for obtaining words and text from images (e.g Xu et al, 2008), we propose using *distributional properties of characters* in images. As such, the method is robust to the typically noisy OCR process. Missing or mis-reading a few characters is highly unlikely to strongly affect the overall distribution of characters obtained from an image.

In terms of article-representation and classification, this work continues along the lines of our own work (Chen et al, 2006; Shatkay et al, 2006) and that of others (e.g. Rafkind et al, 2006), suggesting that defining certain types of images and automatically identifying images of these types within articles is useful both for image retrieval in-and-of itself, and, more importantly as a basis for article classification. Clearly, even within the scope of the research presented here, there is still much room left for further exploration of variants in the choice of vector representations, classifiers and even evaluation measures, which we plan to do as the next step in this work.

As we have noted earlier, the representation used for articles in the articles-classification task relied on manual tagging DNA-rich images, rather than on automated tagging by the image-classifier. Manual tagging of images was used at this stage to focus attention on the merits and shortcomings of the *article*-representation and classification, rather than on the possible issues involved in the image-classification step itself. We plan to combine the image-classifier and the article-classifier into one pipeline that will serve in the curation process for CYRENE. We are also pursuing the integration of the text- and the image-based classifiers. The application of the proposed tools to larger and more diverse datasets is another part of our planned future research.

Acknowledgements

This work was partially supported by HS's NSERC Discovery Award 298292-2009, NSERC DAS 380478-2009, CFI New Opportunities Award 10437 and Ontario's Early Researcher Award ER07-04-085, and by SI's NSF grant 0645955.

References

Brady S, and Shatkay H. 2008. *EpiLoc: a (working) text-based system for predicting protein subcellular location*. Proc. of the Pacific Symposium on Biocomputing (PSB'08), 604-615.

Chen N, Shatkay H, and Blostein D. 2006. *Exploring a new space of features for document classification: figure clustering.* Proc. of the 2006 Conference of the IBM Center for Advanced Studies on Collaborative research. (CASCON'06).

Cohen W, Kou Z, and Murphy RF. 2003. *Extracting Information from Text and Images for Location Proteomics*. Proc. of the 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD'03), 2-9.

Demner-Fushman D, Antani S, Simpson M, and Thoma GR. 2009. *Annotation and Retrieval of Clinically Relevant Images.* International Journal of Medical Informatics: Special Issue on Mining of Clinical and Biomedical Text and Data, 78(12), e59-e67.

Eppig JT, Bult CA, Kadin JA, Richardson JE, and Blake JA. 2005. *The Mouse Genome Database (MGD): From Genes to Mice — A Community Resource for Mouse Biology.* Nucleic Acids Research, 33, (Database Issue), D471-D475.

Gonzalez RC, and Woods RE. 2002. Digital Image Processing. Prentice-Hall.

Haralick RM, Shanmugam K, and Dinstein I. 1973. *Texture features for image classification*. IEEE Trans. On Systems, Man and Cybernetics, SMC-3(6), 610-621.

Hersh WR, Cohen A, Yang J, Bhuptiraju RT, Roberts P, and Hearst M. 2006. *TREC 2005 Genomics Track Overview.* Proc. of TREC 2005, NIST Special Publication. 14-25.

Istrail S, Tarpine R, Schutter K, and Aguiar D. 2010. *Practical Computational Methods for Regulatory Genomics: A cisGRN-Lexicon and cisGRN-Browser for Gene Regulatory Networks.* Methods in Molecular Biology 1, 674, Computational Biology of Transcription Factor Binding, 369 See also: *http://www.brown.edu/Research/Istrail_Lab/pages/cyrene.html*

Jain AK, and Vailaya A. 1998. *Shape-based retrieval: a case study with trademark image databases*. Pattern Recognition, 31(9), 1369-1390.

Kamachi Y, Sockanathan S, Liu Q, Breitman M, Lovell-Badge R and Kondoh H. 1995. *Involvement of SOX Proteins in Lens-Specific Activation of Crystallin Genes.* EMBO J. 14(14), 3510-19.

Krallinger M, Vazquez M, Leitner F, *et al*. 2011. *The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text.* BMC Bioinformatics, 12(Suppl 8):S3.

Murphy RF, Velliste M, Yao J, Porreca G. 2001. *Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Patterns*. Proc. of the 2nd IEEE Int. Symp. on Bio-Informatics and Biomedical Engineering (BIBE'01), 119-128.

Porter MF. 1997. *An Algorithm for Suffix Stripping (Reprint).* Readings in Information Retrieval, Morgan Kaufmann. *http://www.tartarus.org/~martin/PorterStemmer/.*

Qian Y, Murphy RF. 2008. Improved Recognition of Figures containing Fluorescence Microscope Images in Online Journal Articles using Graphical Models. Bioinformatics 24, 569-576.

Rafkind B, Lee M, Chang S, Yu H. 2006. *Exploring Text and Image Features to Classify Images in Bioscience Literature.* Proc. of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL.

Regev Y, Finkelstein-Landau M, Feldman R, Gorodetsky M, Zheng X, Levy S, Charlab R, Lawrence C, Lippert RA, Zhang Q, and Shatkay H. 2002. *Rule-Based Extraction of Experimental Evidence in the Biomedical Domain - the KDD Cup (Task 1).* SIGKDD Explorations, 4(2), 90-91.

Shatkay H, Chen N, and Blostein D. 2006. *Integrating Image Data into Biomedical Text Categorization.* Bioinformatics, 22(11), e446-e453.

Shatkay H, Narayanaswamy R, Nagaral SS, Harrington N, MV R, Somanath G, Tarpine R, Schutter K, Johnstone T, Blostein D, Istrail S. 2012. *OCR-based Image Features for Biomedical Image and Article Classification: Identifying Documents relevant to Cis-Regulatory Elements.* Proc. of the ACM Conf. on Bioinformatics and Computational Biology (BCB). (*To Appear*).

Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, and Ringwald M. 2007. *The Mouse Gene Expression Database (GXD): 2007 Update.* Nucleic Acids Res, 35, D618-D623.

Takasaki N, Kurokawa D, Nakayama R, Nakayama J, and Aizawa S. 2007. Acetylated YY1 Regulates Otx2 Expression in Anterior Neuroectoderm at two cis-sites 90 kb apart. EMBO J. 26(6), 1649-59.

Ulvila J, Arpiainen S, Pelkonen O, Aida K, Sueyoshi T, Negishi M, and Hakkola J. 2004. Regulation of Cyp2a5 Transcription in Mouse Primary Hepatocytes: Roles of Hepatocyte Nuclear Factor 4 and Nuclear Factor I. Biochem J. 381(Pt 3), 887-94.

Wang QF, Lauring J, and Schlissel MS. 2000. c-Myb Binds to a Sequence in the Proximal Region of the RAG-2 Promoter and is Essential for Promoter Activity in T-lineage Cells. Mol Cell Biol. 20(24), 9203-11. Figure 2.

Witten IH, and Frank E. 2005. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann. (Describes Weka: The Waikato Environment for Knowledge Analysis. *http://www.cs.waikato.ac.nz/ml/weka.*)

Xu S, McCusker J, and Krauthammer M. 2008. *Exploring the use of image text for biomedical literature retrieval*. Proc. of the AMIA Annu Symp, 2008, 1186.

Yu H, Liu FF, and Ramesh BP. 2010. Automatic Figure Ranking and User Interfacing for Intelligent Figure Search. PLoS One 5(10), e12983.