OCR-based Image Features for Biomedical Image and Article Classification: Identifying Documents relevant to 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
[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

ABSTRACT

Images form a significant and useful source of information in published biomedical articles, which is still under-utilized in biomedical document classification and retrieval. Much current work on biomedical image retrieval and classification employs simple, standard image features such as gray scale histograms and edge direction to represent and classify images. We have used such features as well to classify images in our early work [5], where we used image-class-tags to represent and classify articles.

In the work presented here we focus on a different literature classification task, motivated by the need to identify articles discussing cis-regulatory elements and modules in the context of understanding complex gene-networks. The curators who try to identify such articles in the vast literature 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 (like those mentioned above) can be 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 precision and recall (over 0.9). Utilizing the occurrence of such DNA-rich images within articles, we train a classifier that identifies articles pertaining to cis-regulatory elements with a similarly high precision and recall. The use of 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, the ability to automatically identify such images has much potential to be widely applicable in other important biomedical document classification tasks.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval; I.4 Image Processing and Computer Vision; I.4.10 Image Representation; I.4.7 Feature Measurement

Keywords

Bioiomedical text classification;Biomedical images; Image classification; Image features; OCR; Optical Character Recognition; cis-regulatory elements; biomedical articles; document classification; information retrieval

1. INTRODUCTION

Classifying biomedical documents based on their relevance with respect to a specific topic is a fundamental step in biomedical database curation; it is also a major component in a variety of biomedical text mining applications. One example is the process used by the Mouse Genome Informatics (MGI) resource at the Jackson labs [1]. Part of the resource includes extensive gene expression information, for which MGI's curators are tasked with identifying published literature containing information about gene expression in the mouse [2]. Doing this requires, as a first step, to obtain all and only articles describing experiments relevant to gene expression in the mouse. The articles are then read, and pertinent information is extracted and curated. Another example is the identification of articles that may contain experimental evidence for protein-protein interaction. Automating the latter task was part of the challenge posed in BioCreative III [4].

Images shown within articles form a rich source of information, and provide significant cues to curators when deciding the relevance of an article to certain biological domains. We are interested in using both images and text to classify biomedical articles, as we have shown in an earlier work [5].

Much research has been done during the past decade on image categorization and content-based retrieval, both within and outside the biomedical domain [5]. Most of the work is concerned with contents-based categorization and retrieval of images (not of documents). To do so, a corpus of images is defined (for testing and training), 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 [17], Haralick's texture-features [18], and values from edge direction histograms [19]. We have used such features as well to classify images in our early work [5] where we used image-class-tags to represent and classify documents.

In the work presented here we focus on a specific and different literature classification task, motivated by the need to identify articles discussing cis-regulatory elements and modules in the context of understanding complex gene-networks. The group working on the *CYRENE cis-regulatory browser project* at Brown University [20,21] noted that to identify such articles in the vast literature, one can use as a major cue a certain type of image showing the DNA and denoting the conserved cis-regulatory elements. An example of such a diagram is shown in Figure 1. We refer to images that show DNA content as *DNA-rich* images.

Automatically identifying such images using common image features (like those mentioned above) proves highly error prone, as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACM-BCB'12, October 7–10, 2012, Orlando, FL, USA.
Copyright © 2012 978-1-4503-1670-5/12/10…$15.00.

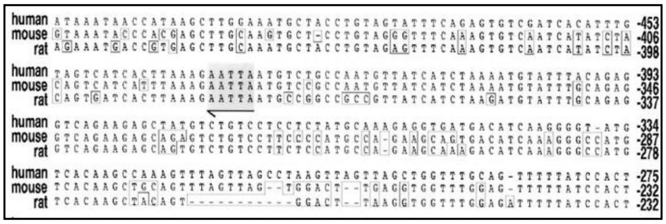
human ATAAATAACCATAAGCTTGGAAATGCTACCTGTAGTATTTCAGAGTGTCGATCACATTTG -453
mouse GTAAATACCCACGAGCTTGCAAGTGCT-CCTGTAGGGTTTCAAAGTGTCAATCATATCTA -406
rat AGAAATGACCGTGAGCTTGCAAATGCTACCTGTAGAGTTTCAAAGTGTCAATCATATCTA -398
human TAGTCATCACTTAAAGAATTAATGTCTGCCAATGTTATCATCTAAAATGTATTTACAGAG -393
mouse CAGTCATCATTTAAAGAATTAATGTCCGCCAATGTTATCATCTAAAATGTATTTGCAGAG -346
rat CAGTGATCACTTAAAGAATTAATGCCGGCCGCCGTTATCATCTAAGATGTATTTGCAGAG -337
human GTCAGAAGAGCTATGTCTGTCCTCCTCTATGCAAAGAGGTGATGACATCAAGGGGT-ATG -334
mouse GTCAGAAGAGCAGAGTCTGTCCTTCCCCATGCCA-GAAGCAGTGACATCAAAGGGCCATG -287
rat GTCAGAAGAGCAGTGTCTGTCCTTCTCCATGCCA-GAAGCAAAGACATCAAAGGGCCATG -278
human TCACAAGCCAAAGTTTAGTTAGCCTAAGTTAGTTAGCTGGTTTGCAG-TTTTTATCCACT -275
mouse TCACAAGCTGCAGTTTAGTTAG--TGGACT--TGAGGTGGTTTGGAG-TTTTTATCCACT -232
rat TCACAAGCTACAGT-----------GGACT--TAAGGTGGTTTGGAGATTTTTATCCACT -232

Figure 1 An example of a DNA-rich diagram of the type that is over-represented in articles discussing cis-regulatory elements. Taken, with permission, from PMID 12972592, Figure 2. [22]

our experiments show. 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 elements or modules with high precision and recall.

While this paper focuses on the specific task of identifying cis-regulatory-related publications, the use of OCR as image features has much potential beyond the current task, to identify other types of biomedical sequence-based images, and 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. Section 2 provides a brief survey on image analysis in biomedical documents highlighting the difference between previous work and the research presented here. Section 3 provides more information about the specific problem we are addressing in the context of the CYRENE project. Section 4 discusses the datasets and the methods we use to process and to represent images and articles. Section 5 presents experiments and results, while Section 6 concludes and outlines future work.

2. BACKGROUND and RELATED WORK

Among the earliest work on using images within biomedical articles is the research by Murphy et al. [6,7,8,9], which uses image categorization for identifying images and articles discussing protein subcellular localization. They provide an extensive in-depth investigation of a specific task: identifying and interpreting a certain type of microscopy image, characteristic of localization experiments. In their image processing they use standard image-features like the ones mentioned above. Notably their tools are centered around the protein-subcellular-localization task, and not around biomedical text/image retrieval as a whole. Work by Rafkind et al. [10] explored retrieval of biomedical images from the literature in a more general context, while work by Shatkay et al. [5] 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 area that focuses on image processing is content-based retrieval of medical images and medical documents. In this type of retrieval one may look, for instance, for *all x-ray images of a fractured wrist*, or for all documents that contain such images. The shared tasks of ImageClef [11] in the past few years have included challenges of this type, and lead to quite a few systems addressing such challenges [e.g. 12]. Again, typically standard image features like the ones mentioned earlier (texture features, gray-scale-based features etc.) are used to represent the images.

Taking advantage of text that is associated with images for document retrieval or for identifying relevant images typically involved using the text of the figures caption (an idea introduced by Regev et al. [23]) or possibly also the text referencing the image from within the article's body [13].

Last, as a way to improve indexing and retrieval of biomedical images, Xu et al. [15] and later Rodriguez-Estaban and Iossifov [[16] proposed to use *optical character recognition (OCR)* to extract text from within biomedical images, using the extracted words/terms to index images [15] or help classify them [[16]. Notably, in contrast to the work presented here, that research viewed OCR as a way to extract text-words associated with images, rather than as an independent source of useful, distributional image-features. This latter idea, which to the best of our knowledge was not pursued before, is the focus of our work as presented here.

3. The Article Classification Task for CYRENE

The CYRENE project [20,21] 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 other regulatory genes. (Primarily in human, mouse, fruit fly, sea urchin and nematode, with some information pertaining to rat, chicken, and zebrafish). To be included in the lexicon, a regulatory mechanism must adhere to the 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 into the database. This paper is concerned with 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 [20], (of which RT,KS and SI are a part), the publications in which the most relevant information is typically found often contain certain types of diagrams and graphs (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, particularly marking the conserved regions, the motifs or the sites 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 [22].

The classification task is thus to identify publications, among a set of candidates already containing basic terms such as "regulation" or published in the relevant journals (such as *Molecular and Cellular Biology*), the documents that are most likely to contain experimentally validated information about cis-regulatory elements and modules. We pursue this task by using both a text-based classifier (briefly mentioned here), and an image-based document classifier, the latter is the main focus of this paper. The data and the methods used for training and testing such a classifier are discussed in the next section.

4. DATA and METHODS

4.1 The Dataset of CYRENE-related Articles

For the purpose of this work, the CYRENE team of curators has initially identified a set of 271 publications as high-quality articles 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. About 90 of the relevant articles came from the first two journals, and additional 80 came from the next five in the above list. A keyword search (using keywords such as *regulatory, transcription, DNA element, DNA motif*) was applied to the many thousands of resulting articles, to further reduce the set to those articles likely to discuss gene regulation. The resulting set (several thousand articles) was examined by the curators to identify the high-quality articles, namely ones that experimentally validate cis-regulatory modules, forming the set of 271 articles. The latter is the *positive set* or the set of *Relevant* articles for our classification training/testing process.

Many of the remaining published articles were rejected from the CYRENE-relevant dataset without further tracing. A small subset of those irrelevant publications, consisting of 78 articles, were identified and kept by the curators, and were provided as a *negative* set of *Irrelevant* articles. As the resulting overall set is highly unbalanced as far as classification goes, (271 positive examples and 78 negative ones), we selected an additional set of 143 negative examples from the Journal of Molecular Cellular Biology – which is a journal in which close to 20% of the 271 relevant articles were found. The negative documents were selected by going into the same volumes from which relevant articles were obtained, and obtaining 10-20 articles from the same volume that were not judged to be relevant by the curators. By selecting irrelevant articles from the same volume from which relevant articles were selected we ensure that the general discourse and style of writing remains consistent across the relevant and the irrelevant articles. That is, there is no shift in time and in the overall areas of current interest between the corpus of relevant articles and the corpus of irrelevant articles. Such a shift, if existed, would over-simplify the learning task of separating between the relevant and irrelevant sets, as separation could then rely on differences in language and style, as opposed to on the difference in actual contents.

The resulting final set 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 describe how the documents are used for testing and training an image-based document-classifier in Sec. 4.3.

4.2 Images, Image Panels, Representation and Classification

It has been noted by multiple groups [6,5,13] that figures in biomedical publications often consist of multiple subfigures or *panels,* as shown in Figure 2. Each panel is typically an individual image, and as such, when considering images, we would like to separate figures into individual panels.

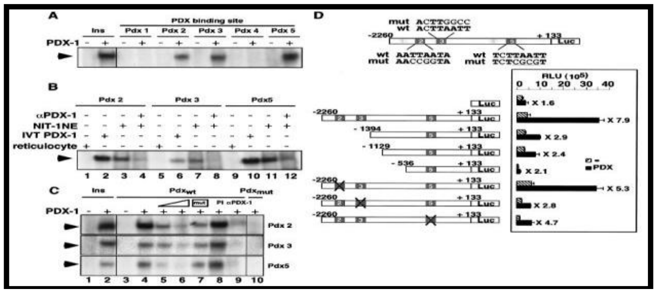

Figure 2 An Example of a composite figure, consisting of multiple image panels. Taken, with permission, from PMID 12972592, Fig. 3. [22]

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 [24]. A full description of this tool and its features is beyond the scope of this paper and will be published separately.

As we noted earlier, image panels containing DNA information, like the one shown in Figure 1, are typically over represented in articles that discuss cis-regulatory modules. As such, we hypothesized that the ability to identify such images automatically, and to identify articles that show an over-abundance of such images would prove helpful in identifying relevant documents for CYRENE database. As before, we refer to this type of image panel, which shows DNA regions, as *DNA-rich image panel*. In order to automatically identify such image panels, we would like to train a classifier that can perform this task, i.e. would distinguish between DNA-rich images and all other images. To achieve this goal we need:

a) To obtain a set positive image panels that contain DNA sequences and a set of negative images panels, which do not contain DNA sequences; and

b) To represent images using a set of features that would 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 the first sub-goal (**a**) above, we identified a set of 88 DNA-rich image panels, and 100 image panels that do not show DNA sequence (although they may show other sequences, such as proteins or RNA). This set of 188 panels is the one to be used for training and testing a classifier that would aim to distinguish between DNA-rich and non-DNA-rich images.

In order to represent images as feature-vectors, so that the panel-classification task could be attempted (aim **b** above), we introduce a novel *OCR-based representation*. To do so, we apply an *optical character recognition* (OCR) tool, ABBYY Finereader [25], 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). Panel A 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, panel B 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, I, and L, 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.

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 maintaining the frequencies of A, C, G, and T. As our results show, 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 experimented with learning a classifier under this representation, as further discussed in Section 5.

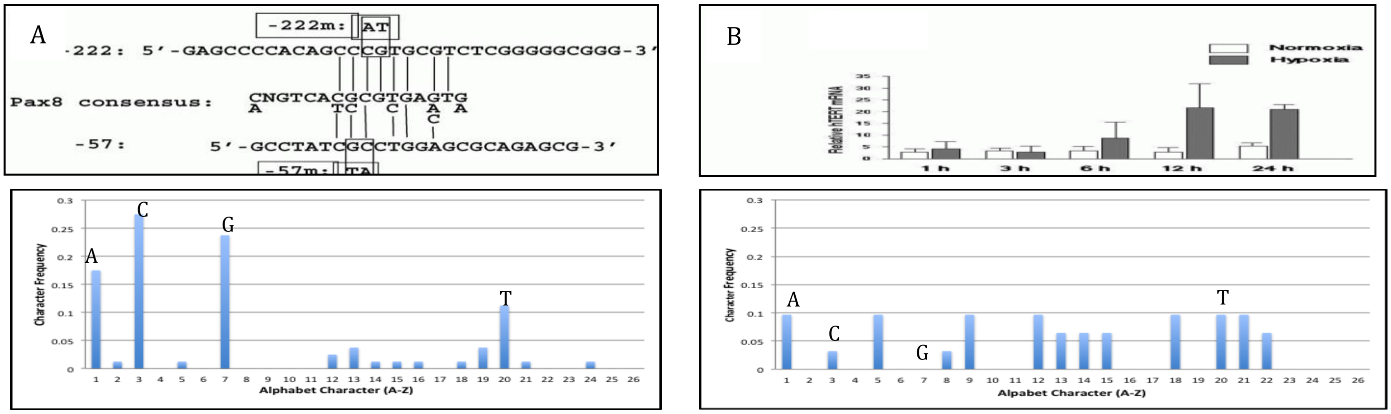

Figure 3 An example of two panels, A ([26], Fig. 5b) and B ([27], Fig. 2b); obtained with permission. The respective character frequency distribution (shown only for the letters a-z) is provided below each image. Panel A shows a DNA-rich image, which translates to peaks on A, C, G and T in the character distribution, while panel B does not.

Each of the 188 image panels is represented as such a vector (under either 5-dimensional or 37-dimensional) representation. To train and test classifiers using these representations, we use the standard WEKA tools [28] to train and test a decision-tree classifier (the J48 algorithm). Further details regarding these experiments are provided in Section 5.

A summary of the datasets discussed above and their respective sizes is shown in Table 1.

4.3 Articles Representation and Classification

While the above paragraphs discussed the representation and the classification of image panels, recall that our ultimate goal is to classify published articles based on their relevance (or there lack-of) to the CYRENE dataset. The dataset of articles we used consists of 271 positive (relevant) examples, and 221 negative (irrelevant) examples, where we have obtained the full PDF text files for 264 positive and 220 negative articles respectively, as discussed in Section 4.1.

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 Section 4.2, 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 Section 6.

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:

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

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 [28].

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 was done before [29], 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 still relatively large number of features involved in such a representation (about 550 terms per vector), we use the WEKA naïve Bayes rather than decision tree, to train/test a classifier from the text representation of articles.

Table 1. Summary of the datasets used for training/testing classifiers. *Positives* are items that satisfy the Dataset condition listed on the left, while *Negatives* are items that do not satisfy the Dataset condition listed on the left.

Dataset	# Positives	# Negatives	Total
CYRENE-related articles	271	221	492
CYRENE-related articles with full-text PDF	264	220	484
DNA-rich panels	88	100	188

5. EXPERIMENTS and RESULTS

5.1 Experimental Setting

Notably, there are two main hypotheses we are examining. The first is whether the OCR-based representation discussed above is indeed an effective representation for automatically distinguishing between DNA-rich image panels and non-DNA-rich panels in articles. The second is whether the proportion of DNA-rich panels within an article can be used as an effective indicator for assessing the article's relevance to the CYRENE dataset.

Accordingly we perform two sets of experiments. The first is concerned with image panel classification using OCR-based representation of image panels. The second is concerned with article classification, using image-based representation of articles. These experiments are described in Sections 5.1.1 and 5.1.2 respectively.

5.1.1 Image-Panel Classification using OCR-based Representation

To evaluate the effectiveness of the OCR-based representation for supporting an automated distinction between DNA-rich and non-DNA-rich image panels, we use the 188 image panels that were manually annotated for this purpose (as discussed in Section 4.2). For each of these image panels we construct three different representations, as follows:

1) A *37-dimentional* feature vector $< w_1^p \ \dots w_{37}^p >$, where the weight in each of the first 36 positions corresponds to the relative abundance of each of the 36 characters (A-Z[1], 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:

$$w_i^p = \frac{\boldsymbol{\#\ of\ times\ character\ c_i\ occurs\ in\ \ panel\ p}}{Total\,\#\,of\ \ \boldsymbol{character\ occurrences\ in\ \ panel\ p}}\ .$$

(See Section 4.2 for further detail and Figure 3 for an example).

2) A *5-dimentional* feature vector $< w_1^p \ \dots w_5^p >$, 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) A simple gray-scale histogram representation. That is a 256-dimensional vector $< w_1^p \ \dots w_{256}^p >$, 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 the WEKA [28] standard tools to train and test a decision tree classifier, using stratified 5-fold cross validation. Under this setting 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 and 1/5 is left out for testing. The process is iterated 5 times with a different 1/5 of the data being left out at each iteration. To ensure stability of the results, we use five separate complete runs of 5-fold-cross-validation for each of the representations (a total of 25 runs per representation).

5.1.2 Article Classification using image-based Representation

To evaluate the effectiveness of the image-based representation for supporting an automated distinction between CYRENE-relevant and non-CYRENE-relevant publications, we represent the 484 pre-classified articles (264 CYRENE-related, 220 non-CYRENE-related, as discussed in Section 4.1) using the simple 2-dimensional representation described by Eq. 1 in Section 4.1.

We again use the WEKA standard 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. As before, we use five separate runs of 5-fold cross validation to ensure stability of the results.

As a point of comparison, we also use a text-based representation of the articles, employing the bag-of-words model of text documents, which is commonly used in information retrieval and document classification applications. The text we use to represent each article is taken only from its title and abstract, rather than the full PDF. This is done for three reasons: 1) The use of full-text leads to very large representations that are both slower to work with and typically lead to sub-optimal results in terms of classification accuracy or clustering coherence. 2) While studies on biomedical *information extraction*, e.g. identifying protein or gene mentions in the literature, suggested that using full-text rather than abstracts allows an application to identify more instances to extract, *no similar study* suggests that document classification improves when using larger full-text documents. Our own experience in another curation-related task [30] supports the notion that text from title-and-abstract fits well for this type of document-classification application. 3) Full-text versions of the articles are not available in ASCII – only in PDF. Converting from PDF to ASCII text is often error-prone, thus introducing noise as an additional factor to consider in a comparative study. This problem does not arise when using titles and abstracts, as they are readily available as ASCII text.

The titles and the abstracts of all 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 [31] 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 documents) 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. Statistical significance of the difference is determined using the Z-score test, as described in our earlier work [29].

The resulting vocabulary of 551 terms is used to represent each article *d* as a 551-dimensional vector of binary values, $< w_1^d \ \dots w_{551}^d >$, where w_i^d =1 if the i^{th} term, t_i, occurs in document *d, i.e.* t_i Î d, and w_i^d =0 otherwise.

As this is a relatively high-dimensional representation, we use the naïve Bayes classifier in the WEKA tools, employing again 5-fold cross validation to train and test the classifier.

5.1.3 Evaluation Measures

To assess the performance of all the classifiers described above, we use the standard measures widely used for classification evaluation, namely: *Precision*, *Recall*, *F-measure,* and overall accuracy (*Acc*) as defined below:

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

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

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 for the panel-classification task, while it is a CYRENE-relevant article for the article classification task.

[1] 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.

5.2 Results

5.2.1 Image-Panel Classification Results

Table 2 summarizes the average results obtained from running *five* separate panel-classification runs of stratified 5-fold cross validation, under each of the three image-panel representation we have used, as described in Section 5.1.1. 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 letter 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 average precision for the top two rows is above 0.9 while the average recall is about 0.9 in both cases. While the second row shows slightly higher values than the first, the differences in performance between the two representations 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 difference in performance with respect to the top two rows is also extremely statistically significant ($p<0.0001$, using the two-sample t-test).

Table 2. Image-panel classification results, 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

5.2.2 Article Classification Results

Table 3 summarizes the average results obtained from running *five* separate article-classification runs of stratified 5-fold 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.

Table 3. Article classification results, averaged over 5 independent runs of 5-fold cross validation. 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)	0.82 (.057)	0.82 (.061)	0.80 (.043)	0.82

On the whole, according to all performance measures, the image-based classifier outperforms the text-based classifier. The differences in Precision, Recall, F-score and Accuracy are visible, and are also highly statistically significant ($p<0.0001$, using the two-sample t-test).

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 this study aims 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 in the next section.

6. DISCUSSION and CONCLUSIONS

The work we described here presents two main contributions. First, we introduced a new method, based on OCR, to represent biomedical images as distributions of characters. Second, we have demonstrated that through the identification and the use of image types, (in this case DNA-rich images vs non-DNA-rich images), one can represent articles quite simply and effectively in support of biomedical document classification.

In terms of image-representation, the results shown in Section 5.2.1 strongly support the notion that OCR-based character distribution provides a very useful - yet simple - representation of images. The proposed approach is particularly suitable, applicable and significant in the context of biomedical publications, because so much of the data has the form of character sequences (RNA, DNA and proteins – which are readily distinguishable from other images based on character distribution), and so many of the images contain text for a variety of reasons ranging from organ- or cell-labels in fluorescence images, through DNA sequences, to tags and marks on graphs and diagrams.

Moreover, by using the distributional properties of characters in the image – as opposed for instance to extracting complete words from it (which was done by others before [15][16]) – the method is robust to the typically noisy OCR process. Missing or mis-reading a few characters in an image is very unlikely to have a strong impact on the overall distribution of characters obtained from the image.

In terms of article-representation and classification, this work continues along the lines of our own work [5,14] and that by others [e.g. 10], suggesting that defining types of images and being able to automatically identify images of certain types within articles is useful not only for image retrieval in-and-of itself, but also as a basis for document classification.

The methods and the results presented here will benefit from further exploration of the possible variants in the specific 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 in Section 4.3, the image-based article representation, used for the article-classification task presented here, relied on the manual tagging of the DNA-rich images, rather than on automated tagging by the image-classifier. We used manual tagging of images to ensure that we indeed focus in that part of the work on the merits and shortcomings of the *article*-representation and classification, rather than on the possible issues involved in the image-classification step. Therefore, another important direction to be pursued in the immediate future is that of assembling the image-classifier and the article-classifier into a single 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.

ACKNOWLEDGMENTS

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

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] Cohen W, Kou Z, 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.

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

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

[10] 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.

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

[12] 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.

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

[14] Chen N, Shatkay H, 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).

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

[16] Rodriguez-Esteban R and Iossifov I. 2009. *Figure Mining for Biomedical research.* Bioinformatics, 25(16), 2082-2084.

[17] Gonzalez RC, Woods RE. 2002. Digital Image Processing. Prentice-Hall.

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

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

[20] 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, 369-399.

[21] CYRENE=*http://www.brown.edu/Research/Istrail_Lab/pages/cyrene.html*

[22] Annicotte JS, Fayard E, Swift GH *et al*. 2003. *Pancreatic-duodenal homeobox 1 regulates expression of liver receptor homolog 1 during pancreas development*. Mol Cell Biol, 23(19), 6713–6724. PMID: 12972592

[23] Regev Y, *et al*. 2002. *Rule-Based Extraction of Experimental Evidence in the Biomedical Domain - the KDD Cup (Task 1).* SIGKDD Explorations, 4(2), 90-91.

[24] Xerox Rossinante= *https://pdf2epub.services.open.xerox.com/*

[25] ABBYY Finereader for OCR. The website is at http://finereader.abbyy.com/

[26] Puppina C,Ivan Prestab I, D'Elia AV *et al*. 2004. Functional interaction among thyroid-specific transcription factors: Pax8 regulates the activity of Hex promoter. Mol Cell Endocrinol, 224(1-2), 117–125. PMID: 15062550

[27] Nishi H, Nakada T, Kyo S *et al*. 2004. *Hypoxia-inducible factor 1 mediates upregulation of telomerase (hTERT).* Mol Cell Biol, 24(13), 6076–6083. PMID: 15199161

[28] Witten IH, 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*.)

[29] Brady S, 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.

[30] Denroche R, Madupu R, Yooseph S, Sutton G, Shatkay H. 2010. *Toward Computer-Assisted Text Curation: Classification Is Easy (Choosing Training Data Can Be Hard...)* Linking Literature, Information, and Knowledge for Biology, Lecture Notes in Computer Science, 6004, 33-42.

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