Proceedings of the 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001), pp. 119-128

Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Patterns

Robert F. Murphy, Meel Velliste, Jie Yao and Gregory Porreca
Department of Biological Sciences, Biomedical and Health Engineering Program, and Center for Automated Learning and Discovery, Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA 15213/U.S.A. Phone: 1.412.268.3480. FAX: 1.412.268.6571. Email: murphy@cmu.edu.

Abstract

There is extensive interest in automating the collection, organization and analysis of biological data. Data in the form of images present special challenges for such efforts. Since fluorescence microscope images are a primary source of information about the location of proteins within cells, we have set as a long-term goal the building of a knowledge base system that can interpret such images in online journals. To this end, we first developed a robot that searches online journals and finds fluorescence microscope images of individual cells. We then characterized the applicability of pattern classification methods we have previously used on images obtained under controlled conditions to images from different sources and to images subjected to manipulations commonly performed during publication. The results indicate the feasibility of developing search engines to find fluorescence microscope images depicting particular subcellular patterns.

Introduction

The dramatic increase in biological knowledge (especially with respect to the sequences and structures of genes and proteins) over the past 20 years, combined with advances in computer technology, has led to the creation of a number of biological databases. These include databases that focus on a particular type of information (e.g., sequences, structures, genome maps) for all organisms, as well as those that combine various types of information for a single organism. The information in these databases is largely incorporated by computer-generated links to relevant entries in other structured databases or entered by hand by scientists in the relevant fields. Such curated databases do not typically capture the supporting evidence for each entry and usually do not allow for uncertainty, alternative views or conflicting evidence. There has therefore been interest recently in the creation of self-populating knowledge bases that can extract and store assertions from published literature in an automated fashion. Such knowledge bases can serve not only as resources for practicing biologists but also as input for systems that can generate new hypotheses using data mining methods. They can also serve as a starting point for modeling and simulation systems that incorporate information beyond that in current structured databases.

A significant part of this task involves information extraction from unstructured text, a burgeoning area of computer science research. Recent work in this area relevant to molecular biology includes a supervised learning approach to extraction of statements regarding protein location from MEDLINE abstracts [1] and the EDGAR system for extracting assertions about drugs and genes related to cancer from the biomedical literature [2]. However, much important information in biological literature is in the form not of text but of figures, and little published work has been done on automating the extraction of information from them. We report here initial work aimed at accomplishing this task for a subset of such figures, fluorescence microscope images.

The starting point for this work is our previous development of numerical features that can describe complex subcellular patterns in such images and neural network classifiers that are capable of recognizing all major subcellular structures in at least one cell type [3-6]. For this purpose, we have developed **S**ubcellular **L**ocation **F**eatures (SLF) that utilize geometric moments, texture measures, and morphological image processing [5, 6].

The work described below addresses two distinct tasks. The first is to extract all figures from articles in on-line journals and to identify those that depict fluorescence microscope images. The second is to identify numerical features that adequately capture information about subcellular location in such images without being inappropriately sensitive to variation in image resolution and any manipulations that the image may have been subjected to during publication.

Collecting Images from Online Journals

On-line Article Downloading and Figure Extraction.

As a first step, we implemented a web robot that allows users to automatically retrieve online journal

articles that may have relevant images. The robot utilizes the search engine of "PubMed" (http://www.pubmed.gov/entrez) to find articles matching a PubMed search string and downloads as PDF files the subset of those articles locally contained in "PubMed Central" (up to a maximum number of articles set by the user). For illustration, we have used the query "Golgi" to search for images that depict the subcellular patterns of Golgi proteins and limited the number of PDF files returned to 100.

Once the PDF files matching the query were downloaded, we extracted all figures and captions using the "PDFtoHTML" tool (http://www.ra.informatik.unistuttgart.de/~gosho/pdftohtml), which we had modified to add the capability of associating figure images with their corresponding captions. Due to the different layouts of PDF files, the tool is not perfect. Some figures could not be correctly extracted and some extracted images did not have their corresponding captions. For the "Golgi" query, visual inspection of the downloaded articles indicated that there were 745 actual figures contained in the 100 articles (each with a caption). The program extracted 581 of what it considered to be figure-caption pairs, but upon visual assessment only 571 of these proved to be correctly matched figure-caption pairs. From this test data, the precision of the extraction process (defined as number of correct predictions divided by number of total predictions) was 98% while the recall (defined as the number of correct predictions divided by the number of actual figure-caption pairs that could have been found) was 77%.

Boundary Detection and Recursive Panel Splitting

Having successfully extracted figures with high precision the next step was to divide the figure images into their constituent components. Nearly all extracted figures contained several panels, since similar conditions are often presented together so that they can be compared and contrasted. More complicated multiple panel figures were also frequently encountered, containing combinations of microscope images, gel pictures, and charts. Therefore, it was necessary to split each figure into its composite panels before we could attempt to display or interpret any particular panel. For this purpose, we utilized a recursive panel splitting method based on boundary detection.

Although figures may have different layouts, they have some properties in common. Most panels differ in intensity from their boundaries. Thus, panels that are largely black (such as fluorescence micrographs) typically have a white boundary and panels that are largely white often have black boundaries. The boundaries consist almost always of straight lines. Accordingly, we used a simple "projection method" for boundary detection.

For each figure, the modal pixel value of the entire image was calculated. If this value was above 127 (the modal pixel value is closer to white than black), the figure was skipped as likely not containing any microscope images. For the remaining figures, largely white outside boundaries were removed by successively deleting rows or columns on the edge of the image when their average pixel value was greater than 180 (all images were retrieved in 8-bit JPEG format so the maximum pixel value was 255).

Projections were then calculated by summing the pixel values of each figure along horizontal and vertical directions. As illustrated in Figure 1, the positions where there are steep changes in the projection correspond to boundaries. To find accurate panel boundaries, we only need to detect the sharply changing positions in the projections, or more conveniently, to find peaks in the differentiated projections.

For figures with complex panel structures like that in Figure 1, some boundaries do not cross the entire image. Therefore, simply cutting the figure through all boundaries will generate more pieces than appropriate. We therefore adopted a recursive approach in which the figure is continuously split at the highest peak position in either the horizontal or vertical direction, the projections recalculated for the pieces, and the process repeated until the highest peak is below a predetermined threshold value.

To evaluate the performance of the panel splitting process, we used the first 100 figures of the 581 figures resulting from the "Golgi" query. The splitting routine yielded 281 panels. Each split panel was displayed alongside the corresponding journal figure and identified visually as either correctly or incorrectly split (we also determined by visual inspection that there were a total of 344 panels). Of the 281 panel images, 205 were considered to be correct single panel images and the remainder were either badly split panels or remnants of edges. This corresponds to a precision of 73% and a recall of 60%.

Identifying Fluorescence Microscope Images

While figures in journals contain various types of images (several kinds of microscope images, charts, gel images etc.), our goal was to collect and analyze fluorescence microscope images. Having split the figure images into their constituent pieces or "panels", the next task was therefore to identify panels containing fluorescence microscope images. We used a k-nearest neighbor (kNN) classifier for this task. For each image, a histogram of pixel intensities was constructed with 64 equally-spaced bins ranging from the minimum to the maximum pixel intensity in that image. For classification, each image was represented by a vector of the frequencies in each of these bins. A threshold value T $(0<T<k)$ was introduced to the kNN algorithm to explore the trade-off

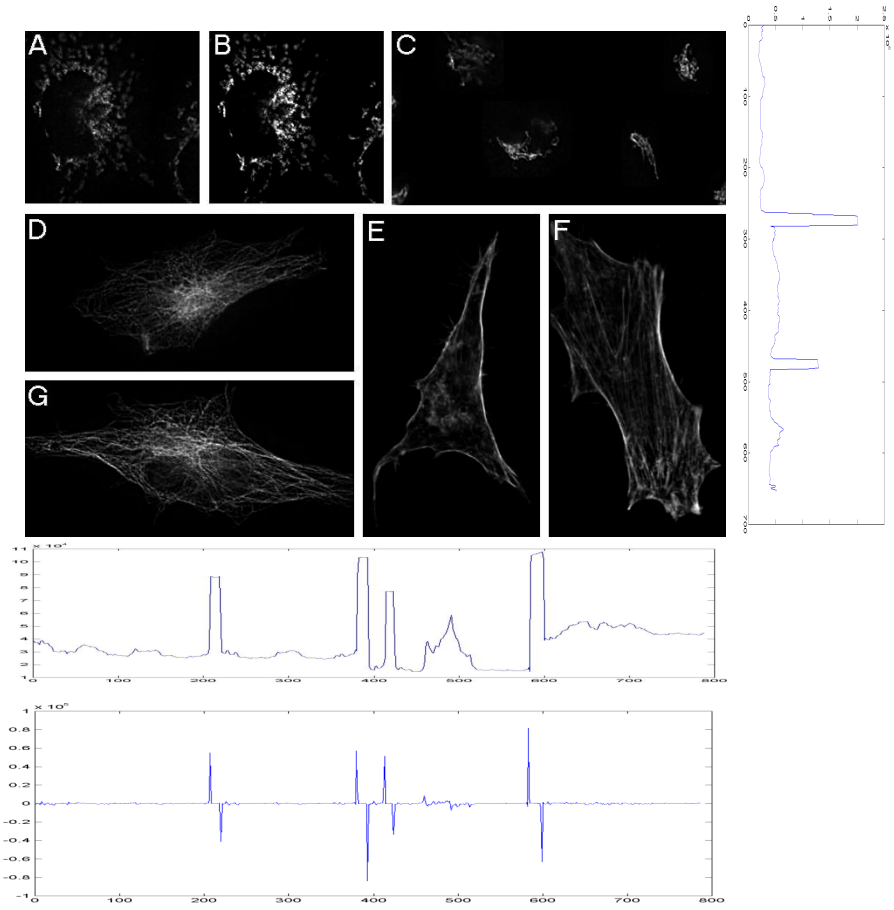

Figure 1. Detecting panel boundaries with projections and recursive splitting. An example multiple panel figure is shown to illustrate the splitting process. The average pixel value across the figure is shown for each position along the horizontal and vertical directions, along with the derivative along the vertical direction. These values were used to recursively separate the figure, first horizontally between panels ABC and DEF, then vertically between panels A and B, B and C, DG and E, and between E and F, and finally again horizontally between panels D and G.

between precision and recall. A test image was considered to be a fluorescence microscope image only if the number of its *k* neighbors (from the training images) that were fluorescence microscope image was larger than *T*.

We applied the leave-one-out cross validation method on a set of pre-labeled 1586 single panel images (788 fluorescence and 798 non-fluorescence). Values of *k* from 3 to 21 were tested, with the best performance occurring for a *k* value of 9. For this value, a recall of 70% was achieved with 100% precision and a precision of 97% was achieved for a recall of 92% (for different *T* values).

We also checked the performance (for *k*=11 and *T*=5) on the first 100 panels from the Golgi query. Of these, 42 actually contained fluorescence microscope images and 38 of these were reported as such. This corresponds to a precision of 100% and a recall of 90%, confirming our *a priori* expectation that identification of fluorescence microscope images is relatively straightforward.

Removing Internal Image Annotations

One of the problems with microscope images from online journals is that nearly all of them contain annotations such as panel labels and arrows. Such annotations can be expected to confuse the pattern analysis algorithms. To remedy this problem, we implemented an algorithm that detects annotations and removes them by filling the area corresponding to the annotation with background pixel values.

The detection relies on finding areas that are bright and have sharp edges. "Bright" was defined as a pixel value greater than or equal to 200 (generating binary image A). Edges were detected using a 3x3 sharpening kernel with 8 in the center and negative one elsewhere.

The sharpened image was thresholded at 200 (yielding binary image B). Image C was generated as the

intersection of Images A and B, thus showing those regions that were bright and had sharp edges.

Image C contained the edges of the annotations as well as some noise. We noticed the noise consisted of short line segments while annotation edges were represented by longer continuous regions. While in principle it may have been possible to remove the noise by simply removing objects consisting of less than a certain number of pixels, that approach would also have removed some annotations because the line segments making up the edges of some annotations (such as a letter A) were sometimes disjoint. We therefore used a three-stage process where we first removed line segments shorter than 4 pixels to eliminate some of the noise. We then closed the binary image using a 4x4 pixel structural element to connect the disjoint sections making up the edges of the annotations. Finally we removed any objects of size 25 pixels or less to delete any remaining noise. This resulted in binary image D which contained simply the edges of annotations in the form of closed boundaries.

In order to eliminate the annotations from the original image it was necessary to have an image of the full region of the annotations rather than just the boundaries. Image E was created by filling image D.

Finally, to eliminate the annotations an intuitive approach might have been to simply set the pixels in the original image corresponding to the "on" pixels in image E to zero (or some other fixed background value). However, we found that it was impossible to find any satisfactory fixed value because while most of the time the annotations were on a dark background, sometimes the background in the immediate area of the annotation was fairly high. We therefore replaced annotations with the most common pixel value in the immediate neighborhood (within three pixels) of that annotation.

To evaluate the annotation removal step, we examined the first 100 reported fluorescence microscope panels from the previous step. Of these, 71 actually contained annotations. The system recognized 70 of these as requiring annotation removal, of which 47 had all annotations correctly removed and an additional 11 had nearly all removed. This corresponds to a precision of 83% and a recall of 82%.

Automated Segmentation of Multi-Cell Images

Since many (if not most) published fluorescence microscope images contain more than one cell (and our methods for classifying location patterns require images of a single cell), our next step was to segment the multi-cell images into single cell images.

At first the gray-level images were transformed to binary images by an automated threshold method. The resulting binary images contain objects which correspond to the cells. The boundary between the cells was generated by the “seeded watershed” algorithm of SDC Morphology Toolbox for Matlab (SDC Information Systems, Naperville, IL).

To evaluate the segmentation method, we utilized an area-based difference measure weighted by image intensity. We generated reference images using manual segmentation and manual seeding. The correctness of an automatically generated region was judged by calculating the percent of fluorescence in the area of overlap between the automated and manual methods. The automated method was considered successful for a cell if 80% of the fluorescence in the manually generated region overlapped with the automatically generated region and if 80% of the fluorescence in the automatically generated region overlapped the manual one.

For the images returned by the “Golgi” query, we chose a random subset of 100 panels from the panels classified as fluorescence microscope images in the previous stage. Of this 100, 36 were ignored as being badly split panels. The remaining 64 panels contained 291 individual cells (cells touching the edge were ignored). The automated segmentation method generated 149 single cell images, of which 93 were considered to be correct using the 80%/80% overlap criterion. This corresponds to 62% precision but only 32% recall. When a similar analysis was done for a "Tubulin" query, the precision was 52% and the recall was 41%. The results indicate the difficulty of developing a general cell segmentation method, and further work will be needed.

Interpreting Fluorescence Microscope Images

The results shown so far demonstrate the feasibility of finding fluorescence microscope images of individual cells in online journals with reasonable precision. The next task is to extract appropriate information from the collected images so that statements can be generated about their meanings. An example of this task is to classify patterns of subcellular distribution, as we have done for images collected under controlled circumstances. There are a number of challenges in doing this for images from online journals.

1) The first challenge comes from differences in the magnification and pixel resolution at which the image was collected.
2) The second is coping with differences in sample preparation, cell type, and microscopy method (wide field, deconvolution or confocal).
3) The third arises from alterations in the images introduced during the publication process, which include image compression, resampling, and intensity transformations.

The following three sections describe our initial efforts to address each of these issues.

Compensating for Differences in Pixel Resolution

The pixel resolution (micrometers per pixel) is one of the most determining factors in how a cell image will be interpreted by an automated analysis method such as a classifier. Out of our previously developed SLF features [6] there are four that are directly dependent on the scale of the image. Luckily, microscope images published in journals often have scale information included in the figures. The images usually contain a scale bar and the respective caption contains the number of micrometers corresponding to the scale bar. Finding the pixel resolution therefore was a matter of locating the scale bar in each image, measuring its length in pixels and extracting the appropriate number from the caption text.

To locate the scale bar, each image was first thresholded using a high threshold (90% of maximum pixel value). Object finding was then used to locate continuous regions of above-threshold pixels. Each object found was then considered as a candidate scale bar. The candidates were filtered using a few simple rules to constrain size and shape. Since a scale bar is almost always a horizontally oblong rectangular bar, the size of the candidates was constrained as follows: width/height ratio of 3 or more, total width not greater than 90% of image width and not less than 5% of image width. An additional constraint was that the object pixels had to cover at least 80% of the area of the bounding rectangle of the object. This was to discard candidates that did not have a nearly perfect rectangular shape even though they met size constraints. Finally, after applying all of the above constraints there was usually either one candidate left (the actual scale bar) or none (frequently only one panel per figure has a scale bar). If more than one candidate was left, then the longest one was chosen, based on the observation that there can be only one scale bar per panel and that erroneous bars were generally smaller than the real scale bar. This resulted in obtaining the pixel width of the bar, which then had to be correlated with the width of the bar in micrometers.

Whenever there is a scale bar in the figure, the caption always contains information about its length in micrometers in the form of a simple string such as “Bar = 20 μm.”, “Bar, 20 μm.” or, in the case where more than one panel has a bar, “Bars, 20 μm.”. Therefore, to obtain the number, the caption was searched for the word “bar” or “Bar”, followed by an optional “s”. Then the following number was taken to be the size of the scale bar, ignoring intervening characters like “,” or “=”. The units were also ignored and always assumed to be micrometers (due to inconsistency of font usage for Greek characters such as mu).

The remaining challenge in assigning a scale to each panel was handling implied assignments. The two simplest cases are: 1) each panel has its own scale bar, in which case each assignment is explicit, or 2) there is only one scale bar per figure, which implies that the same scale bar applies to each and every panel in the same figure. However, microscope image panels are frequently arranged in rows and columns with only one scale bar per row or column. In such cases each scale bar was made a candidate assignee for its respective row and column. Simple conflict resolution was then used to determine whether it should apply to the row or the column. Occasionally one encounters more cryptic implied scale bar assignments such as more than one bar per row and more than one per column. Such cases were ignored and no scale assigned to any panels in such figures.

In order to evaluate the scale bar finding module, we examined a test set of 167 panel-caption pairs in which a scale was present. The scale bar finder reported a scale for 110 of these panels, and the scale was correct for 84 of these. This corresponds to a precision of 76% but a recall of only 50%. Further work will be needed to refine the scale finding algorithm.

Classification of Images from Different Sources

Since the goal of this work is the development of a self-populating knowledge base containing classified fluorescence images of protein localization patterns from journal articles, such a system necessarily must be able to consistently and correctly classify images of different cell types, obtained using different equipment. Additionally, because the classification methods to be used here were previously applied only to homogeneous image sets of a given cell line and microscopy method, their applicability to mixed sets was tested. To this end, three sets of images were merged, representing two cell lines (Chinese Hamster Ovary (CHO), and HeLa) and two microscopy methods (plane deconvolution and laser scanning confocal). Specifically, these were sets of CHO [4] and HeLa images [6] both obtained via deconvolution microscopy, and a set of HeLa images obtained using laser scanning confocal microscopy. We combined the four classes that were common to the three original sets. These were the major, distinct localization patterns of DNA (217 images), lysosomes (196 images), Golgi (215 images), and tubulin (202 images).

We calculated the SLF3 feature set [6] for all images. The images for each class were randomly divided into a training set (100 images per class), stop training set (50 images per class), and test set (46-67 images per class, depending on class). Ten different random subdivisions were generated and each was used to train and test a BPNN (topology 78-20-4). Table 1 shows the confusion matrix averaged across the ten trials. The classifier was able to recognize images of all four classes with accuracies ranging from 88% for DNA to 97% for tubulin. Based on the overall average classification accuracy of 92%, we conclude that the current methods can be applied to train a network that will classify images of different

True Class	Output of the Classifier			
	DNA	**Giantin**	**LAMP2**	**Tubulin**
DNA	**88%**	1%	3%	7%
Giantin	1%	**95%**	3%	2%
LAMP2	1%	8%	**89%**	2%
Tubulin	2%	0%	1%	**97%**

Table 1. Classification results for images from different sources over ten trials using a BPNN with one hidden layer of 20 nodes, and the SLF3 feature set (78 features).

cell lines, obtained using different equipment, with high accuracy (at least for lines as similar as CHO and HeLa).

Simulation of Publication-Associated Image Perturbations

To address the issue of publication-associated manipulations, we have carried out a controlled characterization of the sensitivity of subcellular location analysis to such manipulations. Our strategy was to carry out these manipulations on the set of HeLa cell images used for our previous classification work [5, 6] and evaluate the effect on our SLF features and on the accuracy of classification with those features. The image set contained approximately 85 images for each of ten classes of subcellular patterns (nuclear DNA, a nucleolar protein, F-actin, tubulin, a mitochondrial protein, an endoplasmic reticulum protein, a lysosomal protein, an endosomal protein and two Golgi proteins).

When images are prepared for publication (either by authors or publishers), they are frequently subjected to the following operations:

1) **Resizing**. Images that are collected with a particular number of pixels often must be resized to conform to the column width of a journal. For most image editing packages, resizing involves low-pass filtering before the image is resampled. This means high-frequency information is lost. Also, error is introduced into the pixel values by the interpolation process.

2) **Intensity modification**. When integer intensity values are rescaled (e.g. to adjust brightness and contrast or to perform gamma correction), quantization error is introduced since the new intensity values are also stored as integer values. This error becomes larger as the bit-depth decreases.

3) **Lossy Compression**. To save space, images in online journals are often significantly compressed relative to the image originally provided by the authors. The most frequently used method is JPEG compression, which discards high-frequency information. This is because JPEG was designed for "natural" images which typically do not contain many sharp edges.

We created sets of images which each had one of these operations performed on it to a particular degree. All operations were carried out with the Matlab Image Processing Toolbox. JPEG compression was carried out using quality settings of 90%, 70% and 50%. Intensity scaling was performed by integer division by 2, 4, 8 and 16 (a majority of the images had bit-depths ranging from 5 to 9). Lastly, image resizing was done to 70%, 40%, 20% and 10% of the original pixel width of the image.

The full set of our 84 features (termed set SLF4) was calculated for each perturbed image and two types of comparisons of the features from the perturbed and unperturbed images were performed.

Univariate Comparisons of Features from Perturbed and Unperturbed Images

We first examined potential effects on individual features. For each perturbation and for each class, we performed a t-test to test the hypothesis that the distribution of each feature was unchanged from that for the unperturbed images. Using a confidence level of 0.05, we identified those features for which this null hypothesis was accepted for all ten classes of images. The results are summarized in Table 2.

The results show that almost all of the features are sensitive to image size rescaling, indicating that they are potentially not suitable in their current form for comparison of images with different magnification. The Haralick texture features are quite sensitive to all image manipulations, as is to be expected given that they are calculated from the correlations between neighboring pixel intensities. By contrast, the Zernike moment features are robust against compression and modest intensity scaling. Intensity scaling by factors of 8 or 16 was too large given the signal in the original images, since they led to large numbers of blank images. Even in these cases, the values of some of the Zernike moments were preserved. In addition, many of the features derived from morphological image processing (e.g., number of objects, average object size) are also robust against modest intensity scaling.

Classification Using Features from Perturbed and Unperturbed Images

We evaluated the overall robustness of the features to perturbations by comparing the performance of a BPNN classifier on the features from the perturbed images against performance on the unperturbed images.

	JPEG 90%	JPEG70%	JPEG50%	I/2	I/4	I/8	I/16	Size 70%	Size 40%
Object number		+	+	+					
Euler number	+	+	+	+	+				
Average object size	+	+	+	+	+				
Variance of obj. size	+	+		+	+				
Max/min ratio obj. sz.	+	+	+	+					
avg. distance to COF	+	+	+	+					
Variance of distance	+	+		+	+	+			
Max/min ratio of dist.		+		+	+				
Fract. Along edge	+	+	+	+					
Edge homogeneity								+	
Edge direction ratio1									
Edge direction ratio2									
Edge direction diff.									
Fract. Of convex hull				+				+	
Convex hull shape				+	+			+	
Convex hull eccent.				+	+			+	+
avg. distance to DNA	+	+	+						
var. distance to DNA	+	+		+					
m/m ratio dist. DNA	+	+		+					
Prot-DNA COF dist.	+	+	+	+					
Prot-DNA area ratio								+	+
Prot-DNA overlap								+	
49 Zernike moments	1	all	all	2	3	4	5		
13 Haralick features	6			6					

Table 2. Robustness of SLF4 features to image manipulations. For each condition, those features that were considered indistinguishable (at a confidence level of 0.05) from the corresponding values for the unperturbed images are shown with a "+".

1) All unchanged except Z1,1 and Z2,0
2) All unchanged except Z0,0
3) All unchanged except Z0,0 and Z12,2
4) 29 of 49 unchanged
5) 14 of 49 unchanged
6) Only the "correlation" feature was considered to be unchanged

For each type of perturbation a BPNN classifier (topology 84-20-20-10) was trained and tested for ten trials. For each trial, images from each class were randomly divided into a training set (40 images per class), a stop training set (20 images per class) and a test set (13-38 images per class depending on class). The results were reported as the average across the 10 trials of the correct classification rate on the test set. When the classifier was trained using unperturbed features, an average accuracy of 79% (with a 95% confidence interval of 5%) was obtained (Table 3). As we have reported previously, all ten classes can be distinguished.

Classification of Perturbed Image Sets

The same approach was used to train and test classifiers using each set of perturbed features. The results are summarized as average classification accuracy in Table 4. Note that the classification accuracy figures have a 95% confidence interval of about 5%, which means that 80% accuracy for Intensity/2 and 79% for Intensity/4 can be considered statistically the same. The same goes for all of the JPEG results.

The results show that modest intensity modifications make little difference to classification accuracy, whereas

True Class	Output of the Classifier									
	DNA	**ER**	**Gia**	**GPP**	**LAM**	**Mit**	**Nuc**	**Act**	**TfR**	**Tub**
DNA	**99%**	1%	0%	0%	0%	0%	0%	0%	0%	0%
ER	0%	**92%**	0%	0%	2%	2%	0%	0%	2%	3%
Giantin	0%	2%	**75%**	16%	0%	1%	2%	0%	3%	0%
GPP130	0%	0%	16%	**77%**	1%	1%	3%	0%	2%	0%
LAMP2	0%	1%	4%	1%	**67%**	3%	3%	0%	20%	1%
Mitochondria	0%	15%	0%	0%	1%	**68%**	0%	2%	5%	11%
Nucleolin	1%	0%	1%	2%	1%	0%	**95%**	1%	1%	0%
Actin	0%	0%	0%	0%	1%	1%	0%	**86%**	1%	12%
TfR	0%	2%	1%	0%	20%	6%	0%	7%	**52%**	11%
Tubulin	0%	4%	0%	0%	0%	10%	1%	5%	3%	**78%**

Table 3. Classification results for unperturbed HeLa cell data over ten trials using a BPNN with two hidden layers of 20 nodes each and the SLF4 feature set (84 features). For each trial, the images for each class were randomly divided into a training set, a stop training set, and a test set. The results on the test set were averaged over the ten trials. Instances of confusion greater than 15% are shaded.

JPEG compression compromises the accuracy by a small amount. As was observed for the univariate tests described above, intensity reduction by 8 or 16 resulted in large numbers of blank images. This prevented training of the classifier for these conditions. Resizing to 70% of original size did not seem to have an adverse effect on classification, whereas resizing of 40% or more resulted in incomputable Haralick and Convex Hull features for many of the images and there were too few "usable" images left for the classifier to be trained. We conclude that the SLF features are robust to a minor resizing but are not necessarily usable for greater reductions in image size or microscope magnification.

Perturbation Type	Classification Accuracy
Unperturbed	79%
Intensity/2	79%
Intensity/4	81%
Resize 70%	79%
JPEG 90%	72%
JPEG 70%	72%
JPEG 50%	74%

Table 4. BPNN classification accuracies for the perturbed image sets (averaged over 10 trials for each set and over the ten classes).

Classification of Perturbed Image Sets Mixed with the Unperturbed Set

Next we tested the hypothesis that a classifier can be made more robust against a given type of perturbation by training it with a with a set of examples that includes several levels of that perturbation. We therefore trained classifiers with mixed sets composed of the unperturbed set plus succesive levels of each perturbation. An advantage of this approach is that while for the individual perturbed set classification the sets I/8 and I/16 as well as R40 to R10 could not be used due to too few "usable" images being left, here they could be put to use.

As reported in the previous section, the correct classification rate for just the unperturbed set is 79%. Interestingly, the classification accuracy for all levels of mixed sets (Table 5) for intensity and resizing perturbations ranges from 78% to 80%, which is statistically the same as 79%, given the 95% confidence interval of 5%. For JPEG perturbations the accuracy is 4% lower than the unperturbed set even for just U+J90. However, even this difference is within the 95% confidence interval, and therefore it can be considered that JPEG perturbed images can be recognized with nearly the same accuracy as unperturbed images. We can conclude that it is possible to train robust classifiers that can recognize subcellular location patterns in images that come at a range of different resolutions and that have been subjected to various degrees of intensity quantization and lossy compression. Since these classifiers are based on the SLF4 features, it must also be true that the SLF4 feature set itself is robust to the named perturbations. This result may be counter-intuitive considering the results from the

Sets Combined	Classification Accuracy
U	79%
U+I/2	78%
U+I/2+I/4	78%
U+I/2+I/4+I/8	78%
U+I/2+I/4+I/8+I/16	79%
U+R70	80%
U+R70+R40	80%
U+R70+R40+R20	80%
U+R70+R40+R20+R10	78%
U+J90	75%
U+J90+J70	76%
U+J90+J70+J50	75%

Table 5. BPNN classification accuracies for the combinations of the perturbed image sets (averaged over 10 trials for each set).
U – Unperturbed
I/x – Intensity reduction by x
Rx – Resizing to x percent of original size
Jx – JPEG compression at quality level x

univariate comparisons of feature values for different levels of perturbations, where it was seen that the distribution of many features changed considerably. Apparently data points from different classes remain as separable clusters in feature space even though the distribution of the points is changed by the perturbations.

It should be noted that before mixing the Resized sets, four of the SLF4 features that are explicitly size-dependent ("Average object size", "Variance of obj. size", "Avg. Distance of objects to COF" and "Variance of object to COF distance") were rescaled to normalize them to the same range as the unperturbed set.

Based on the comparisons of the individual and combined perturbed sets, and in accordance with the results from the univariate comparisons of feature values, we conclude that the SLF4 feature set is robust to moderate levels of lossy compression and intensity scaling. They are not, in the present form usable for images of different resolution. We also conclude that a somewhat paradoxical improvement in classification accuracy can be obtained by including perturbed images during training.

Using the Classifier to Search for Online Images

Having found that our classifier is robust to pattern changes due to different microscopy methods and cell types, and that it can reasonably withstand resolution changes, intensity scaling and lossy compression, we next wondered whether it would be possible to use it in combination with the web robot to find and classify fluorescence microscope image of a defined subcellular pattern.

As an initial test, we performed a search for "tubulin" using the online journal search robot described above, and then used a BPNN to rank the resulting images in order of their "likelihood" of depicting tubulin. The BPNN (topology 65-20-2) used the SLF6 feature set (SLF4 minus DNA features and minus Haralick texture features), because DNA features were not computable for online images and Haralick texture features had shown more sensitivity to image perturbations than other features. This network had just two outputs, one for class "tubulin" and one for class "other", so that it could be optimized for the task of identifying one single class apart from everything else. The network was trained with the set of perturbed HeLa cell images where all perturbations had been mixed together in an effort to provide maximum robustness. In addition, to provide more fault tolerance, ten independent BPNN-s were trained, each with different randomly chosen subset of training data.

The ranking of the output images from the online image collector was done by computing the SLF6 features from each image, applying those to each of the ten networks, and then converting the network outputs for each image to a single numerical score. This way the images could be ranked by their numerical score. The scoring was based on the notion that the outputs of a BPNN approximate the probability of the input pattern belonging to respective class of each output node. The overall score for an image was a sum of two terms: 1) The average of the scores for the 10 networks, 2) Negative three times the standard deviation of the scores for the 10 networks. The first term was designed to give high scores to images for which the "tubulin" output was dominant and the second term was designed to penalize those images for which the 10 different networks were not in good agreement with each other (in other words the confidence was low).

The "tubulin" query returned 587 images that the program considered to be single cells (note that not all of these images would be expected to be of tubulin, since any other fluorescence microscope images in articles about tubulin would also be returned). The first 200 of these images were manually labeled as "tubulin" or "other" by reading the figure captions. 31 of the images (16%) were of the "tubulin" class. Figure 2 shows the cumulative percentage of images that were actually of tubulin as a function of their rank. Eight of the 10 highest ranked images were found to depict a tubulin pattern, which is much higher than expected for random ranking.

Conclusions

We have described a totally automated method to find and interpret the fluorescence microscope images contained in articles in online journals. Our system includes 1) a web robot to download articles matching a query, 2) a tool to extract figure images and the

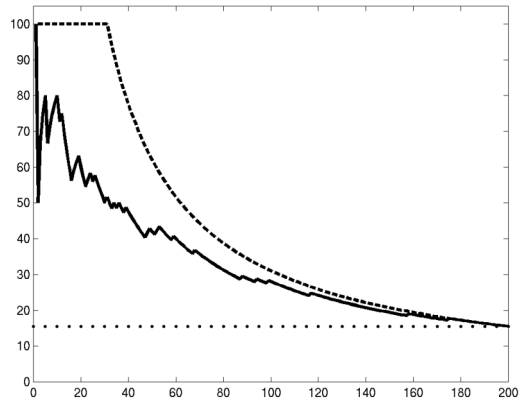

Figure 2. Performance of the combined system using a query for articles containing the word "tubulin." Single cell images returned from online articles were classified and ranked with a binary classifier as described in the text. The cumulative percentage of images returned that were judged to actually depict tubulin is shown as a function of rank. The dotted line shows expected performance for random ranking while the dashed line shows ideal performance.

corresponding captions from PDF files, 3) a program for splitting figures into individual panels, 4) a filter program for identifying fluorescence microscope images among the panel images, 5) a program for removing annotations from images, 6) a segmentation program based on the watershed algorithm to isolate individual cells from images containing multiple cells, and 7) a minimal program to find scale information from the images and captions. Evaluations of each of these steps showed good precision and reasonable recall.

We have also characterized the sensitivity of the features that we have developed for interpreting subcellular patterns to manipulations that may occur during publication. The sensitivity was assayed in two ways, using either univariate comparison of feature values for perturbed and unperturbed features, or testing performance with neural network classifiers. The results clearly indicate that our SLF features are robust to intensity rescaling and image resizing as well as modest levels of lossy compression. We also investigated the classification of images from multiple sources. We showed that despite the variation of patterns across two different types of microscopy and two different cell types the classifier was able to distinguish four of the major classes with high accuracy. Combining these findings with the encouraging results from the perturbation sensitivity analysis, we conclude that it will be possible to improve our Subcellular Location Feature sets such that they can be used for automated analysis of cell images from the most heterogenous set imaginable - online journals and databases. Encouraging preliminary results in this direction were obtained using a binary classifier trained on our images of tubulin to find other images of tubulin in journal articles.

In future work, we plan to further characterize the sensitivity of the features to different microscopy methods and cell types, but also to different specimen preparation methods (e.g., live vs. fixed cells) and different cell states (e.g., mitosis vs. interphase). Ultimately, the methods we describe here will be used to create a knowledge base for protein location that will contain supported assertions automatically generated from full-text journal articles and other online sources.

Acknowledgements

The work described in this article was supported in part by NIH grant CA83219 and by NSF Science and Technology Center grant MCB-8920118. M.V. was supported by a fellowship from the Center for Automated Learning and Discovery (NSF grant REC-9720374). G.P. was supported by a Summer Scholar award from the Merck Computational Biology and Chemistry Program through a grant from the Merck Company Foundation.

Literature Cited

[1] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," *ISMB*, vol. 7, pp. 77-86, 1999.

[2] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, "EDGAR: extraction of drugs, genes and relations from the biomedical literature," *Pac. Symp. Biocomput.*, vol. 5, pp. 517-28., 2000.

[3] M. V. Boland, M. K. Markey, and R. F. Murphy, "Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy," presented at the 19th Annu. Intl. Conf. IEEE Eng. Med. Biol. Soc., Chicago, IL, USA, 1997.

[4] M. V. Boland, M. K. Markey, and R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry*, vol. 33, pp. 366-375, 1998.

[5] R. F. Murphy, M. V. Boland, and M. Velliste, "Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images," *ISMB*, vol. 8, pp. 251-259, 2000.

[6] M. V. Boland and R. F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics*, vol. 17, in press, 2001.