Journal of VLSI Signal Processing 35, 311–321, 2003

Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images*

ROBERT F. MURPHY, MEEL VELLISTE† AND GREGORY PORRECA**
Departments of Biological Sciences and Biomedical Engineering, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA

Received December 12, 2002; Revised February 17, 2003

Abstract. The ongoing biotechnology revolution promises a complete understanding of the mechanisms by which cells and tissues carry out their functions. Central to that goal is the determination of the function of each protein that is present in a given cell type, and determining a protein's location within cells is critical to understanding its function. As large amounts of data become available from genome-wide determination of protein subcellular location, automated approaches to categorizing and comparing location patterns are urgently needed. Since subcellular location is most often determined using fluorescence microscopy, we have developed automated systems for interpreting the resulting images. We report here improved numeric features for describing such images that are fairly robust to image intensity binning and spatial resolution. We validate these features by using them to train neural networks that accurately recognize all major subcellular patterns with an accuracy higher than any previously reported. Having validated the features by using them for classification, we also demonstrate using them to create Subcellular Location Trees that group similar proteins and provide a systematic framework for describing subcellular location.

Keywords: protein localization, subcellular location features, fluorescence microscopy, pattern recognition, location proteomics

Introduction

The genome of each organism encodes tens of thousands of proteins, some of which are made in all cells of that organism and some of which are made only in specific cell types. Recent years have seen the determination of the entire DNA sequences of a number of genomes, leading to a major paradigm shift in biological research. This is from research groups each studying many aspects of an individual protein (primarily by manual methods) to larger scale projects in which a specific aspect of all proteins expressed in a given cell type are studied primarily by automated methods. Aspects that have received significant attention include the amount of each protein expressed, their three-dimensional structures, the ways in which they are chemically modified after synthesis, and their ability to bind to other proteins.

Animal and plant cells have a number of subcompartments (such as lysosomes) and subcellular structures (such as the cytoskeleton) that play distinct and essential roles in cell functions. When characterizing a protein, determining its location within cells is critical to understanding its function, since each subcompartment has a different biochemical environment.

*Based on "Robust Classification of Subcellular Location Patterns in Fluorescence Microscope Images" by Robert F. Murphy, Meel Velliste, and Gregory Porreca, which appeared in the Proceedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing. © 2002 IEEE.

†Present position: Postdoctoral fellow, Department of Neurobiology, University of Pittsburgh, PA, USA.

**Present position: Graduate student, Biological and Biomedical Science Program, Harvard University, USA.

A comprehensive, systematic approach to determining, describing and/or predicting the subcellular location of proteins is needed for this purpose.

A number of factors have limited progress in this area in the past. The first is ambiguity in the words used to describe subcellular locations. Different investigators use different terms to describe the same pattern and the same term is often used to describe protein patterns known not to be identical. Second, there have been no reliable, automated methods to map between images depicting patterns and words describing them. Third, comprehensive knowledge of all possible locations (and combinations of locations) that proteins may exhibit does not exist. Restated, what is missing is a grouping of all proteins such that the proteins in each group all have an identical distribution in cells (and thus the unique identifier of that group can be assigned to each of the members). This is necessary if location information is to be satisfactorily included in biological databases, as has been done for protein sequence and structure families. Fourth, while methods for comparing protein and nucleotide sequences (and structures) are well established and can be used directly from database entries, comparing protein locations from database entries is not yet possible. The concepts of hierarchical organization and distance have not yet been developed for location analysis.

Recently, progress has been made towards overcoming these problems. Laudable efforts towards addressing the first problem have been made by the Gene Ontology Consortium. However, words do not currently exist to describe the full complexity of subcellular location. Our group has addressed the second problem by developing automated methods for determining subcellular location from fluorescence microscope images. Fluorescence microscopy is the most commonly used method for analyzing subcellular location, and it is well-suited to high throughput automation. We began by developing sets of numerical features that describe protein patterns in fluorescence microscope images. We validated these descriptors by using them to develop automated classifiers capable of determining subcellular location from previously unseen images [1–4]. In this paper, we describe improved sets of features that are robust to differences in intensity and spatial resolution between images. We also use these features to begin to address the third and fourth problems discussed above by presenting the first systematic framework for capturing complexity and similarity in protein subcellular location.

Sources of Data on Subcellular Location

With the development of automated, digital fluorescence microscopes over the past 15 years, it is possible to collect the large numbers of fluorescent images for diverse proteins that are required in order to develop, test and use automated interpretation methods. We have created three image datasets to this end (summarized in Table 1). The 2D HeLa dataset has been our primary reference point for developing features and testing methods. It covers all major subcellular structures and organelles and was generated using fluorescent probes that bind to molecules known to be located in those structures: a probe that binds to DNA to label the nucleus, a probe that binds to microfilaments to label the actin cytoskeleton, and antibodies against proteins located in the endoplasmic reticulum, the Golgi apparatus, lysosomes, endosomes, mitochondria, nucleoli and microtubules.

In addition to these datasets, a major National Cancer Institute-funded project led by Jonathan W. Jarvik, Peter B. Berget and Robert F. Murphy is beginning to provide images of the subcellular location of randomly-tagged proteins in 3T3 cells. Preliminary wide-field 2D images (40×, pixel size 0.475 microns) have been acquired for approximately 100 clones produced by random CD-tagging [5–7]. By sequencing DNA adjacent to the tag, the tagged gene has been identified (if it is present in current sequence databases). A high resolution 3D image database for these clones

Table 1. Image sources for development of methods for analysis of protein subcellular location.

Dataset	Number of classes	Number of images per class	Microscopy method	Objective	Size of pixel region in original field	Reference
2D CHO	5	33–97	Deconvolution	100×	0.23 μm	Boland et al. [2]
2D HeLa	10	73–98	Deconvolution	100×	0.23 μm	Boland and Murphy [4]
3D HeLa	11	50–58	Confocal	100×	0.0488 μm	Velliste and Murphy [16]

is currently being acquired by spinning disk confocal microscopy.

Other approaches to random-tagging of genes have also been described [8–10] and thus we can anticipate increased availability of detailed information on protein location over the next few years.

Subcellular Location Features

We have designed a number of sets of numerical features—which we term SLF for Subcellular Location Features—to describe protein subcellular distributions [4]. The types of feature used include Haralick texture features, Zernike moment features, features derived from morphological image processing, and, in some cases, features derived from comparison with a reference image of the DNA distribution in the same cell (see Table 2). The feature sets have been used with several classification methods (linear classifiers, decision trees, k-nearest neighbor classifiers, and one- and two-hidden-layer backpropagation neural networks or BPNNs) and two-hidden-layer BPNNs were found to produce the best results for the 2D HeLa dataset [3] (although the improvement over one-hidden layer networks was within the estimated error). A BPNN with 30 nodes in both hidden layers achieved an average correct classification accuracy of 84% using a set of 37 features (termed SLF5) that was selected by stepwise discriminant analysis [11] from a larger set of 84 features (SLF4) that includes all four types of features.

This represents the best accuracy so far achieved for single 2D images of all major organelle patterns. By analyzing *sets* of images taken from the same slide, the accuracy for these classes was improved to 98% [3, 4]. It is important to note that this classification approach can distinguish two Golgi proteins with very similar patterns that cannot be distinguished by a human observer.

Recently, exciting confirmation that protein subcellular location patterns can be distinguished by automated classifiers with high accuracy has been reported by Danckaert et al. [12]. They employed a Modular Neural Network classifier (MNN, a topological variation on the back-propagation network), to classify 2D images from confocal microscope stacks representing six different subcellular location classes. Instead of using features, the input to the MNN was composed of raw pixel values from a downsampled version of the original image. Each module covered a specific area of

Table 2. Comparison of subcellular location feature sets. All features that measure length or area are calculated in pixels that are 0.23 μm square in the sample plane.

Feature description	SLF3	SLF7
Morphological features: Number of fluorescent objects in image, Euler number of image, average object size, variance of object size, ratio of largest to smallest object size, average object distance to cell center of fluorescence, variance of object distance to cell center, ratio of largest to smallest object distance to cell center	SLF1.1 through SLF1.8	SLF1.1 through SLF1.8
Edge-related features: Fraction of above-threshold pixels along edge, measure of edge gradient intensity homogeneity, measure of edge direction homogeneity 1, measure of edge direction homogeneity 2, measure of edge direction difference	SLF1.9 through SLF1.13	SLF7.9 through SLF7.13 (minor error corrections)
Convex hull features: Fraction of convex hull occupied by above-threshold pixels, roundess of convex hull, eccentricity of convex hull	SLF1.14 through SLF1.16	SLF1.14 through SLF1.16
Zernike moment features through order 12, calculated for a unit circle with radius equal to the average radius of the cell type being analyzed (150 pixels or 34.5 μm for HeLa)	SLF3.17 through SLF3.65	SLF3.17 through SLF3.65
Haralick texture features: angular second moment, contrast, correlation, sum of squares variation, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, info. measure of correlation 1, info. measure of correlation 2	SLF3.66 through SLF3.78	SLF7.66 through SLF7.78 (after downsampling to 1.15 μm/pixel and 256 gray levels)
Fraction of non-object fluorescence	–	SLF7.79
Skeleton features (see text)	–	SLF7.80 through SLF7.84

the organelle image. They found that the trained classifier can recognize individual 2D images from previously unseen 3D image stacks with 84% accuracy. It is worth noting that the image set of Danckaert et al. consisted of images from four different cell types. We had previously observed that classifiers can be trained to recognize images from two cell types and two modes of microscopy [13].

Improving the Subcellular Location Features

The classification accuracy achieved when using the previously described SLF features is already high considering the amount of within-class variability of patterns and the similarity between some pairs of classes. However, when creating a systematics of all proteins in a cell type where thousands of proteins have to be placed within an overall hierarchy, it is desirable to have the best possible representation of the location patterns. We have therefore developed an improved version of our previously developed SLF3 features [4] in combination with six new features described below.

SLF3 includes 13 Haralick texture features that we have subsequently found to be overly sensitive to image pixel resolution and number of gray levels. We therefore characterized the contribution of Haralick features to overall classification accuracy after resampling in various ways (Table 3). To do this, we started with the 34 (out of 37) features from SLF5 that did not require a parallel DNA image and determined average classification accuracy using a BPNN with a single hidden layer of 20 nodes. We did this for only eight of the ten classes in the 2D HeLa dataset, since we found that the original images of the two Golgi classes (giantin and gpp130) are sufficiently different in intensity scale that the original Haralick features can artificially discriminate them on that basis alone. The result was 86.4%. We then determined average accuracy using a set in which the values of the Haralick features were scrambled between images so that the number of features used was still the same (34) but any information in the Haralick features was lost. The result was 81.4%. Thus, the 12 Haralick features included in SLF5 provided a net benefit of 5.0%. Average accuracies for sets containing Haralick features calculated on downsampled images were also found and converted to net benefit. (Note that in all cases the non-Haralick features were calculated using the original images.) We found that Haralick features at a resolution of 1.15 μm/pixel and 32 gray levels were nearly as informative as those calculated on images of higher resolution. Perhaps surprisingly, rebinning to 1.15 μm/pixel but keeping 256 gray levels had a net benefit greater than that of the original features. The Haralick features calculated this way have the advantage of being essentially insensitive to the original spatial resolution of an image, because fluorescence microscope images generally come at a resolution higher than 1.15 μm/pixel and therefore can always be down-sampled to this "standard" pixel size. In addition, these features are essentially insensitive to the original intensity resolution of an image because images are expected to have more than 256 graylevels, and can therefore be re-quantized to the "standard" 256 levels. Even if an original image at high (e.g., 0.23 μm/pixel) spatial resolution only had gray-level values in the range of 0 to 17 (the lowest observed in the 2D HeLa dataset), the down-sampled version will be expected to have more than 256 gray levels since the gray level counts of around 25 pixels would be summed in the course of spatial down-sampling.

Table 3. Relative percent benefit of Haralick features calculated on downsampled and rebinned 2D HeLa images (excluding the giantin and gpp130 classes). At the original resolution of 0.23 μm/pixel and 256 gray-levels the classification accuracy was 86.4%. After scrambling the Haralick features the accuracy decreased to 81.4%, giving a relative benefit of 5.0% (shaded). A similar comparison was made after reducing the resolution (using bi-linear interpolation) and/or the number of gray levels (by division and integer rounding).

	Number of gray-levels		
Pixel size (μm)	256	32	16
0.23	5.0	4.4	3.8
0.69	5.4	3.1	2.9
0.92	5.8	3.1	2.9
1.15	6.2	3.6	3.5

We therefore chose to define a new feature set, SLF7, incorporating the 78 features of SLF3 but with Haralick features from images downsampled to 1.15 μm/pixel and 256 gray levels. We also added six new features:

SLF7.79: The fraction of cellular fluorescence not included in objects

SLF7.80: The average length of the morphological skeleton of objects

SLF7.81: The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects

SLF7.82: The fraction of object pixels contained within the skeleton

Table 4. Confusion matrix for classification of images from the 2D HeLa dataset using SLF8 with a BPNN with a single layer of 20 hidden units over 10 cross-validation trials. The average correct classification rate was 86%.

	Output of the classifier (%)									
True class	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	**98**	2	0	0	0	0	0	0	0	0
ER	0	**87**	0	0	3	5	0	0	1	3
Giantin	0	0	**73**	25	0	0	0	0	1	0
GPP130	1	0	26	**70**	0	0	1	0	1	0
LAMP2	0	3	0	0	**84**	1	0	0	12	0
Mitochond.	0	4	0	0	2	**88**	0	0	4	2
Nucleolin	2	0	2	0	3	1	**93**	0	0	0
Actin	0	0	0	0	0	0	0	**99**	0	1
TfR	0	3	1	0	14	5	0	0	**74**	3
Tubulin	0	2	0	1	0	1	0	1	3	**93**

SLF7.83: The fraction of object fluorescence contained within the skeleton

SLF7.84: The ratio of the number of branch points in the skeleton to the length of skeleton

SLF7.79 was added to measure the amount of fluorescence that is not contained in discrete objects. An object in SLF3 is defined as a contiguous region of above threshold pixels. The relatively dim fluorescence from small vesicles or other structures dispersed throughout the cytoplasm is excluded from objects. The new feature is expected to provide an important distinction between proteins that localize mainly to the same organelle but have different amounts in the cytoplasm. Features SLF7.80 through 7.84 were defined based on the morphological skeleton of objects obtained by thinning using a homotopic interval. This was implemented using the "mmthin" function from the SDC Morphology Toolbox for MATLAB (SDC Information Systems, Naperville, IL, USA). (During the testing of the new features, we discovered that the code for the edge features, SLF1.9 through SLF1.13, had a very minor error; the code was corrected and the corrected features are referred to as SLF7.9 through SLF7.13.)

We validated these features (for all ten classes) using the 2D HeLa dataset by selecting a subset (termed SLF8) via stepwise discriminant analysis and then training and testing using a one-hidden layer BPNN. The 32 features selected were SLF1.3, SLF7.74, SLF3.19, SLF7.79, SLF7.71, SLF7.76, SLF3.23, SLF7.9, SLF1.2, SLF1.6, SLF7.68, SLF3.59, SLF1.8, SLF7.11, SLF3.47, SLF7.70, SLF7.82, SLF1.1, SLF3.24, SLF7.66, SLF7.80, SLF7.69, SLF3.50, SLF1.5, SLF7.84, SLF7.77, SLF7.10, SLF7.73, SLF3.26, SLF7.78, SLF7.72, and SLF1.7. The results (Table 4) indicate a modest gain in performance over our previous best, but, more importantly, demonstrate that this performance can be obtained with fewer, more robust features that are suitable for images from different image sources. We also determined that if rebinning to only 32 gray levels is used, the average accuracy is still 85% (using 27 selected features, data not shown).

We have previously described features SLF2.17 through SLF2.22 that describe protein patterns relative to a parallel DNA image [4]. By combining these with SLF7 and performing stepwise discriminant analysis, we obtained a set of 31 features that we defined as SLF13. The set consists of SLF1.3, SLF7.74, SLF2.22, SLF3.19, SLF7.79, SLF7.71, SLF3.23, SLF7.76, SLF7.9, SLF7.68, SLF1.6, SLF2.19, SLF7.11, SLF1.2, SLF3.37, SLF3.24, SLF7.82, SLF3.60, SLF2.21, SLF7.70, SLF1.1, SLF3.50, SLF7.77, SLF1.5, SLF7.66, SLF2.17, SLF7.84, SLF1.8, SLF7.10, SLF7.69, and SLF7.67. Using SLF13, we obtained an average accuracy of 88% (Table 5).

Comparison with Human Classification

Human ability to visually recognize patterns in real world scenes like faces or road traffic is so far unsurpassed by any automated system. Not surprisingly humans are frequently used as role models in computer vision research. Of course in the case of protein location

Table 5. Confusion matrix for classification of images from the 2D HeLa dataset combined with a parallel DNA image. The SLF13 feature set was used with a BPNN with a single layer of 20 hidden units over 10 cross-validation trials. The average correct classification rate was 88%.

	Output of the classifier (%)									
True class	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	**99**	1	0	0	0	0	0	0	0	0
ER	0	**89**	0	0	4	4	0	0	1	2
Giantin	0	0	**76**	20	0	1	1	0	1	0
GPP130	0	0	23	**73**	0	1	2	0	1	0
LAMP2	0	2	0	0	**83**	1	0	0	13	0
Mitochond.	0	5	0	0	2	**90**	0	0	1	2
Nucleolin	0	0	0	0	0	0	**98**	0	0	0
Actin	0	0	0	0	0	0	0	**99**	0	1
TfR	0	3	0	0	16	3	0	1	**75**	2
Tubulin	0	2	0	0	0	2	0	0	3	**93**

patterns the images do not exactly represent ordinary scenes from the real world that a typical person is likely to grow up looking at. Since a human would have to be specifically trained to recognize such patterns, it is not clear whether humans would perform better than automated systems on this task. Yet most current knowledge on protein subcellular location is based on human interpretation. Therefore it is a natural question to ask whether automated classifiers using SLF features can do as well as humans at recognizing subcellular location patterns, or if they can do better.

To address this question a Matlab script was set up to train and test a human subject. In training mode the script presented a series of images from the 2D HeLa set to the subject and after each image allowed the subject to guess the class of the image. It then told the subject whether the answer was correct, and if not then what the correct class was. In testing mode the script presented another series of images, let the subject classify each image and recorded the responses. For the training set 30 images out of the total of approximately 90 in each class were randomly chosen. Another subset of 30 images per class (not overlapping with the training set) was chosen for testing. The process of training followed by testing was repeated, each time with a different randomly chosen subset of training/test images, until the subject's classification accuracy on the test images stopped improving. The results from the final testing round were taken as the end result of the experiment, i.e. the best possible performance by the subject.

This procedure was carried out by one of the authors (G.P.), a biologist with no prior experience of seeing fluorescence microscope images depicting protein subcellular location patterns but who was well aware of cellular structure and the shapes of all organelles. The classification accuracy reached a plateau after ten rounds of training and testing, achieving a final classification accuracy of 83%. Table 6 shows the confusion matrix from the final test run. The overall classification accuracy was similar to that for the automated system (83% and 86%, respectively). Comparison of Tables 4 and 6 reveals that visual classification was much worse for distinguishing the two Golgi proteins Giantin and gpp130 than automated classification. Human classification performed somewhat better on mitochondria, nucleolin, and TfR, while the computer did somewhat better on LAMP2. Accuracies for the other classes were not significantly different between human and computer. Since Giantin and gpp130 are very similar patterns, it can be concluded that the computer is better than a human at detecting small, almost invisible differences in location, while there is modest room for improvement of the automated system for classifying some patterns that seem clear to humans.

Testing SLF8 on Downsampled Images

The goal in designing SLF7 and SLF8 was to be able to compare and classify images collected with moderately different pixel sizes. As an initial test of whether this goal had been met, we determined the accuracy of classifiers trained with 2D HeLa images that had been downsampled to larger pixel sizes. To avoid potential problems introduced by interpolating during

Table 6. Confusion matrix for human classification of images from the 2D HeLa dataset. The average correct classification rate was 83%.

	Output of the classifier (%)									
True class	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	**100**	0	0	0	0	0	0	0	0	0
ER	0	**90**	0	0	3	6	0	0	0	0
Giantin	0	0	**56**	36	3	3	0	0	0	0
GPP130	0	0	53	**43**	0	0	0	0	3	0
LAMP2	0	0	6	0	**73**	0	0	0	20	0
Mitochond.	0	3	0	0	0	**96**	0	0	0	0
Nucleolin	0	0	0	0	0	0	**100**	0	0	0
Actin	0	0	0	0	0	0	0	**100**	0	0
TfR	0	13	0	0	3	0	0	0	**83**	0
Tubulin	0	3	0	0	0	0	0	3	0	**93**

downsampling, we simply summed 2×2 or 3×3 regions of the original images (resulting in images identical to what would have been obtained had a camera with pixels twice or three times as large been used). When calculating features for these images, we normalized any values involving length or area by the downsampling factor such that the values would be comparable to those for the original 0.23 μm square pixels. We then determined classification accuracies for these images in two ways. First, the set of images for a single pixel size were used for training and testing as was done in Table 4. Second, mixed sets were created by randomly choosing which magnification would be included for each original image. The average correct classification rates over all ten classes are shown in Table 7. The rates for the individual classes are shown graphically in Fig. 1. Since it can be seen that the ability to distinguish between the two closest pairs (giantin/gpp130 and LAMP2/TfR) is dramatically reduced at lower resolution, we also calculated accuracies for an 8 class system created by just merging the corresponding rows and columns in the confusion matrix. It can be seen in Table 7 that accuracies over 80% can be obtained even for mixtures including 0.69 μm pixels.

At least two conclusions can be drawn from these results. The first is that the SLF8 normalization was quite successful, in that mixtures of images with 0.23 μm and 0.46 μm pixels can be classified with only 4–6% loss in accuracy compared to classifying just the original images (as seen in Fig. 1(B) and (D), most classes show little loss in accuracy between these two pixel sizes). The second is that while the ability to distinguish the two most similar pairs of classes is progressively lost with larger pixels, the SLF8 set permits a single classifier to be trained to recognize with over 80% accuracy the basic 8 classes in images with pixel sizes varying over a factor of 3.

Table 7. Classification accuracy for images with varying pixel sizes. Images originally acquired at 0.23 μm were downsampled by a factor of 2 or 3. SLF8 features were calculated for each image, with normalization of the feature values to correct for the change in pixel size. Shown are the average correct classification rates over 10 cross-validation trials with a BPNN as in Table 4 (10 class) for each set of images separately and for mixed sets (see text). Also shown are results for mixtures of downsampled images that were not normalized for pixel size. Average correct classification rates are also shown for 8-class classifiers formed from each 10 class classifier by merging the giantin and gpp130 classes and the LAMP2 and TfR classes.

Image set	10 class (%)	8 class (%)
0.23 μm	84.4	92.9
0.46 μm	77.2	89.4
0.69 μm	73.1	84.9
0.23 μm + 0.46 μm	79.9	87.3
0.23 μm + 0.46 μm + 0.69 μm	69.8	80.1
Mixed, unnormalized	74.5	85.4

The task for these classifiers was to recognize patterns in images for which the pixel size was known, and for which the features could therefore be properly normalized. We also explored whether it was feasible to train classifiers for these patterns in which the input images varied in pixel size (0.23, 0.46, or 0.69 μm/pixel) but the system was not provided with the pixel size for either training or test images (and therefore normalization could not be done). Perhaps surprisingly, the

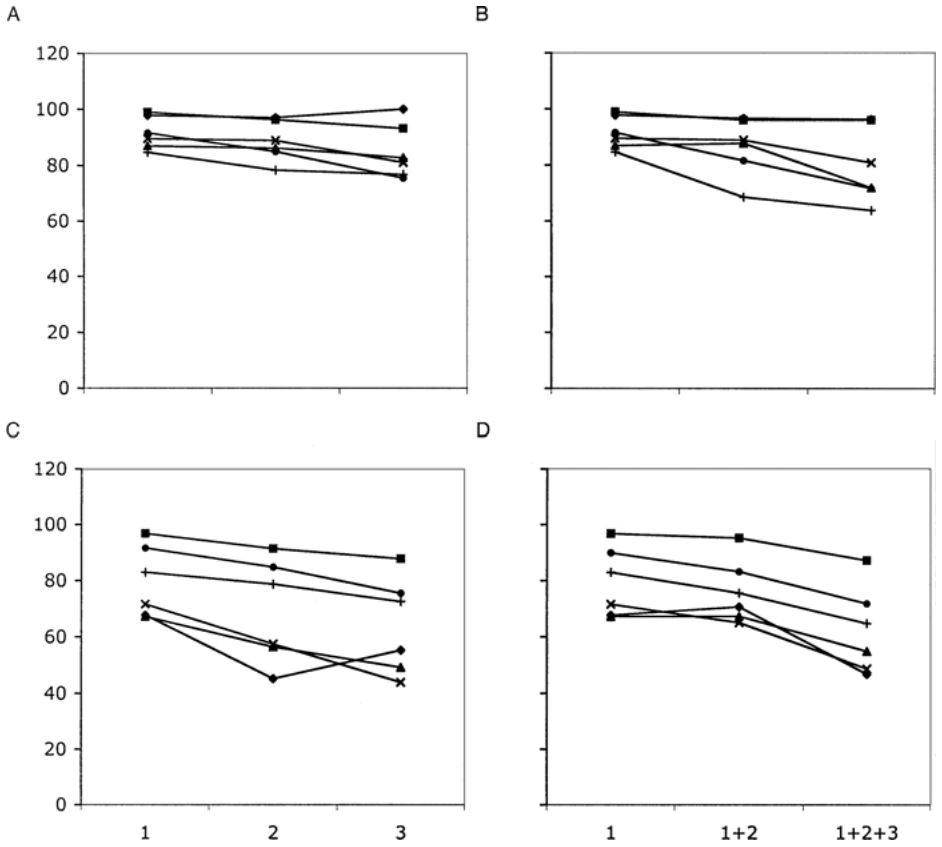

Figure 1. Classification accuracy for individual location classes determined for images with different pixel sizes. The 2D HeLa images, which were originally collected at 0.23 μm/pixel (set 1), were downsampled to 0.46 μm/pixel (set 2, roughly comparable to reducing magnification to 50×) or 0.69 μm/pixel (set 3, comparable to 33× magnification) by averaging 2 × 2 and 3 × 3 regions, respectively. Classification using the SLF8 features was carried out for each pixel size separately (A, C) or for mixed sets in which a pixel size was randomly chosen for each of the original images (B, D). The percent of the test images that was correctly classified is displayed for (A, B) DNA (diamond), ER (triangle), mitochondria (+), nucleoli (×), actin (square), and tubulin (circle) or (C, D) giantin (diamond), gpp130 (triangle), LAMP2 (+), TfR (×), a merged class consisting of the two Golgi proteins (square) and a merged class combining LAMP2 and TfR (circle).

performance of such systems for either 8 or 10 classes was quite good. However, given the way this experiment was performed, it is possible that the classifier learned three distinct sets of decision boundaries (one for each pixel size) and that such a classifier would not perform as well on images with pixel sizes in between these values. Further exploration is needed in this area.

An Example Subcellular Location Tree

The demonstration that the SLF features can adequately describe the major organelle patterns (and also distinguish closely related patterns) allows them to be used to create a systematic framework for protein location. Just as comparison of DNA sequences can be used to create phylogenetic trees that group similar sequences, the SLF features can be used to create "subcellular location trees" that group similar location patterns. To create such trees, we need a measure of the degree of similarity between each pair of classes. For this purpose, we have used the new feature set SLF8. We calculated a feature covariance matrix for all of the images combined and a mean feature vector for each class. We then calculated the Mahalanobis distance between each pair of classes, which is the multivariate distance between the mean feature vectors weighted by the overall covariance matrix. These distances were used to create a dendrogram or hierarchical tree (Fig. 2), in which the distance between adjacent nodes is proportional to the Mahalanobis distance between them.

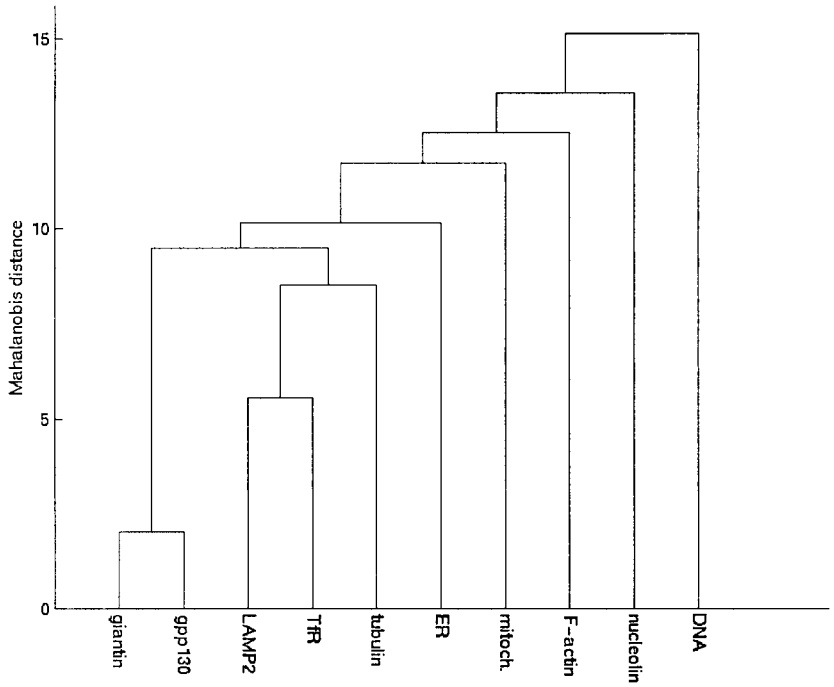

Figure 2. Example subcellular location tree for the 2D HeLa dataset.

As expected, the two Golgi proteins giantin and gpp130 were grouped together, as were the similar patterns of LAMP2 (lysosomes) and transferrin receptor (endosomes). Further examination of Fig. 2 confirms that it is consistent with biological knowledge about the major organelle patterns. For example, the compartments that have a diffuse distribution throughout the cytoplasm (lysosomes and endosomes) that is thought to involve traffic along microtubules are grouped together with tubulin. While Fig. 2 only reflects ten subcellular patterns and we cannot realistically imagine that the arrangement of branches will remain unchanged as more classes are added, it illustrates the utility of generating subcellular location trees (SLT) to organize information about protein location.

Conclusions

We have shown previously that protein subcellular locations can be determined automatically from fluorescence microscope images based on numeric descriptors. We report some improvements in reliability of classification of 2D images, primarily by making features less sensitive to image spatial and intensity resolution and adding new skeleton features. These improvements represent an important step towards generalizing the approaches we have described to other cell types and image sources.

Since the SLF features have been validated by using them to achieve good classification accuracy for subcellular location patterns, it is possible to use them as a basis for building trees to systematize protein subcellular location. We have presented an example Subcellular Location Tree that is consistent with current biological knowledge. We anticipate that our introduction of the concepts of pattern hierarchy and distance measurements to subcellular location will enable new directions in proteomics. Distance measures could, for example, be used to create "location neighbors" in databases. Location distance measures could also be combined with quantitative measures of sequence similarity as part of efforts to understand the sequence motifs that determine subcellular locations.

It should also be noted that it is possible to use the SLF features for other automated analyses of fluorescence microscope images, such as for automated selection of representative images from a set [14], rigorously comparing two sets of images [15], and finding and interpreting fluorescence microscope images in journal articles or web pages [13]. The confluence of genomics, protein tagging methods, automated

microscopy and pattern interpretation methods is opening a new frontier in computational biology.

Acknowledgments

This work was supported in part by NIH grant R33 CA83219 and by a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund. G.P. was supported by a Summer Scholar award from the Merck Computational Biology and Chemistry Program made possible by a grant from the Merck Company Foundation.

References

1. M.V. Boland, M.K. Markey, and R.F. Murphy, "Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy," in *19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, USA, 1997, pp. 594–597.
2. 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, 1998, pp. 366–375.
3. 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," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, San Diego, 2000, pp. 251–259.
4. 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, 2001, pp. 1213–1223.
5. J.W. Jarvik, S.A. Adler, C.A. Telmer, V. Subramaniam, and A.J. Lopez, "CD-Tagging: A New Approach to Gene and Protein Discovery and Analysis," *Biotechniques*, vol. 20, 1996, pp. 896–904.
6. C.A. Telmer, P.B. Berget, B. Ballou, R.F. Murphy, and J.W. Jarvik, "Epitope Tagging Genomic DNA Using a CD-Tagging Tn10 Minitransposon," *Biotechniques*, vol. 32, 2002, pp. 422–430.
7. J.W. Jarvik, G.W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P.B. Berget, "In Vivo Functional Proteomics: Mammalian Genome Annotation Using CD-Tagging," *BioTechniques*, vol. 33, 2002, pp. 852–867.
8. M.M. Rolls, P.A. Stein, S.S. Taylor, E. Ha, F. McKeon, and T.A. Rapoport, "A Visual Screen of a GFP-Fusion Library Identifies a New Type of Nuclear Envelope Membrane Protein," *J. Cell Biol*., vol. 146, 1999, pp. 29–44.
9. A. Kumar, K.-H. Cheung, P. Ross-Macdonald, P.S.R. Coelho, P. Miller, and M. Snyder, "TRIPLES: A Database of Gene Function in Saccharomyces Cerevisiae," *Nucleic Acids Research*, vol. 28, 2000, pp. 81–84.
10. G. Habeler, K. Natter, G.G. Thallinger, M.E. Crawford, S.D. Kohlwein, and Z. Trajanoski, "YPL.db: The Yeast Protein Localization Database," *Nucleic Acids Research*, vol. 30, 2002, pp. 80–83.
11. R.I. Jennrich, "Stepwise Discriminant Analysis," in *Statistical Methods for Digital Computers*, K. Enslein, A. Ralston, and H.S. Wilf (Eds.), John Wiley & Sons: New York, 1977, pp. 77–95.
12. A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes, "Automated Recognition of Intracellular Organelles in Confocal Microscope Images," *Traffic*, vol. 3, 2002, pp. 66–73.
13. R.F. Murphy, M. Velliste, J. Yao, and G. Porreca, "Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Locations," in *Proceedings of the 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001)*, Bethesda, MD, USA, 2001, pp. 119–128.
14. M.K. Markey, M.V. Boland, and R.F. Murphy, "Towards Objective Selection of Representative Microscope Images," *Biophys. J.*, vol. 76, 1999, pp. 2230–2237.
15. E.J.S. Roques and R.F. Murphy, "Objective Evaluation of Differences in Protein Subcellular Distribution," *Traffic*, vol. 3, 2002, pp. 61–65.
16. M. Velliste and R.F. Murphy, "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," in *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002)*, Bethesda, MD, USA, 2002, pp. 867–870.

Robert F. Murphy earned an A.B. in Biochemistry from Columbia College in 1974 and a Ph.D. in Biochemistry from the California Institute of Technology in 1980. He was a Damon Runyon-Walter Winchell Cancer Foundation postdoctoral fellow with Dr. Charles R. Cantor at Columbia University from 1979 through 1983, after which he became an Assistant Professor of Biological Sciences at Carnegie Mellon University in Pittsburgh, Pennsylvania. He received a Presidential Young Investigator Award from the National Science Foundation shortly after joining the faculty at Carnegie Mellon in 1983 and has received research grants from the National Institutes of Health, the National Science Foundation, the American Cancer Society, the American Heart Association, the Arthritis Foundation, and the Rockefeller Brothers Fund. He has co-edited two books and published over 90 research papers. His research group at Carnegie Mellon focuses primarily on the application of fluorescence methods to problems in cell biology, with particular emphasis on automated interpretation of fluorescence microscope images. He has a long-standing interest in computer applications in biology, and developed the Computational Biology curriculum that he has taught at Carnegie Mellon since 1989. In 1984, he co-developed the Flow Cytometry Standard data file format used throughout the industry and he is Chair of the Cytometry Development Workshop held each year in

Asilomar, California. He is Professor of Biological Sciences and Biomedical Engineering and Director of the Merck Computational Biology and Chemistry Program at Carnegie Mellon.
murphy@cmu.edu

Meel Velliste received a bachelors degree in Computer Science and Cybernetics from the University of Reading, UK, in 1998. He earned a doctorate in Biomedical Engineering from Carnegie Mellon University, USA in 2002, working under the direction of Robert F. Murphy on automated methods for interpretation of subcellular location patterns in fluorescence microscope images. He is presently doing post-doctoral research on neuroprosthetics and the primate motor cortex in the Andrew B. Schwartz laboratory at the University of Pittsburgh. His research interests include Signal/Image Processing, Computer Vision, Machine Learning, Neural Systems and Neuroprosthetics.

Gregory J. Porreca earned a B.S. in Biology and Computer Science from the College of New Jersey in 2002 and is currently a first-year Ph.D. student at Harvard Medical School in the Biological and Biomedical Sciences Program of the Division of Medical Sciences. As an undergraduate, he worked at the Protein Data Bank operated by the Research Collaboratory for Structural Bioinformatics at Rutgers University, and participated in the Summer Undergraduate Research Program at Carnegie Mellon University where he was supported by a grant from the Merck Company Foundation.