Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images

Robert F. Murphy, Michael V. Boland and Meel Velliste

Department of Biological Sciences, Biomedical Engineering Program,
and Center for Light Microscope Imaging and Biotechnology,
Carnegie Mellon University
4400 Fifth Avenue
Pittsburgh, PA 15213
murphy@cmu.edu

Abstract

Determination of the functions of all expressed proteins represents one of the major upcoming challenges in computational molecular biology. Since subcellular location plays a crucial role in protein function, the availability of systems that can predict location from sequence or high-throughput systems that determine location experimentally will be essential to the full characterization of expressed proteins. The development of prediction systems is currently hindered by an absence of training data that adequately captures the complexity of protein localization patterns. What is needed is a systematics for the subcellular locations of proteins. This paper describes an approach to the quantitative description of protein localization patterns using numerical features and the use of these features to develop classifiers that can recognize all major subcellular structures in fluorescence microscope images. Such classifiers provide a valuable tool for experiments aimed at determining the subcellular distributions of all expressed proteins. The features also have application in automated interpretation of imaging experiments, such as the selection of representative images or the rigorous statistical comparison of protein distributions under different experimental conditions. A key conclusion is that, at least in certain cases, these automated approaches are better able to distinguish similar protein localization patterns than human observers.

Introduction

As the initial sequencing of a number of eukaryotic genomes is completed, a major effort to determine the structure and function of all expressed proteins is beginning. Significant current work in this new area (often referred to as functional genomics or proteomics) is devoted to the analysis of gene and protein expression patterns (in different adult tissues and during development) and the prediction and determination of protein structure. More limited efforts have been made to predict aspects of protein function. Overlooked in much of this work is the importance of subcellular location for proper protein function.

The organelle or structure where a protein is located provides a context for it to carry out its role. Each organelle provides a unique biochemical environment that may influence the associations that a protein may form and the reactions that it may carry out. For example, the concentrations of protons, sodium, calcium, reducing agents, and oxidizing agents vary dramatically between organelles. Thus we can imagine that knowledge of the location(s) in which a previously uncharacterized gene product will be found would be of significant value when attempting to determine (or predict) its properties.

Such knowledge can be obtained by two approaches: experimental determination or prediction. Experimental determination of subcellular location is accomplished by three main approaches: cell fractionation, electron microscopy and fluorescence microscopy. As currently practiced, these approaches are time consuming, subjective, and highly variable. While each method can yield important information, they do not provide unambiguous information on location that can be entered into databases.

There have been pioneering efforts to predict subcellular location from protein sequence (Eisenhaber and Bork 1998; Horton and Nakai 1997; Nakai and Horton 1999; Nakai and Kanehisa 1992). These efforts have been modestly successful, correctly classifying approximately 60% of proteins whose locations are currently known. A major limitation of the usefulness of these systems is that only broad categories of subcellular locations were used (see Table 1). This limitation is a reflection of the nature of the available training data: the subcellular location is only approximately known for most known proteins.

For example, the categories used by YPD (Garrels 1996)are overlapping but not explicitly hierarchical (the

PSORT	YPD
	Bud neck
	Cell wall
	Centrosome/spindle pole body
chloroplast	
cytoplasm	Cytoplasmic
	Cytoskeletal
endoplasmic reticulum	Endoplasmic reticulum
	Endosome/Endosomal vesicles
outside	Extracellular (excluding cell wall)
Golgi body	Golgi
	Lipid particles
lysosome/vacuole	Lysosome/vacuole
	Microsomal fraction
mitochondria	Mitochondrial
mitochondria inner membrane	Mitochondrial inner membrane
mitochondria intermembrane space	Mitochondrial intermembrane space
mitochondria matrix space	Mitochondrial matrix
mitochondria outer membrane	Mitochondrial outer membrane
nucleus	Nuclear
	Nuclear matrix
	Nuclear nucleolus
	Nuclear pore
	Nuclear transport factor
	Other vesicles of the secretory/endocytic pathways
microbody (peroxisome)	Peroxisome
plasma membrane	Plasma membrane
	Secretory vesicles
	Unspecified membrane

Table 1. Categories of subcellular localization used for previous prediction systems. The categories used by PSORT (Nakai and Kanehisa 1992) for yeast, animal and plant cells are shown. The categories used in YPD (Garrels 1996) have been used to test other prediction approaches.

microsomal fraction is expected to include much of the endoplasmic reticulum, endosomes, and the Golgi apparatus). They also do not provide sufficient resolution to determine if two proteins can be expected to show the same distribution. For example, a protein predominantly located in the *cis* Golgi protein would have a different subcellular location pattern than one in the *trans* Golgi (and would be expected to show different protein sorting motifs). Similarly, endosomes in animal cells can be resolved into (at least) early endosomes, sorting endosomes, recycling endosomes, and late endosomes.

The limitations of current knowledge on subcellular location can be further verified by inspection of the "subcellular location" field of the Swiss-PROT database. The contents of the field fall into three broad categories. For most proteins, it is empty. For many, it consists of a brief, standardized but very general description, such as "integral membrane protein" or "cytoplasmic." For the remaining proteins, the field contains unstructured text that varies from being very general to quite specific. The ambiguities in database descriptions and reports of experiments reflect imprecision and investigator-to-investigator variation in terminology (especially in the endomembrane system), uncertainty about the actual location of many proteins (e.g., whether a Golgi protein is in cis or medial cisternae), and the fact that many proteins cycle between different locations. What is needed is a systematic approach to describing subcellular location that can

- incorporate information obtained by diverse methodologies
- address differences in cell morphology and organelle structure between cell types
- provide sufficient accuracy and resolution that proteins with similar but not identical subcellular locations can be distinguished
- reflect the possibility that subcellular localization patterns can be formed from weighted combinations of simpler patterns (e.g., a protein may be found in both the endoplasmic reticulum and the Golgi complex).

Developing a Systematics

The first requirement for creating a systematics for subcellular location is a means of obtaining the set of all possible localization patterns (initially in one cell type but eventually in others). This can be accomplished by random tagging of all expressed genes (Jarvik et al. 1996; Rolls et al. 1999) and then collecting fluorescence microscope images showing the distribution of each tagged gene product. Criteria are then needed for deciding whether two proteins show the same localization pattern or whether each should be assigned to its own class. This could potentially be decided by imaging both proteins in a single cell (by labeling each with a different fluorescent probe), but the number of pairwise combinations of the estimated 10,000 to 100,000 proteins expressed in a single cell make this effectively impossible.

As an alternative, we suggest a heuristic procedure for exploring the space of possible localization patterns. This procedure starts by numerically describing all known protein localization patterns and then attempting to cluster the proteins into essentially non-overlapping groups. Each group is then be examined to determine whether any existing knowledge (i.e., from cell fractionation experiments or visual inspection of images) suggests that it should be split. Additional numeric image descriptors would then be sought to resolve the subpopulations and the process repeated until either all existing knowledge has been accounted for or until the limitations of fluorescence microscopy are reached. As will be discussed below, our results suggest that automated pattern analysis is more sensitive than human observation.

As alluded to above, one benefit of having a systematic scheme is that the location of known proteins can be more accurately described so that systems for predicting location from sequence can be better trained. New motifs responsible for localization may be discovered directly by such systems (by examining the decision rules of a successful classifier) or separately by using unsupervised learning programs such as MEME (Bailey and Elkan 1995) on subsets of proteins identified as having the same localization pattern.

There are additional benefits, however. Being able to quantitatively describe localization patterns such that all (or most) patterns can be distinguished provides a objective means for comparing images.

Our progress towards numerical description and classification of protein localization patterns is described in this paper. The overall goal of the work described below is to enable comparison and analysis of microscope images to become as automated, accepted, objective, reliable, and statistically-sound as comparison of protein and nucleotide sequences.

Image Datasets

We have followed a data-driven approach to design and selection of numerical features to describe fluorescence microscope images. In order to attempt to demonstrate the feasibility of classification of protein localization patterns, we created a database of images of five different subcellular patterns in Chinese hamster ovary cells (Boland et al. 1997; Boland et al. 1998). Based on the encouraging results obtained with this set, we then generated a collection of images of ten subcellular patterns in HeLa cells (Boland, M.V., and Murphy, R.F., submitted). This collection was designed to include all major organelles and to include pairs of similar patterns to enable testing of the sensitivity of various features and classifiers. Representative images for each pattern are shown in Figure 1. There is a high degree of similarity between the patterns of the two Golgi proteins (Fig. 1B,C), both of which a located in a tight structure near the nucleus. The patterns of the lysosomal and endosomal proteins (Fig. 1D,H) are somewhat similar, with both proteins being concentrated to one side of the nucleus but also showing punctate staining through the cytoplasm. The endoplasmic reticulum (ER) and mitochondrial proteins (Fig. 1A,E) are both distributed around the nucleus in a fairly symmetric manner.

Numerical Features

Traditional pattern recognition applications (such as industrial parts recognition or military target recognition) most frequently make use of model-based approaches in which spatial models of the target are fitted to the image. However, many cell types (including HeLa cells) display a very large degree of heterogeneity both in overall cell morphology and in the distribution of organelles within cells. This is illustrated in Figure 2, which shows representative images of the pattern of transferrin receptor (primarily found in endosomes). Given this heterogeneity, we have chosen to describe protein locations using numerical features that capture essential characteristics of the patterns, rather than to try to develop spatial models of the pattern displayed by each class.

We have utilized three types of these numerical features to describe subcellular location patterns. All features were chosen to be invariant to position and rotation of the cell within the image, and to be insensitive to changes in the scale of intensity values. The first set is the Zernike moment features (Teague 1980) through order 12. These 49 features are calculated from the moments of each image relative to the Zernike polynomials, an orthogonal basis set defined on the unit circle. The calculation of these moments therefore requires that we supply the radius of a typical cell. Positional invariance is achieved by defining the origin of the unit circle to be the center of fluorescence of the image. The second set is the Haralick texture features (Haralick 1979). These 13 features can be related to intuitive descriptions of image texture, such as coarseness, complexity and isotropy. The procedures we used to calculate the Zernike and Haralick features have been described previously (Boland et al. 1998).

While we have shown that these features are valuable for classifying cell images, they are not sufficient by themselves to distinguish all cell patterns we have tested. For this reason, as well as the fact that it is difficult (especially for the Zernike moments) to relate them to the ways in which cell patterns are usually described by biologists, we have created a new set of 22 features derived from morphological and geometric analysis that correspond better to the terms used by biologists.

Thirteen of these features are derived from object finding and edge detection in combination with an automated thresholding method (Ridler and Calvard 1978). These include the number of objects, the ratio of the size of the largest object to the smallest object, the average distance of an object from the center of fluorescence, and the fraction of above-threshold pixels along an edge. Three features are derived from the convex hull of the fluorescence distribution (e.g., the fraction of the area of the convex hull that is occupied by above-threshold pixels).

Since biologists frequently use the cell nucleus as a landmark for assessing subcellular patterns (resulting in terms such as perinuclear), we have developed six features that relate each protein pattern to a superimposed DNA image of the same cell. These features include the average distance of protein-containing objects to the center of fluorescence of the DNA image (the center of the nucleus) and the fraction of protein fluorescence that colocalizes with DNA fluorescence (i.e., is inside the nucleus).

The calculation and properties of these features will be described in detail elsewhere (Murphy, R.F., Boland, M.V., and Valdes-Perez, R., in preparation).

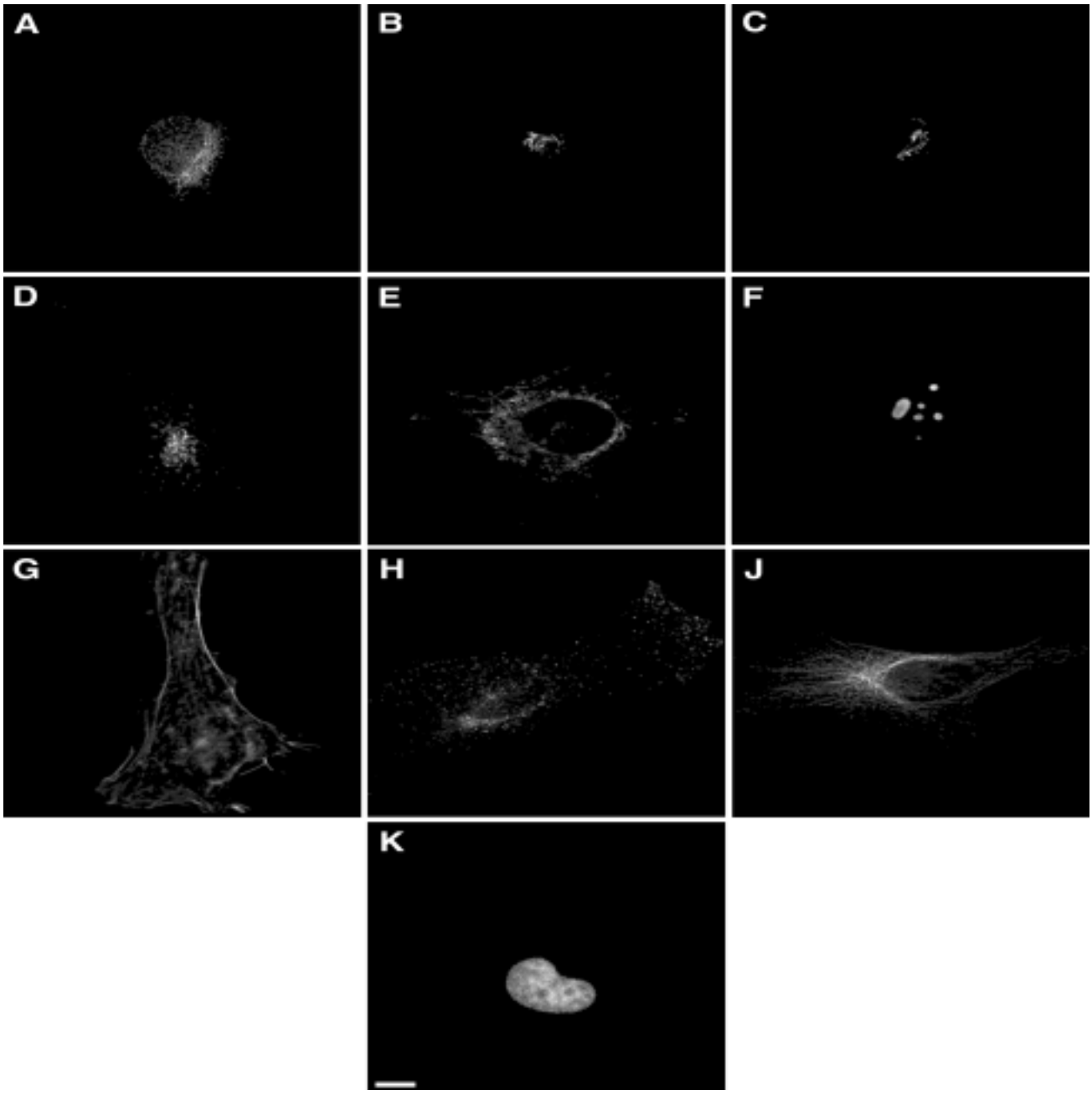

Figure 1. Representative images from the HeLa cell database. The most typical image was chosen from each class using the HTFR method (Markey, et al. 1999) which uses Haralick texture features to rank images based on their Mahalanobis distance from a robust estimate of the multivariate mean. Images are shown for cells labeled with antibodies against an endoplasmic reticulum protein (A), the Golgi protein giantin (B), the Golgi protein GPP130 (C), the lysosomal protein LAMP2 (D), a mitochondrial protein (E), the nucleolar protein nucleolin (F), transferrin receptor (primarily found in endosomes) (H), and the cytoskeletal protein tubulin (J). Images are also shown for filamentous actin labeled with rhodamine-phalloidin (G) and DNA labeled with DAPI (K). Scale bar = 10 μm.

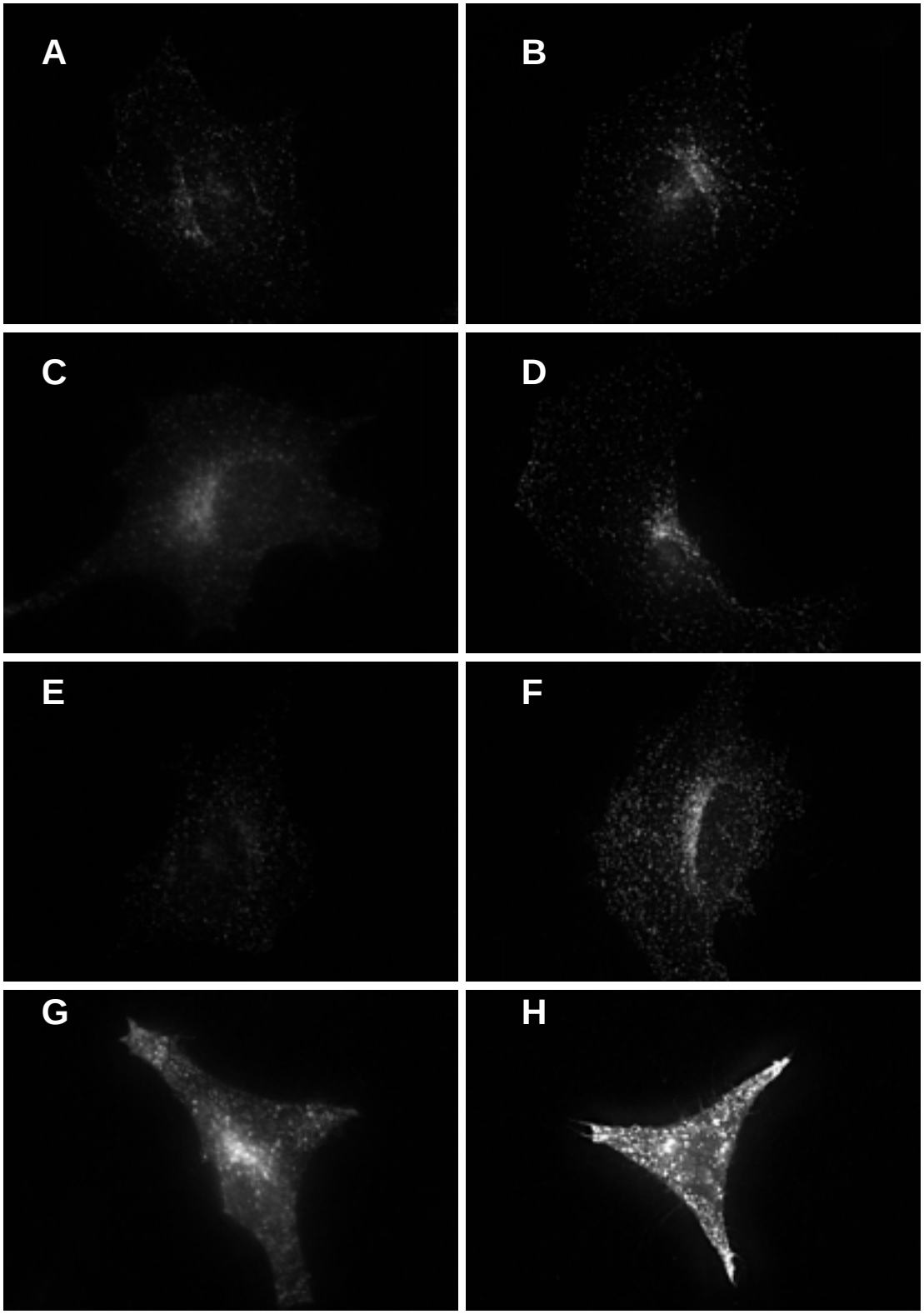

Figure 2. Variation in the transferrin receptor distribution in HeLa cells. The 91 images in the dataset were ranked in order of their typicality using a modification of the approach of Markey et al. (1999). The 84 features describing each image were used to calculate 16 principal components comprising 90% of the total variance. These were used to rank the images by their Mahalanobis distance from a robust estimate of the multivariate mean. The most (A) and least (H) typical image are shown, along with every 13th image in between (B-G). Note that the more typical images show a more rounded morphology. All images are shown at the same magnification and intensity scale.

Classification with All Features

In our initial work (Boland et al. 1998), we used Zernike features to describe the five patterns from CHO cells and demonstrated that neither linear discriminators nor classification trees were able to provide satisfactory results. However, a back-propagation neural network (BPNN) with a single hidden layer of 20 nodes was able to correctly classify an average of 87% of previously unseen images (compared to only 65% for a classification tree). Using Haralick features instead, a single hidden layer network was able to correctly classify 88%, indicating that both the Zernike and Haralick feature sets contain sufficient information to differentiate the five patterns in the CHO dataset. The results also indicated that even for a small number of relatively distinct patterns, the more complex decision boundaries available in a neural network provide better classification performance.

Initial analysis of the ability of a BPNN to resolve the ten classes in the HeLa cell dataset revealed that Zernike and Haralick features (either separately or combined) were unable to distinguish many of the classes. We therefore combined the Zernike, Haralick and the new localization features described above and tested BPNN with various numbers of hidden nodes. The average correct classification rate increased from 75% for 5 hidden nodes to 81% for 20 hidden nodes, but did not increase further for 30 hidden nodes. The number of hidden nodes was therefore set at 20 for further experiments.

A confusion matrix for such a network is shown in Table 2. The average correct classification rate was 81 ± 4.8% (mean ± 95% confidence interval). The most important conclusion from this table is that all ten classes can be resolved with better than 50% accuracy, including the three pairs of classes expected to be difficult to distinguish. We conclude not only that the features we have chosen capture essential aspects of the subcellular patterns, but that the majority of images of each class can be easily separated from the others in the feature space.

	Output of Classifier									
	DN	**ER**	**Gia**	**GP**	**LA**	**Mit**	**Nuc**	**Act**	**TfR**	**Tub**
DNA	**99**	1	0	0	0	0	0	0	0	0
ER	0	**86**	3	0	0	5	0	0	0	5
Giantin	0	0	**77**	19	0	1	2	0	1	0
GPP130	0	0	18	**78**	2	0	2	0	1	0
LAMP2	0	1	3	2	**73**	1	2	0	17	1
Mitoch.	0	9	2	0	4	**77**	0	0	2	6
Nucleolin	2	0	1	2	1	0	**94**	0	0	0
Actin	0	0	0	0	0	3	0	**91**	0	6
TfR	0	5	3	1	25	3	0	5	**55**	5
Tubulin	0	5	0	0	1	7	1	4	5	**77**

Table 2. Average percent correct classification for HeLa data over ten trials using a BPNN with a single hidden layer of 20 nodes and all 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 10% are shaded.

Given the large number of weights (1880) to be adjusted in this network, we considered the possibility that a simpler classifier could perform as well or better. We therefore tested *k* nearest neighbor classifiers with *k* ranging from 1 to 10. The best average classification rate was 68 ± 5.7%, indicating that (as with the CHO dataset) the more complex decision boundaries of a BPNN provide better classification.

We next considered whether *increasing* the complexity of the classifier could improve performance further. Two-hidden-layer (2HL) networks can form even more complex decision boundaries than one-hidden-layer (1HL) networks (Beale and Jackson 1990). While the decision boundaries of 1HL networks are restricted to convex hulls, 2HL networks can use any combinations of such convex hulls, providing arbitrary hulls. There is no further advantage to be gained by adding further layers because 2HL provide the maximum decision boundary complexity.

We first explored the performance of 2HL networks with various numbers of nodes in the first and second layer. As shown in Table 3, performance reached a plateau above approximately 15 nodes in each. The best performance, 79%, was obtained with 20 hidden nodes in each layer. Since this performance is comparable to that of the 1HL network (81%), we conclude tentatively that those observations in each class that can be distinguished from the other classes can be enclosed by a convex decision boundary and that no separable subpopulations exist (or at least any subpopulations are so subtle that more training samples would be required to define them).

	No. of nodes in 1st hidden layer					
	5	10	15	20	25	30
5	71	75	75	74	75	74
10	73	77	77	77	78	77
15	76	76	77	78	78	78
20	74	78	78	79	78	77
25	74	76	78	77	78	78
30	75	77	77	77	78	77

Table 3. Average percent correct classification for HeLa data using BPNN with two hidden layers of various sizes and all 84 features. Values shown are averages for test data over ten trials as described in Table 2. Each value has a 95% confidence interval of approximately 5%. Percentages at or above 78% are shaded.

Classification with Selected Features

Since improved classification performance is often obtained by reducing the number of input features, we sought to choose a subset of the 84 features that preserves their ability to resolve the classes. We used stepwise discriminant analysis (Jennrich 1977) for this purpose. An

True Class	Output of Classifier									
	DN	ER	Gia	GP	LA	Mit	Nuc	Act	TfR	Tub
DNA	**99**	1	0	0	0	0	0	0	0	0
ER	0	**87**	2	0	1	7	0	0	2	2
Giantin	0	1	**77**	19	1	0	1	0	1	0
GPP130	0	0	16	**78**	2	1	1	0	1	0
LAMP2	0	1	5	2	**74**	1	1	0	16	1
Mitoch.	0	8	2	0	2	**79**	0	1	2	6
Nucleolin	1	0	1	2	0	0	**95**	0	0	0
Actin	0	0	0	0	0	1	0	**96**	0	2
TfR	0	5	1	1	20	3	0	2	**62**	6
Tubulin	0	4	0	0	0	8	0	1	5	**81**

Table 4. Average percent correct classification for HeLa data over ten trials using a BPNN with a single hidden layer of 20 nodes and the 37 "best" features. Values shown are averages for test data from 10 trials as described in Table 2. Instances of confusion greater than 10% are shaded.

F-statistic was calculated for each feature to test the hypothesis that any difference in that feature's values between two classes could have arisen randomly even if those classes were drawn from the same population. The features for which this hypothesis could be rejected at the 0.0001 confidence level were selected, giving a set of 37. These included 11 of the 49 Zernike features, 12 of the 13 Haralick features, and 14 of the 22 biologically motivated features.

The ability of this set of features to resolve the ten classes was tested using a 1HL BPNN in a similar manner to that used for the full feature set. The average correct classification rate was 83 ± 4.6%. As shown in Table 4, there was an improvement of 5% in classification of actin images, 7% for transferrin receptor, and 4% for tubulin. Thus, overall classification accuracy was improved by 2% while reducing the size of the network by 50% (from 1880 weights to 940 weights).

No. of nodes in 2nd HL	No. of nodes in 1st hidden layer							
	5	10	15	20	25	30	35	40
5	79.0	81.2	81.4	81.8	81.6	81.6	80.2	81.1
10	79.3	82.5	83.2	83.0	83.1	83.2	83.0	83.7
15	81.1	82.3	83.2	83.6	83.2	83.7	83.8	83.5
20	80.3	83.3	83.4	82.7	83.2	83.2	83.2	84.0
25	79.8	82.9	82.8	83.4	83.2	84.0	83.1	83.8
30	81.5	83.2	83.4	83.6	83.6	84.4	83.7	83.8
35	79.2	83.3	83.4	83.2	83.6	83.6	83.5	83.3
40	80.0	82.6	83.1	83.3	83.6	83.3	83.4	83.7

Table 5. Average percent correct classification for HeLa data using BPNN with two hidden layers of various sizes at the 37 "best" features. Values shown are averages for test data over ten trials. Each value has a 95% confidence interval of approximately 5%. Percentages at or above 84% are shaded.

True Class	Output of Classifier									
	DN	ER	Gia	GP	LA	Mit	Nuc	Act	TfR	Tub
DNA	**98**	1	0	0	0	0	0	0	1	0
ER	0	**87**	2	0	1	5	0	0	1	3
Giantin	0	0	**84**	12	1	1	1	0	1	0
GPP130	0	0	20	**72**	1	2	3	0	2	0
LAMP2	0	0	5	1	**74**	0	3	0	15	2
Mitoch.	0	8	1	0	0	**81**	0	0	5	5
Nucleolin	0	0	0	1	1	0	**98**	0	0	0
Actin	0	0	0	0	0	1	0	**96**	1	3
TfR	0	2	2	0	18	4	0	2	**65**	7
Tubulin	0	2	1	0	2	7	0	1	5	**84**

Table 6. Average percent correct classification for HeLa data over ten trials using a BPNN with 30 nodes in two hidden layers and the 37 "best" features. Values shown are averages for test data from 10 trials. Instances of confusion greater than 10% are shaded.

As we did for the full feature set, we next explored the performance of 2HL networks with various numbers of nodes in the first and second layer. As shown in Table 5, performance reached a plateau above approximately 15 nodes in each. The best performance, 84.4%, was obtained with 30 hidden nodes in each layer. This is marginally higher than the performance of the 1HL network (83%), confirming our previous conclusion that most of the observations in each class that can be distinguished from the other classes can be enclosed by a convex decision boundary. Comparing the confusion matrix for this 2HL network (Table 6) with the results for the corresponding 1HL network (Table 4), we observe that the 2HL network performs 2-3% better on average for mitochondria, nucleolin, transferrin receptor and tubulin, 7% better for giantin, and 6% worse for gpp130. Note that these two classes that changed the most are the most difficult to distinguish even for experienced cell biologists. The combination of the feature subset and the 2HL classifier provides the best overall performance we have been able to achieve for classifying individual images.

Classification of Sets of Images

While the performance of the classifier is impressive given the subtlety of the differences between many of the classes, it is possible to improve that performance even more by considering how biologists frequently draw conclusions when using a microscope. This is by scanning over many fields of view to be able to integrate information from more than one cell. We can mimic this by attempting to classify *sets* of images that are all known to be taken from the same class (i.e., were collected from the same microscope slide). The principle is to use the single cell classifier on each image in a set and then classify the set as belonging to the class that receives a plurality (if any). For the 1HL classifier using the 37 "best" features, this approach results in an average classification accuracy of 98% (Boland, M.V., and Murphy, R.F., submitted). When sets that are

classified as unknown (no class received a plurality) are eliminated, the average increases to 99% and each class except transferrin receptor has an accuracy of 99% or higher (the accuracy for transferrin receptor is 93%). By using even larger sets, arbitrary accuracy can be achieved whenever it is possible to acquire multiple images from a population of cells expressing the same labeled protein.

Extension to Three Dimensions

The discussion above has been restricted to the analysis of two-dimensional images. Modern fluorescence microscopes (such as confocal, laser scanning microscopes) are able to acquire stacks of two-dimensional slices to form a full three-dimensional cell image. Since proteins in cells are distributed in three dimensions, not two, it will be important in the future to extend the approaches described here to images obtained by 3D microscopy. Such images obviously contain more information than 2D images, therefore presumably providing improved discrimination between similar classes of protein localization patterns. It is also worth noting that visual interpretation of 3D images is even more difficult than it is for 2D images, making computational analysis even more potentially valuable for 3D images. For cell types with extensive 3D structure, such as columnar epithelial cells, it will be especially crucial to use 3D pattern analysis because any single 2D section cannot be representative of the whole cell.

It is anticipated that the features we have used can all be extended to three dimensions with minimal difficulty. Three-dimensional analogues of the previously developed 2D features are not enough to capture all of the information in 3D images. Many cell types display a specific orientation with respect to a basement membrane (or the surface of a culture dish). Therefore, additional 3D features that reflect z-axis directionality (while still being invariant to rotation in the horizontal plane) will be needed.

Conclusions

The most important conclusion to be drawn from the work described above is that the set of features that we have developed captures the essential characteristics of subcellular structure. The validation of these features indicates that they can be used as a basis for defining localization classes, and then those classes can be correlated (where possible) with currently used (or new) descriptive terms.

Since many articles containing fluorescence micrographs are currently published in journals that have full-text, on-line versions, an interesting potential application of the single cell classifiers we have described is for cataloging (and perhaps reinterpreting) published images depicting subcellular location. We anticipate the possibility of creating a multimedia knowledge base with links to statements regarding localization derived from journal text and links to available images that support these statements (along with results of systematic analysis of these images). The creation of such a knowledge base can also be expected to aid improvement in description and classification efforts by providing new patterns for learning and testing.

The ability to describe cell images using validated features suggests additional applications beyond classification. These are based on the use of features to measure similarity between images. The first application is the objective selection of a representative image from a set. Representative images are useful for presentation, publication and for providing pictorial summaries of localization classes (e.g., on a summary page for an image database). We have described a system for choosing representative images using the Haralick and Zernike features (Markey et al. 1999) and established a web service (http://murphylab.web.cmu.edu/services/Typic) that will rank uploaded images in order of their "typicality." This service is currently being improved by incorporating the new features and including feature reduction methods (such as the approach in Figure 2).

A second application is the automated interpretation of imaging experiments. Cell biologists often wish to determine whether the distribution of a particular protein is altered by the expression of another protein or the addition of a drug. We have recently developed a system that can answer such questions with any desired statistical accuracy by comparing the distributions of the features for images with and without treatment (Roques, E.J and Murphy, R.F., in preparation).

The results from this application emphasize the apparently higher level of sensitivity to pattern changes of the automated approach we have described compared with human observers. Informal comparisons reveal that even trained cell biologists are unable to distinguish related patterns as well as our classification systems. It appears that while human observers can discern a great deal of information from individual images, they have difficulty retaining and comparing that information across large sets of fairly similar images. We are planning a more formal study to clarify this point.

In order to provide the initial framework for a systematics of protein localization, we plan to use our approach to generate a database for randomly-tagged proteins in HeLa cells (with links to existing protein sequence databases). The methods we have described should be useful to enable the systematic classification and cataloging of the subcellular location of expressed proteins in a variety of cell types and organisms.

Acknowledgments

We thank David Casasent, Raul Valdes-Perez, Mark Craven and Tom Mitchell for helpful discussions. The research discussed in this article was supported in part by research grant RPG-95-099-03-MGO from the American Cancer Society (R.F.M.), by NSF grant BIR-9217091, and

by NSF Science and Technology Center grant MCB-8920118. M.V.B. was supported by NIH training grant T32GM08208 and by NSF training grant BIR-9256343. M.V. was supported by a fellowship from the Center for Automated Learning and Discovery (NSF grant REC-9720374).

References

Bailey, T. L., and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning Journal*, 21:51-83.

Beale, R., and Jackson, T. 1990. *Neural Computing: An Introduction.*, Philadelphia: Hilger.

Boland, M. V., Markey, M. K., and Murphy, R. F. 1997. Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 594-597. Chicago, IL, USA

Boland, M. V., Markey, M. K., and Murphy, R. F. 1998. Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. *Cytometry*, 33:366-375.

Eisenhaber, F., and Bork, P. 1998. Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology*, 8:169-170.

Garrels, J. I. 1996. YPD--A database for the proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 24:46-49.

Haralick, R. M. 1979. Statistical and Structural Approaches to Texture. *Proceedings of the IEEE*, 67:786-804.

Horton, P., and Nakai, K. 1997. Better Prediction of Protein Cellular Localization Sites with the *k* Nearest Neighbors Classifier. *Intelligent Systems for Molecular Biology*, 5:147-152.

Jarvik, J. W., Adler, S. A., Telmer, C. A., Subramaniam, V., and Lopez, A. J. 1996. CD-Tagging: A New Approach to Gene and Protein Discovery and Analysis. *Biotechniques*, 20:896-904.

Jennrich, R. I. 1977. Stepwise Discriminant Analysis. *Statistical Methods for Digital Computers*, Vol. 3, K. Enslein, A. Ralston, and H. S. Wilf, eds., John Wiley & Sons, New York, 77-95.

Markey, M., Boland, M., and Murphy, R. 1999. Towards Objective Selection of Representative Microscope Images. *Biophysical Journal*, 76:2230-2237.

Nakai, K., and Horton, P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24:34-35.

Nakai, K., and Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14:897-911.

Ridler, T. W., and Calvard, S. 1978. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-8:630-632.

Rolls, M. M., Stein, P. A., Taylor, S. S., Ha, E., McKeon, F., and Rapoport, T. A. 1999. A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *Journal of Cell Biology*, 146:29-44.

Teague, M. R. 1980. Image Analysis via the General Theory of Moments. *Journal of the Optical Society of America*, 70:920-930.