BIOINFORMATICS

Vol. 17 no. 12 2001
Pages 1213–1223

A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells

*Michael V. Boland and Robert F. Murphy**

Center for Light Microscope Imaging and Biotechnology, Biomedical and Health Engineering Program, and Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Ave., Pittsburgh, PA 15213, USA

Received on March 9, 2001; revised on June 14, 2001; accepted on August 1, 2001

ABSTRACT

Motivation: Assessment of protein subcellular location is crucial to proteomics efforts since localization information provides a context for a protein's sequence, structure, and function. The work described below is the first to address the subcellular localization of proteins in a quantitative, comprehensive manner.

Results: Images for ten different subcellular patterns (including all major organelles) were collected using fluorescence microscopy. The patterns were described using a variety of numeric features, including Zernike moments, Haralick texture features, and a set of new features developed specifically for this purpose. To test the usefulness of these features, they were used to train a neural network classifier. The classifier was able to correctly recognize an average of 83% of previously unseen cells showing one of the ten patterns. The same classifier was then used to recognize previously unseen sets of homogeneously prepared cells with 98% accuracy.

Availability: Algorithms were implemented using the commercial products Matlab, S-Plus, and SAS, as well as some functions written in C. The scripts and source code generated for this work are available at http://murphylab.web.cmu.edu/software.

Contact: murphy@cmu.edu

INTRODUCTION

An important part of the characterization of a protein is the determination of the subcellular organelles or structures to which it localizes. This information is valuable because it provides a context for the protein's structure and function. For example, two proteins that are hypothesized (based on sequence similarity) to possess similar structure and function may in fact localize to different compartments within the cell and therefore be involved in distinct cellular processes.

*To whom correspondence should be addressed.

The most common method for determining subcellular location is interpretation of fluorescence microscope images, either of cells stained with monoclonal antibodies against a specific endogenous protein or of cells expressing a GFP-tagged protein from a transfected construct. Currently, the interpretation is performed visually by the investigator. Such subjective interpretations may be influenced by investigator bias (either conscious or unconscious), cannot be easily confirmed by other investigators, do not lend themselves to statistical analysis, and do not provide a systematic description that can be entered in databases.

An automated system for interpreting images of localization patterns would therefore have a number of advantages over current practice. These would include objectivity, reliability, and repeatability. Since we have found no prior work on the numerical analysis of protein localization patterns, we have worked to develop and test methods for *quantitatively* describing such patterns. To this end, we initially demonstrated the feasibility of creating an automated system to distinguish five subcellular patterns in Chinese hamster ovary cells (Boland *et al.*, 1998). When we attempted to apply the features used in that system to a larger number of patterns in HeLa cells, many of the patterns could not be distinguished. In the work described here, we developed new features and classification approaches to address more challenging questions: can all major classes of localization patterns (e.g. organelles) be distinguished by an automated system? How visually subtle can differences between patterns be such that they can still be distinguished by such a system? The work we describe will be immediately useful in a number of biotechnology applications, including pharmaceutical screening for drugs that affect protein location, large scale proteome characterization efforts, and microscope-based automated functional assays.

SYSTEMS AND METHODS

Immunofluorescence microscopy

HeLa cells were grown to sub-confluent levels on collagen-coated microscope coverslips, fixed in paraformaldehyde, and permeabilized with saponin. They were then incubated with one of eight monoclonal antibodies or rhodamine phalloidin (a label for filamentous actin). Monoclonal antibodies directed against an ER antigen (clone RFD6; DAKO, Carpinteria, CA, USA), the Golgi protein giantin (Linstedt and Hauri, 1993), the Golgi protein GPP130 (Linstedt *et al.*, 1997), the lysosomal protein LAMP2 (Mane *et al.*, 1989), a mitochondrial outer membrane protein (clone H6/C12; Serotec, Oxford, England), the nucleolar protein nucleolin (Deng *et al.*, 1996), transferrin receptor (clone 236-15 375; O.E.M. Concepts, Toms River, NJ, USA), and beta tubulin (clone 2-28-33; Sigma) were used as primary antibodies in separate labeling experiments. Working dilutions of antibody stock solutions were obtained by empirically optimizing for low background in the presence of adequate specific signal. Filamentous actin was labeled with 53 nM rhodamine phalloidin (Molecular Probes). Cells incubated with primary antibodies (except the anti-nucleolin antibody, which was directly conjugated with Cy3) were subsequently incubated with a Cy5 conjugated secondary antibody (Jackson Immunoresearch, West Grove, PA, USA). All cells were also labeled with the DNA intercalating dye DAPI. After fixation, permeabilization, and labeling, the coverslips were mounted on microscope slides. The coverslips were scanned manually using differential interference contrast microscopy to identify cells that were well spread and separated from their neighbors. The focus was adjusted, while viewing in DIC mode, until most cellular organelles appeared in focus. Two stacks of three images separated axially by 0.237 μm (above, at and below the focal plane chosen by DIC) were collected for each field of view. One of these stacks was collected for Cy5, Cy3 or rhodamine fluorescence and one stack for DAPI fluorescence.

Image processing

The out-of-focus component of the fluorescence in the central plane of each stack was reduced via nearest neighbor deconvolution (Agard *et al.*, 1989). The remaining background fluorescence (defined as the most common pixel value in the image) was then subtracted from the deconvolved image and small, isolated spots of fluorescence were removed with a majority filter (Pratt, 1991, p. 457)—the Matlab bwmorph function with the 'majority' option. This filter sets a given pixel to 1 if at least five of its immediate eight neighbors are 1 and to 0 otherwise. All pixels whose value was below a threshold chosen by an automated method (Ridler and Calvard, 1978) were set to zero, and single cells were isolated using a manually defined polygon. The intensity values for each cell image were scaled to the range 0–1 by dividing by the highest intensity value for that cell.

Zernike and Haralick features

Zernike moments and Haralick texture features were calculated from the processed protein localization images as previously described (Boland *et al.*, 1998). Briefly, Zernike moments (Zernike, 1934) are calculated using an orthogonal basis set, the Zernike polynomials, which are defined over the unit circle. For reference, plots of the Zernike polynomials are available at http://murphylab.web.cmu.edu/services/SLF. The amplitudes of these complex-valued moments were used as features in subsequent pattern recognition. Translation invariance was incorporated into the Zernike features by calculating them about the center of fluorescence of the image. The Haralick texture features (Haralick, 1979), on the other hand, are statistics calculated on the gray-level co-occurrence matrix derived from each image.

Subcellular location features

Additional sets of features designed to capture information about subcellular location were developed for this study. The first of these, termed Subcellular Location Features, set 1 (SLF1) contains 16 features that are calculated from the processed protein image only.

SLF1.1—the number of fluorescent objects in the image. Objects were identified by applying the Matlab bwlabel function to a binarized version of the processed image. The bwlabel function defines an object as a contiguous group of non-zero pixels in an eight-connected environment (i.e. a given pixel is adjacent to each of its eight neighbors) (Haralick and Shapiro, 1992, pp. 28–48). No restrictions were placed on the size of an object identified using this method.

SLF1.2—the Euler number of the image. The Matlab imfeature function was used to calculate the Euler number, the number of objects in the image minus the number of holes (Gonzalez and Woods, 1992, p. 505). A hole is defined as a contiguous group of zero-valued pixels contained entirely within an area of non-zero pixels. This feature is intended to distinguish reticular or mesh-like patterns from those that are more uniformly distributed.

SLF1.3—the average number of above-threshold pixels per object. The mean number of non-zero pixels per object was calculated for the binarized image. Along with some features below, this number was intended to capture information about the sizes of fluorescent objects in the cell.

SLF1.4—the variance of the number of above-threshold pixels per object. The variance of the number of non-zero pixels per object was also calculated. This feature was included to help quantify the homogeneity of fluorescent object sizes in the image.

SLF1.5—the ratio of the size of the largest object to the smallest. This was defined as the number of pixels in the largest object divided by the number of pixels in the smallest object. Like SLF1.4, this feature was included as a means of assessing the distribution of fluorescent object sizes.

SLF1.6—the average object distance to the cellular center of fluorescence. The Center Of Fluorescence (COF) of the whole cell was calculated and used to determine distances to the centers of fluorescence of each object in that cell. Centers of fluorescence were calculated as:

$$\bar{x}_c = \frac{\sum_x \sum_y x f(x, y)}{\sum_x \sum_y f(x, y)}, \quad \bar{y}_c = \frac{\sum_x \sum_y y f(x, y)}{\sum_x \sum_y f(x, y)}$$

where x and y are the coordinates of each pixel (in either the entire cell or a particular object), and $f(x, y)$ is the intensity of the pixel at (x, y). This feature provides information about how the individual fluorescent objects are distributed throughout the cell.

SLF1.7—the variance of object distances from the image COF. The variance was calculated using the COF determined for SLF1.6. As with SLF1.8, this feature is included to capture information about the distribution of objects around a central point.

SLF1.8—the ratio of the largest to the smallest object to image COF distance. This ratio was calculated as the distance from the image COF to the furthest object in the cell divided by the distance from the image COF to the closest object. This feature was also included to characterize the distribution of object distances from a central point.

SLF1.9—the fraction of the non-zero pixels in a cell that are along an edge. Edge detection was performed on each image using the Canny method (Canny, 1986) as implemented in the Matlab edge function. Canny's method calculates the gradient of the image using the derivative of a Gaussian filter. It then assigns edges to strong and weak categories. Weak edges are only included in the final output if they are connected to strong edges. This approach is less sensitive to noise in the image than other edge detection methods. The area of the binarized edge image was then divided by the area of the binarized cell image. In a biological sense, this feature is included to help distinguish proteins that localize along edges (i.e. the membrane of an organelle or along a filament or tubule) from those that do not.

SLF1.10—measure of edge gradient intensity homogeneity. Each image (**I**) was convolved separately with the kernels **N** and **W**

$$\mathbf{N} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

to find the intensity gradients in two orthogonal directions ($\mathbf{G_N} = \mathbf{I} \otimes \mathbf{N}$ and $\mathbf{G_W} = \mathbf{I} \otimes \mathbf{W}$). The intensity of the gradient at all points in the image was calculated using

$$A(x, y) = \sqrt{G_N^2(x, y) + G_W^2(x, y)}$$

and a four-bin histogram was calculated for the values in this edge intensity image. The final feature was the fraction of all values that fall in the first bin of this histogram. This feature was designed to capture the homogeneity of edge gradients. In other words, are the edges primarily 'steep' or more gradually sloping.

SLF1.11—measure of edge direction homogeneity 1. The edge direction gradient at each point in the image **G** was then calculated from the convolved images, $\mathbf{G_N}$ and $\mathbf{G_W}$, used in SLF1.10:

$$G(x, y) = \tan^{-1}\left(\frac{G_N(x, y)}{G_W(x, y)}\right).$$

The value of each pixel in the image **G** is therefore the direction (from $-\pi$ to π) of the intensity gradient at that point in the image, **I**. An eight-bin histogram was calculated using all of the values in the gradient image **G**. The final feature was calculated as the ratio of the largest to smallest value in the histogram. This feature was designed to capture the homogeneity of edge direction, i.e. are the edges primarily in one direction or are they more evenly distributed? Images with patterns containing edges oriented predominantly along a particular direction (some patterns of actin filaments, for example) result in edge gradient histograms in which a few bins will dominate. Histograms of edge direction are not completely insensitive to rotation because of quantization error. Because the edges in biological images are not as regularly oriented as those in images of man made patterns, it was decided to avoid the smoothing techniques previously described (Jain and Vailaya, 1996) so that any information in the histogram was not futher degraded by the smoothing operation.

SLF1.12—measure of edge direction homogeneity 2. The ratio of the largest to the next largest value in the eight-bin histogram used for SLF1.11 was calculated. This feature was included to overcome problems that may arise with values of SLF1.11 becoming very large when the minimum value of the histogram is small.

SLF1.13—measure of edge direction difference. For the eight-bin histogram used for SLF1.11, the difference between the bins for an angle and for that angle plus π was calculated by summing bins 1–4 and subtracting the sum of bins 5–8. This difference was normalized by the sum of all eight bins. This feature is intended to distinguish patterns in which there are parallel edges or in which the edge directions are uniformly distributed (i.e. the difference between the first four and last four bins of the histogram is small) from patterns in which the edges are primarily in one direction.

SLF1.14—the fraction of the convex hull area occupied by protein fluorescence. The convex hull of the protein localization image was calculated using the convhull function in Matlab and converted to a binary image. The area of the binarized protein image was then divided by the area of the convex hull image. This feature has been described as the 'transparency' of the image (Eakins *et al.*, 1998).

SLF1.15—the roundness of the convex hull. The roundness of an arbitrary shape is defined as $\frac{(\text{Perimeter})^2}{4\pi \cdot \text{Area}}$ (Sonka *et al.*, 1993, p. 227), which approaches 1 as the shape approaches a circle. We applied this calculation to the convex hull.

SLF1.16—the eccentricity of the convex hull. The eccentricity of the ellipse that is equivalent, based on second order moments, to the protein image convex hull was calculated using the following (from Prokop and Reeves, 1992):

$$\frac{\sqrt{(\text{Semimajor Axis})^2 - (\text{Semiminor Axis})^2}}{(\text{Semimajor Axis})},$$

where

$$\text{Semimajor Axis} = \sqrt{\frac{2\left[\mu_{20} + \mu_{02} + \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}\right]}{\mu_{00}}}$$

$$\text{Semiminor Axis} = \sqrt{\frac{2\left[\mu_{20} + \mu_{02} - \sqrt{(\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2}\right]}{\mu_{00}}}$$

and μ_{xy} are the central moments of the protein image convex hull. This feature is intended to distinguish patterns that are elongated from those that are more circular.

A second set of features, SLF2, was defined to include all of the features of SLF1 as well as six features calculated using both the processed protein image and the corresponding DNA image. The image of DNA distribution was included in the analysis because it provides a common reference point for each cell, and it may help to overcome issues related to the variability of cell size and shape.

SLF2.17—the average object distance from the COF of the DNA image. As for SLF1.6, the distances from a reference point of objects in the protein image are calculated. However, in this case the COF of the DNA image is used in place of the COF of the protein image.

SLF2.18—the variance of object distances from the DNA COF. This feature is analogous to SLF1.7 except that the DNA COF is used as the reference point.

SLF2.19—the ratio of the largest to the smallest object to DNA COF distance. This feature is analogous to SLF1.8 except that the DNA COF is used as the reference point.

SLF2.20—the distance between the protein COF and the DNA COF. The distance between the COF of a protein image and its corresponding DNA image is calculated. This feature was designed to capture information about how the protein is distributed relative to the nucleus.

SLF2.21—the ratio of the area occupied by protein to that occupied by DNA. The number of pixels in the binarized protein image is divided by the number of pixels in the binarized DNA image. This feature describes the area occupied by the protein distribution relative to the size of the nucleus.

SLF2.22—the fraction of the protein fluorescence that co-localizes with DNA. The fraction of pixels in the binarized protein image that overlap with pixels in the binarized DNA image. As with SLF2.20, this feature captures information about the distribution of the protein with respect to the nucleus.

Lastly, SLF3 was defined as the combination of SLF1 with the Zernike and Haralick features (a total of 78 features that can be calculated from a protein image only) and SLF4 was defined as the combination of SLF2 with the Zernike and Haralick features (a total of 84 features that are derived from a protein image and a corresponding DNA image).

Neural networks

Back-Propagation Neural Networks (BPNNs) were implemented using the Netlab (http://www.ncrg.aston.ac.uk/netlab/) scripts for Matlab. A fixed number of instances from each class were randomly assigned to the training (40 instances from each class), stop (20 instances from each class), and test (remaining 13–38 instances from each class) sets. The mean and standard deviation of each feature were calculated using the instances assigned to the training set. These values were then used to normalize the training data to have a mean of 0 and a variance of 1. In order to avoid biasing the classification system, the mean and standard deviation of the training data were also used to normalize the stop and test sets. This choice simulates the situation in which a previously trained

classifier is applied to data not available at the time of training (and which must therefore be converted using the same transformation as the training data).

The normalized training and stop sets were used to train a BPNN with a number of inputs equal to the number of features being evaluated, 20 hidden nodes, and 10 output nodes. The momentum and learning rate were 0.9 and 0.001, respectively. The target outputs of the network for each instance were defined to be 0.9 for the node representing the correct class of that instance and 0.1 for the other outputs. After each epoch of training, the stop data were passed through the network and a sum of squared error was calculated for the difference between the actual network outputs and the target outputs. When this error term for the stop data reached a minimum, training was halted. At that point, the test data were applied to the network and the outputs recorded. Starting with random assignment of the feature data, all of the steps above were repeated 10 times. The result was 10 networks, each created with a unique combination of training, stop and test data, and 10 corresponding sets of network output data.

Pairwise feature comparisons

To gain insight into the basis for distinguishing similar classes, all pairs of features in SLF2 were tested for their ability to discriminate a given pair of classes. This exercise was carried out only for the two pairs of classes that are most similar, giantin/gpp130 and transferrin receptor/LAMP2. The values for each pair of features for all observations in the two classes were used as both training and test data for the classify function of Matlab. This function implements a minimum Mahalanobis distance classifier for the case where the covariance matrices of the two classes are not assumed to be equal (Duda and Hart, 1973, p. 30). The percentage of images that were correctly classified was calculated, and the pair of features giving the highest percentage was then found. For this pair of features, the decision boundary (the line separating points classified into one class from those classified as the other class) was found as the solution to the equation

$$M(\mathbf{f}, \boldsymbol{\mu_1}, \mathbf{C}_1) = M(\mathbf{f}, \boldsymbol{\mu_2}, \mathbf{C}_2)$$

where $\mathbf{f}$ represents a feature vector (of length 2), $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ represent the mean feature vectors for the first and second classes, $\mathbf{C}_1$ and $\mathbf{C}_2$ represent the covariance matrices of the features for those classes, and M represents the Mahalanobis distance function. Since this approach is for visualization purposes only, we choose two features that can be plotted versus each other rather than three or more features. It is not intended to replace the more accurate BPNN. It is important to note also that using a minimum Mahalnobis distance classifier does not assume that the populations are multivariate Gaussian but only that the decision boundary is optimal only for the multivariate Gaussians with the same covariance matrices.

IMPLEMENTATION

Image collection

For our initial work on pattern classification, we created a collection of microscope images depicting five subcellular patterns in Chinese Hamster Ovary (CHO) cells (Boland *et al.*, 1998). Since more monoclonal antibodies against human proteins are available than against hamster proteins, we chose to create a new database of images of HeLa cells (a human cultured cell line) that included all major classes of subcellular structures. HeLa cells have the additional advantage for microscopy of being larger and better spread than CHO cells. Briefly, fluorescent dyes were targeted to nine specific proteins and DNA. Images depicting the localization of those dye molecules, and hence the localization of the target protein (or DNA) were then collected by fluorescence microscopy.

The antibodies chosen targeted the major classes of subcellular structures: the Endoplasmic Reticulum (ER), the Golgi complex, lysosomes, endosomes, mitochondria, the actin cytoskeleton, the tubulin cytoskeleton, nucleoli, and nuclei. Pairs of antibodies expected to produce patterns difficult to distinguish visually were purposely included. We anticipated that testing the ability of our numeric descriptors to distinguish these pairs would help to understand how successful these methods would be with patterns from *subcompartments* of organelles and with organelles possessing similar localization patterns.

Representative images selected from each of the ten classes of localization patterns using a systematic method (the HTFR typicality method, Markey *et al.*, 1999) are shown in Figure 1.

Feature extraction

Arguably the most important step in pattern recognition is the appropriate choice of numbers (features) to represent an image. Given that subconfluent, unpolarized cells on coverslips have arbitrary location and orientation, all features used to describe them should be invariant to the translation and rotation of cells within a field of view. Since a long term goal of this work is a system that is able to distinguish the localization of many proteins (not just the ten patterns used in this study), we therefore used two sets of 'general purpose' features that we previously showed were useful for distinguishing subcellular patterns (Boland *et al.*, 1998). Zernike moments (Teague, 1980) have found application in pattern recognition (Bailey and Mandyam, 1996; Khotanzad and Hong, 1990; Perantonis and Lisboa, 1992). We used Zernike moments up to degree 12, providing 49 numbers describing each image.

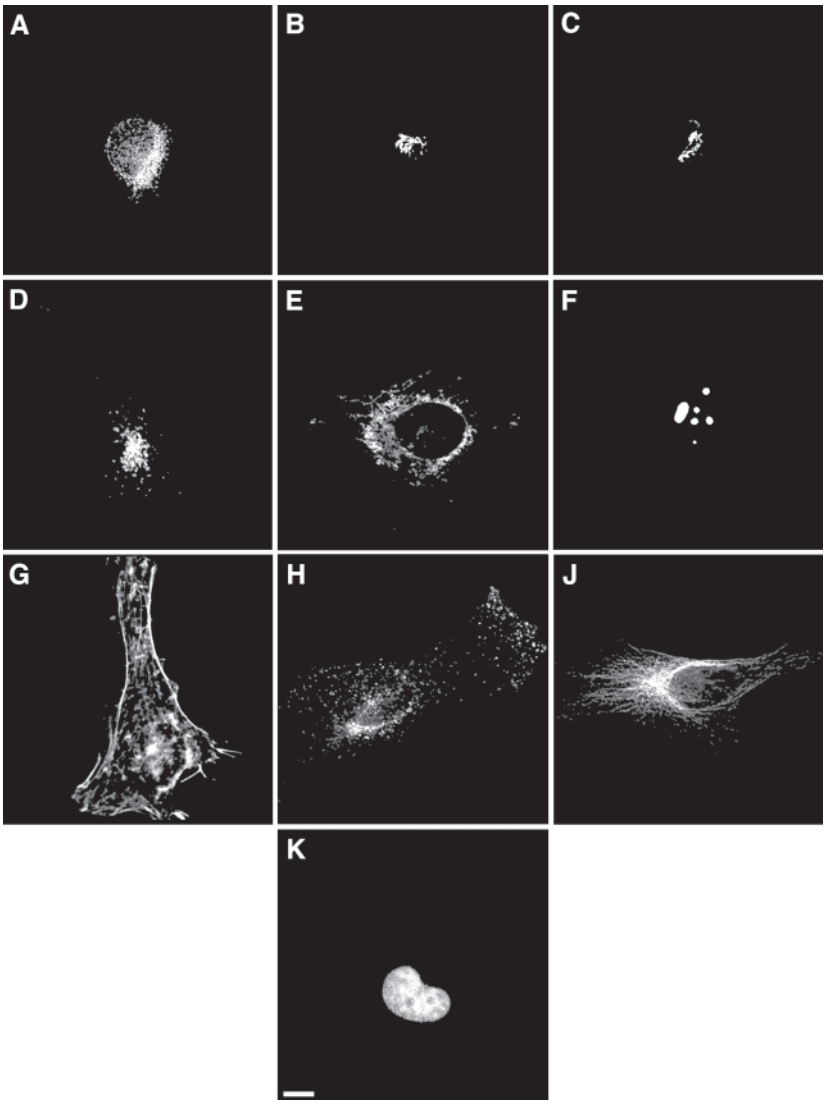

Fig. 1. Representative images from the classes used as input to the classification systems described in the text. These images have had background fluorescence subtracted and have had all pixels below a threshold set to 0. Images are shown for cells labeled with antibodies against an ER protein (A), 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 (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.

A fundamentally different set of features that measure image texture (Haralick, 1979) were also used. These features describe more intuitive aspects of the images (e.g. complexity, coarseness, isotropy, etc.) using statistics of the gray-level co-occurrence matrix for each image.

A new set of features was also developed using some features designed specifically for this problem and some features previously used for other pattern recognition and image processing applications. The goal for these features was to capture some of the criteria used by biologists to describe the localization of proteins. The first set of these SLF1, includes measures of object distance from the center of the cell (where an object is a contiguous group of fluorescent pixels and may represent all or part of an organelle), the distribution of object sizes, the degree to which the protein distribution overlaps the nucleus, the diffuseness of the localization pattern, the edge content of the image, and others. To help provide some standardization between cells, which are inherently heterogeneous in morphology, some features were designed to take advantage of a DNA image collected along with each protein localization image. The intent

Table 1. The 37 features selected from SLF4 using stepwise discriminant analysis to maximize discrimination between the ten classes of images used in this study. The features are shown in decreasing order of Wilks' λ statistic. This set is defined as SLF5

1.	SLF1.3: the average number of pixels per object
2.	$Z_{4,0}$
3.	SLF2.22: the fraction of the protein fluorescence that co-localizes with DNA
4.	$Z_{2,0}$
5.	Haralick information measure of correlation 1
6.	SLF1.6: the average object distance to the COF
7.	SLF1.2: the Euler number of the image
8.	Haralick sum entropy
9.	SLF1.14: the fraction of the convex hull occupied by protein fluorescence
10.	SLF1.9: the fraction of non-zero pixels that are along an edge
11.	SLF2.19: the ratio of the largest to the smallest object to DNA COF distance
12.	$Z_{8,0}$
13.	$Z_{12,2}$
14.	$Z_{12,0}$
15.	Haralick information measure of correlation 2
16.	Haralick correlation
17.	$Z_{7,1}$
18.	$Z_{4,2}$
19.	SLF1.5: the ratio of the size of the largest object to the smallest
20.	SLF1.11: the ratio of the largest to smallest value in a histogram of gradient direction
21.	SLF2.17: the average object distance from the DNA COF
22.	Haralick angular second moment
23.	Haralick contrast
24.	Haralick sum variance
25.	Haralick sum average
26.	SLF1.1: the number of fluorescent objects in the image
27.	Haralick difference variance
28.	SLF1.8: the ratio of the largest to smallest object—COF distance
29.	$Z_{10,0}$
30.	$Z_{1,1}$
31.	SLF1.7: the variance of object distances from the COF
32.	$Z_{11,1}$
33.	SLF1.4: the variances of the number of above-threshold pixels per object
34.	Haralick sum of squares
35.	Haralick difference entropy
36.	Haralick inverse difference moment
37.	$Z_{8,8}$

was to use the DNA pattern, which is fairly consistent among cells, as a common landmark to which the protein localization pattern could be referred. Feature set SLF2 was defined as the combination of the features in SLF1 with the six additional features that are defined in relation to the DNA distribution.

Feature subset selection

It is accepted in the pattern recognition community that simply adding more descriptive features to a system will not necessarily increase the ability of that system

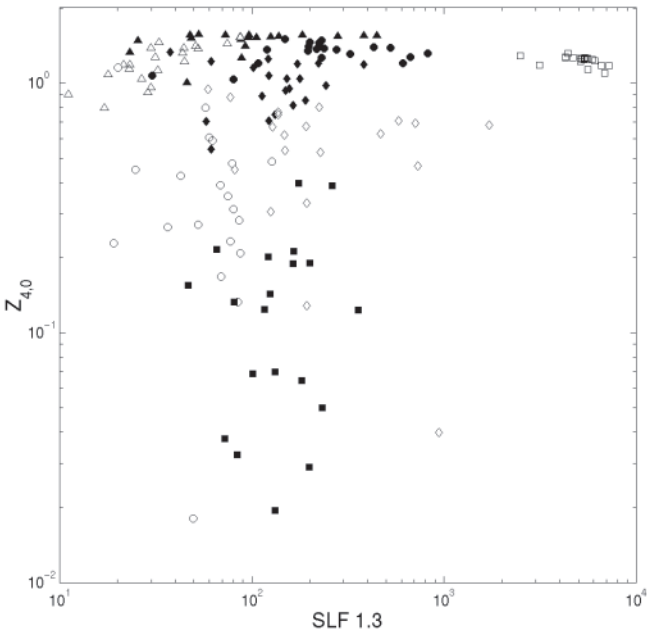

Fig. 2. Resolving power of the most discriminating features identified by stepwise discriminant analysis. A scatterplot displaying Zernike moment $Z_{4,0}$ versus SLF1.3 (average number of pixels per object) is shown for twenty observations each for DAPI (□), an ER protein (♦), a mitochondrial protein (○), giantin (▲), F-actin (■), tubulin (◇), nucleolin (●), and LAMP-2 (△). To simplify the plot, values for gpp130 and transferrin receptor are not shown (they are similar to giantin and LAMP-2, respectively). Note that many of the classes overlap but that most can be roughly distinguished using only these two features.

to correctly recognize patterns. In an attempt to optimize the dimensionality of the feature set, a subset of features was selected from the SLF4 feature set (the combination of Zernike, Haralick, and SLF2 features) via stepwise discriminant analysis (Jennrich, 1977) using the STEPDISC function of SAS (SAS Institute, Cary, NC, USA). This method uses Wilks' λ statistic to iteratively determine which features are best able to separate the classes from one another in the feature space. Since it is not possible to identify a subset of features that are optimal for classification without training and testing classifiers for all combinations of the input features, optimization of Wilks' λ was chosen as a reasonable alternative. The 37 (out of 84) features that were most statistically significant in terms of their ability to separate the ten classes identified using this approach are listed in Table 1. This set is referred to as SLF5 and was used in the classification phase of the work. A scatter plot for the two most distinguishing features is shown in Figure 2. While there is significant overlap in the distributions of these features for the various classes, the scatter plot gives

Table 2. Average performance of BPNNs for classifying previously unseen *single images* using the SLF5 feature set. The average rate of correct classification is 83 ± 4.6% (mean ± 95% confidence interval). The number of test samples per class is indicated in parentheses. This number of samples was randomly selected 10 times and classified by 10 different networks to generate this table. Not all rows sum to 100% due to rounding

True classification (no. samples)	Output of classifier									
	DNA (%)	ER (%)	Giantin (%)	GPP130 (%)	LAMP2 (%)	Mitochondria (%)	Nucleolin (%)	Actin (%)	TfR (%)	Tubulin (%)
DNA (87)	**99**	1	0	0	0	0	0	0	0	0
ER (86)	0	**87**	2	0	1	7	0	0	2	2
Giantin (87)	0	1	**77**	19	1	0	1	0	1	0
GPP130 (85)	0	0	16	**78**	2	1	1	0	1	0
LAMP2 (84)	0	1	5	2	**74**	1	1	0	16	1
Mitochondria (73)	0	8	2	0	2	**79**	0	1	2	6
Nucleolin (80)	1	0	1	2	0	0	**95**	0	0	0
Actin (98)	0	0	0	0	0	1	0	**96**	0	2
TfR (91)	0	5	1	1	20	3	0	2	**62**	6
Tubulin (91)	0	4	0	0	0	8	0	1	5	**81**

a rough indication that at least most of the classes are likely to be distinguishable using the SLF5 feature set.

Classification of single cells

A BPNN was chosen as a classifier primarily because of its ability to generate complex decision boundaries in a multidimensional feature space (Hornik *et al.*, 1989). Neural networks were chosen for use as classifiers after prior studies demonstrated the inferiority of other approaches including linear discriminant analysis, decision trees, and k-nearest neighbor classifiers (data not shown). A BPNN with a single hidden layer of 20 nodes was used to classify the ten classes of images described above. The choice of 20 hidden nodes was made by testing networks with 5–30 hidden nodes and assessing their ability to classify patterns (data not shown). The choice of training algorithm, back propagation with momentum, was essentially arbitrary but was intended to demonstrate the utility of a straightforward neural network in this particular application.

Forty samples were taken randomly from each class and their features were used to train the BPNN. Features from another 20 samples from each class were then collectively used to decide when to stop the training process. Finally, the features from the remaining 13–38 images from each class were classified using the trained network. This process, starting with random assignment of the training samples, was repeated 10 times to produce the confusion matrix in Table 2. An ideal classifier would produce a confusion matrix in which the diagonal elements were all 100% and all off-diagonal elements were 0%. The matrix in Table 2 is clearly not ideal, but most classes of images are well resolved from each other. The poorest performance is on the classes that were expected to be easily confused, but even images in these classes were correctly classified at a rate of at least 62%. Confused pairs of patterns include those for LAMP2 and the transferrin receptor, and those for the ER and mitochondrial proteins. Surprisingly, the system was able to distinguish the patterns of the two Golgi proteins to a significant extent. The basis for this distinction will be discussed below.

An alternative method for reducing the dimensionality of the original feature set is to calculate principal components that capture a specified fraction of the total variance. To test this approach, we first converted all features to zero mean and unit variance and then calculated principal components. The first 7 principal components captured approximately 68% of the total variance while the first 32 captured approximately 95%. When BPNN classifiers were trained and tested using either the first 7 or 32 principal components, the average correct classification rate was lower (73%) than for SLF5 (83%). We conclude that stepwise discriminant analysis, while known to be a suboptimal method, performs better than principal component calculation in our case.

Classification of populations

The results for classification of single cells based on protein localization patterns are very good given the high degree of heterogeneity within the individual classes, but are not as impressive as pattern recognition results from other fields in which classification rates approach 100%. While the systematic approach to description of protein localization described here is valuable even if classification accuracies cannot be improved beyond those in Table 2, there are applications of this work in which one would want the classification accuracies to be as high as possible. The primary example is experiments involving screening for cells expressing a particular protein localization pattern. Specifically, it is common for an investigator to conduct an experiment

Table 3. Average performance of BPNNs for classifying previously unseen *sets of ten images* using the SLF5 feature set. Each set was assigned a single classification based on the class to which a plurality of its members were assigned by the BPNNs whose performances are summarized in Table 2. The 10 networks trained to generate Table 2 were each tested on 1000 sets of 10 images, with plurality rule at the output, to generate this table. The numbers in parentheses are calculated using only those sets not classified as unknown. The average rate of correct classification for all trials is 98% and the average for sets that were not classified as unknown (numbers in parentheses) is 99%

True classification	Output of classifier										
	DNA (%)	ER (%)	Giantin (%)	GPP130 (%)	LAMP2 (%)	Mitochondria (%)	Nucleolin (%)	Actin (%)	TfR (%)	Tubulin (%)	Unknown (%)
DNA	**100** (**100**)	0	0	0	0	0	0	0	0	0	0
ER	0	**100** (**100**)	0	0	0	0	0	0	0	0	0
Giantin	0	0	**98** (**99.5**)	0	0	0	0	0	0	0	1
GPP130	0	0	0	**99** (**99.7**)	0	0	0	0	0	0	1
LAMP2	0	0	0	0	**97** (**99**)	0	0	0	1	0	2
Mitochondria	0	0	0	0	0	**100** (**100**)	0	0	0	0	0
Nucleolin	0	0	0	0	0	0	**100** (**100**)	0	0	0	0
Actin	0	0	0	0	0	0	0	**100** (**100**)	0	0	0
TfR	0	0	0	0	6	0	0	0	**88** (**93**)	0	6
Tubulin	0	0	0	0	0	0	0	0	0	**99.9** (**100**)	0

using many populations of cells where each of those populations has been grown under different conditions. The goal may then be, for example, to distinguish those populations in which a particular protein is found in the Golgi from those in which that protein is in the ER.

For this purpose, improvements in classification can be achieved by assigning a single classification to a *set* (or population) of homogeneously prepared cells. Groups of cells that have been subject to the same preparation procedures (i.e. they were in the same culture dish throughout the experiment) can be assumed to belong to the same class for the purposes of assessing protein localization. Classifying *populations* rather than *individual* cells parallels the practice of cell biologists, who often scan across many fields before drawing a conclusion.

To test this method of classifying populations experimentally, the same networks trained and tested for single cell classification (above) were used to classify random sets of ten images each drawn from a single class of the test data. That is, the ten images all depicted different instances of one of the localization patterns described above. The entire set of ten images was then assigned to the class to which a plurality of its constituents were assigned by the classifier. Sets for which no plurality existed were classified as 'unknown'. This procedure was repeated 1000 times (using different sets of randomly chosen images) for each of the ten neural networks trained using different permutations of the feature data. The accuracy of classification of the resulting 10 000 sets of ten images is shown in Table 3. Note that Table 3 includes correct classification rates for all sets derived from a given class as well as the correct classification rate for those sets not classified as unknown. As expected, the classification system performs better when we allow it to say 'I don't know' and avoid making a classification.

The most important result found in Table 3 is that the average classification rate (defined as the average of the diagonal elements) is much higher (98%) than that obtained for the single cell case (83%). Furthermore, there is little or no confusion between the pairs of classes that were problematic when looking at cells one at a time (giantin and GPP130, transferrin receptor and LAMP2). In fact, looking at the results shown in parentheses in Table 3 (sets for which a plurality existed, i.e. those not classified as 'unknown'), there is essentially *no* confusion between any of the classes except for transferrin receptor and LAMP2.

Basis of distinction between similar classes

The fact that the various classifiers we tested were able to distinguish image classes expected to be difficult to separate necessarily implies that decision boundaries can be drawn in the high-dimensional feature space that separate (at least to a large degree) each class. While statisticians or computer scientists may be satisfied by this knowledge, biologists may reasonably ask what features enable an automated system to separate classes that they may not be able to distinguish by eye. Although it is not possible to adequately describe or visualize the decision boundaries of a high-dimensional neural network, examination of scatter plots of pairs of features may provide insight into the basis for separation. We therefore searched for pairs of features that could provide the best discrimination between particular classes and generated scatter plots of these features (Figure 3). It can be seen in Figure 3a that the endosome and lysosome classes are somewhat distinguishable based on the number of fluorescent objects in each cell (with the lysosomal protein showing fewer objects) and based on the average distance of an object to the COF (with lysosomes being closer to the center). This finding recapitulates a common description of the difference between endosomal and lysosomal patterns (endosomes are more numerous and peripheral). Similarly, Figure 3b shows that the distributions of the two Golgi proteins can be partially distinguished based on their relationship to nuclear DNA and on their shape. Giantin's distribution is more circular than gpp130's (at least as estimated by the convex hull) and it overlaps with the nucleus somewhat more than the distribution of gpp130. Of course, this overlap is expected to arise from Golgi elements either above or below, rather than inside, the nucleus. The biological significance of the distinction between these Golgi protein distributions remains to be determined.

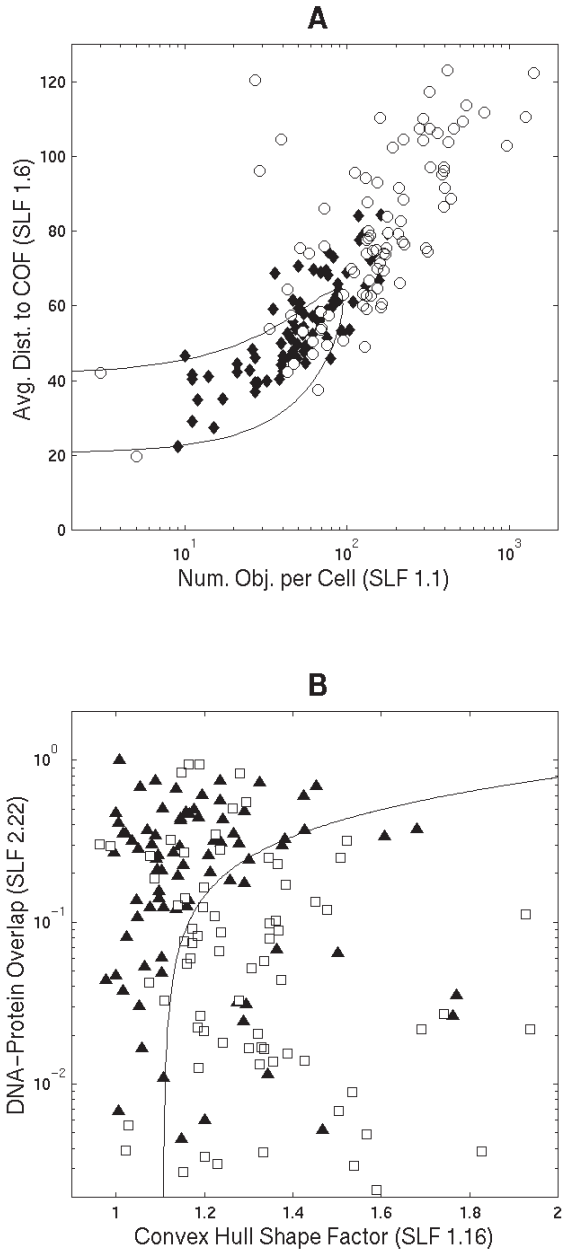

Fig. 3. Partial basis for distinguishing between similar classes. All pairs of features in SLF2 were tested for their ability to discriminate similar classes using a minimum-Mahalanobis-distance classifier (with non-equal covariance matrices for the two classes). (a) A scatterplot of SLF1.6 versus SLF1.1 is shown for images of transferrin receptor (♦) and LAMP-2 (○). The decision boundary for the classifier is also shown; 74.9% of the images are correctly classified using this boundary. While the best performance was actually obtained using SLF1.6 with SLF1.12 (77.7% correct classification), SLF1.6 and SLF1.1 are shown since their basis for distinguishing the two classes is more understandable from a biological perspective. (b) A scatterplot for the pair of features best able to discriminate giantin (♦) and gpp130 (□) is shown. The decision boundary shown correctly classifies 75.6% of the images.

DISCUSSION AND CONCLUSIONS

We have described an approach to quantitative description of protein localization patterns and demonstrated that this approach can produce classifiers that can reliably distinguish the patterns of all major organelles in HeLa cells. It is worth noting that the approach should work equally well on proteins that move between organelles and those that remain largely in a single organelle, since it is the *steady state* pattern, which includes contributions from all organelles weighted by the average fraction of time that the protein spends in each, which is being analyzed. Protein movement between organelles adds additional components to the overall steady state pattern beyond those of the individual organelles themselves. Quantitative analysis of the localization of proteins provides an objective method of describing proteins that is complementary to those that currently exist (e.g. amino

acid sequence, hydrophobicity, functional motifs, etc.) Ultimately, it will be possible to quantify the degree of similarity in the localization of two proteins, just as it is now possible to quantitatively describe the degree of similarity between two amino acid sequences. A benefit of such quantitative analysis will be the ability to obtain and archive novel information about new or existing proteins; a list of proteins with the same or similar localization characteristics, for instance. These techniques form an ideal adjunct to methods for randomly tagging expressed proteins (Jarvik *et al.*, 1996; Rolls *et al.*, 1999).

In addition, automated screening of microscope images is becoming an increasingly important tool in a variety of fields, including biology and pharmacology (Giuliano and Taylor, 1998). As an example, pharmaceutical companies have a large arsenal of compounds that are potentially marketable drugs. It is a significant effort to identify those few that have a desired effect on a system (e.g. those compounds that prevent translocation of a transcription factor to the nucleus). Techniques like those described here, for the automated screening of protein localization patterns, are potentially useful in this process.

ACKNOWLEDGEMENTS

We thank Drs David Casasent, Raúl Valdés-Pérez and Frederick Lanni for helpful discussions, Dr Adam Linstedt for the donation of antibodies, and Mr Meel Velliste for critical reading of this manuscript. The research discussed in this article was supported in part by research grant RPG-95-099-03-MGO from the American Cancer Society, by NSF grant BIR-9217091, by NSF Science and Technology Center grant MCB-8920118, and by NIH grant R33 CA83219. M.V.B. was supported by NIH training grant T32GM08208 and by NSF training grant BIR-9256343.

REFERENCES

Agard,D.A., Hiraoka,Y., Shaw,P. and Sedat,J.W. (1989) Fluorescence microscopy in three-dimensions. *Meth. Cell Biol.*, **30**, 353–377.

Bailey,R.R. and Mandyam,S. (1996) Orthogonal moment features for use with parametric and non-parametric classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI **18**, 389–399.

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.

Canny,J. (1986) A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI **8**, 679–698.

Deng,J.S., Ballou,B. and Hofmeister,J.K. (1996) Internalization of anti-nucleolin antibody into viable HEp-2 cells. *Mol. Biol. Rep.*, **23**, 191–195.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Eakins,J.P., Boardman,J.M. and Graham,M.E. (1998) Similarity retrieval of trademark images. *IEEE Multimedia*, **5**, 53–63.

Giuliano,K.A. and Taylor,D.L. (1998) Fluorescent-protein biosensors: new tools for drug discovery. *Trends Biotechnol.*, **16**, 135–140.

Gonzalez,R.C. and Woods,R.C. (1992) *Digital Image Processing*. Addison-Wesley, Reading, MA.

Haralick,R.M. (1979) Statistical and structural approaches to texture. *Proc. IEEE*, **67**, 786–804.

Haralick,R.M. and Shapiro,L.G. (1992) *Computer and Robot Vision*. Addison-Wesley, Reading, MA.

Hornik,K., Stinchcombe,M. and White,H. (1989) Multilayer feedforward networks are universal approximators. *Neural Netw.*, **2**, 359–366.

Jain,A.K. and Vailaya,A. (1996) Image retrieval using color and shape. *Pattern Recognit.*, **29**, 1233–1244.

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. In Enslein,K., Ralston,A. and Wilf,H.S. (eds), *Statistical Methods for Digital Computers*, Vol. 3, Wiley, New York, pp. 77–95.

Khotanzad,A. and Hong,Y.H. (1990) Invariant image recognition by Zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI **12**, 489–497.

Linstedt,A.D. and Hauri,H.P. (1993) Giantin, a novel conserved Golgi membrane protein containing a cytoplasmic domain of at least 350 kDa. *Mol. Biol. Cell*, **4**, 679–693.

Linstedt,A.D., Mehta,A., Suhan,J., Reggio,H. and Hauri,H.P. (1997) Sequence and overexpression of GPP130/GIMPc: evidence for saturable pH-sensitive targeting of a type II early Golgi membrane protein. *Mol. Biol. Cell*, **8**, 1073–1087.

Mane,S.M., Marzella,L., Bainton,D.F., Holt,V.K., Cha,Y., Hildreth,J.E. and August,J.T. (1989) Purification and characterization of human lysosomal membrane glycoproteins. *Arch. Biochem. Biophys.*, **268**, 360–378.

Markey,M.K., Boland,M.V. and Murphy,R.F. (1999) Towards objective selection of representative microscope images. *Biophys. J.*, **76**, 2230–2237.

Perantonis,S.J. and Lisboa,P.J.G. (1992) Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers. *IEEE Trans. Neural Netw.*, **3**, 241–251.

Pratt,W.K. (1991) *Digital Image Processing*. Wiley, New York.

Prokop,R.J. and Reeves,A.P. (1992) A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP, Graph. Models Image Process.*, **54**, 438–460.

Ridler,T.W. and Calvard,S. (1978) Picture thresholding using an iterative selection method. *IEEE Trans. Syst. Man Cybern.*, 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. *J. Cell Biol.*, **146**, 29–44.

Sonka,M., Hlvac,V. and Boyle,R. (1993) *Imaging Processing, Analysis and Machine Vision*. Chapman and Hall, London.

Teague,M.R. (1980) Image analysis via the general theory of moments. *J. Opt. Soc. Am.*, **70**, 920–930.

Zernike,F. (1934) Beugungstheorie des Schneidenverfarhens und seiner Verbesserten Form, der Phasenkontrastmethode. *Physica*, **1**, 689–704.