Automated Interpretation of Protein Subcellular Location Patterns

Implications for Early Cancer Detection and Assessment

ROBERT F. MURPHY

Departments of Biological Sciences and Biomedical Engineering, and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

ABSTRACT: Fluorescence microscopy is a powerful tool for analyzing the subcellular distributions of proteins, but that power has not been fully utilized because most analysis of those distributions has been done by visual examination. This limitation can be overcome using automated pattern recognition methods widely used in other fields. This article summarizes work demonstrating that automated systems can recognize the patterns of major organelles in both two- and three-dimensional images of cultured cells, and that these systems can distinguish similar patterns better than visual examination. The basis for these systems are sets of Subcellular Location Features that capture the essence of subcellular patterns without being sensitive to the extensive variation that occurs in the size, shape, and orientation of cells in microscope images. These features can also be used to make sensitive, statistical comparisons of the distribution of a protein between two conditions, such as in the presence and absence of a drug. The possible use of automated pattern analysis methods for improving detection of abnormal cells in cancerous or precancerous tissues is also discussed.

KEYWORDS: Subcellular Location Features; protein localization; fluorescence microscopy; pattern recognition; location proteomics; image similarity; protein distribution comparison

LOCATION PROTEOMICS

Basic research in biology has been revolutionized by the advent of ***genomics***, defined as the study of entire genomes rather than individual genes. As the sequences of complete genomes (and lists of suspected genes making up those genomes) have become available, the focus of much biological research has shifted from ***genomics*** to ***proteomics*** in order to understand the behavior and function of all proteins and the roles they play in development and disease. Most proteomics efforts to date have

Address for correspondence: Robert F. Murphy, Departments of Biological Sciences and Biomedical Engineering, and Center for Automated Learning and Discovery, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213. Voice: 412-268-3480; fax: 412-268-6571.
murphy@cmu.edu

**Ann. N.Y. Acad. Sci. 1020: 124–131 (2004). © 2004 New York Academy of Sciences.
doi: 10.1196/annals.1310.013**

focused on methods for determining protein *sequence*, *structure*, *abundance*, and *interactions*. Far less attention has been paid to determining and understanding the ***locations*** of protein within cells, although knowledge of the subcellular location of a protein is critical to understanding how it functions. The primary method by which information has been obtained about the organelles and other subcellular structures that contain a specific protein is by labeling that protein with a fluorescence probe (e.g., using a monoclonal antibody) and collecting images of cells using a fluorescence microscope. The main reason for the absence of prior systematic, large-scale efforts to determine subcellular location for all proteins has been the difficulty of automatically and quantitatively describing subcellular location in cells with varying sizes and shapes.

A major goal in my group in the past few years has therefore been to perform ***automated interpretation of fluorescence microscope images*** depicting the subcellular distribution of proteins.[1–7] While the primary motivation behind this work has been to enable the new field of ***location proteomics***, the work also has potential applications in cancer detection, assessment, and treatment.

This chapter will therefore review previous work demonstrating that changes in the subcellular distributions of proteins and organelles can be recognized in fluorescence microscope images in a fully automated manner.

DEVELOPMENT OF SUBCELLULAR LOCATION FEATURES AND CLASSIFICATION OF SUBCELLULAR PATTERNS

The most critical component of our work to date has been the development of sets of numerical features that capture the essence of subcellular distributions without being overly sensitive to the position or rotation of a cell within an image.[1,3,6] We have used these features to create automated *classifiers* that can recognize the patterns of all major subcellular structures in 2D images.[3,6]

The input was a collection of images of HeLa cells that were labeled with antibodies against protein markers for various organelles. Examples of the images used in these studies are shown in FIGURE 1. We specifically included markers whose distributions are quite similar: the proteins giantin and GPP130 are both found primarily in the Golgi apparatus, and the patterns of LAMP2 (primarily lysosomal) and transferrin receptor (TfR, primarily endosomal) are difficult to distinguish visually.

The general pattern recognition problem is to learn the patterns present in two or more *classes* of image, where each class is known to differ from the others by at least one descriptor external to the image (called a *label*), such that the class of new images not used in the learning can be predicted correctly. The accuracy of prediction for new images is usually assessed by dividing any available *labeled* images into a *training* set used to train the classifier and a *test* set used to evaluate performance by comparing the class predicted by the classifier to the known class. There are two basic approaches to recognizing patterns in images. The first involves learning a model of the distribution of the pixels in each class so that the model can be compared to the pixel values in new images. The second involves describing the images using numerical features and learning rules to associate the feature values to the classes.

For recognizing protein patterns, the variability of cell size, shape, and orientation within the microscope field, and the variability in the number, position, and

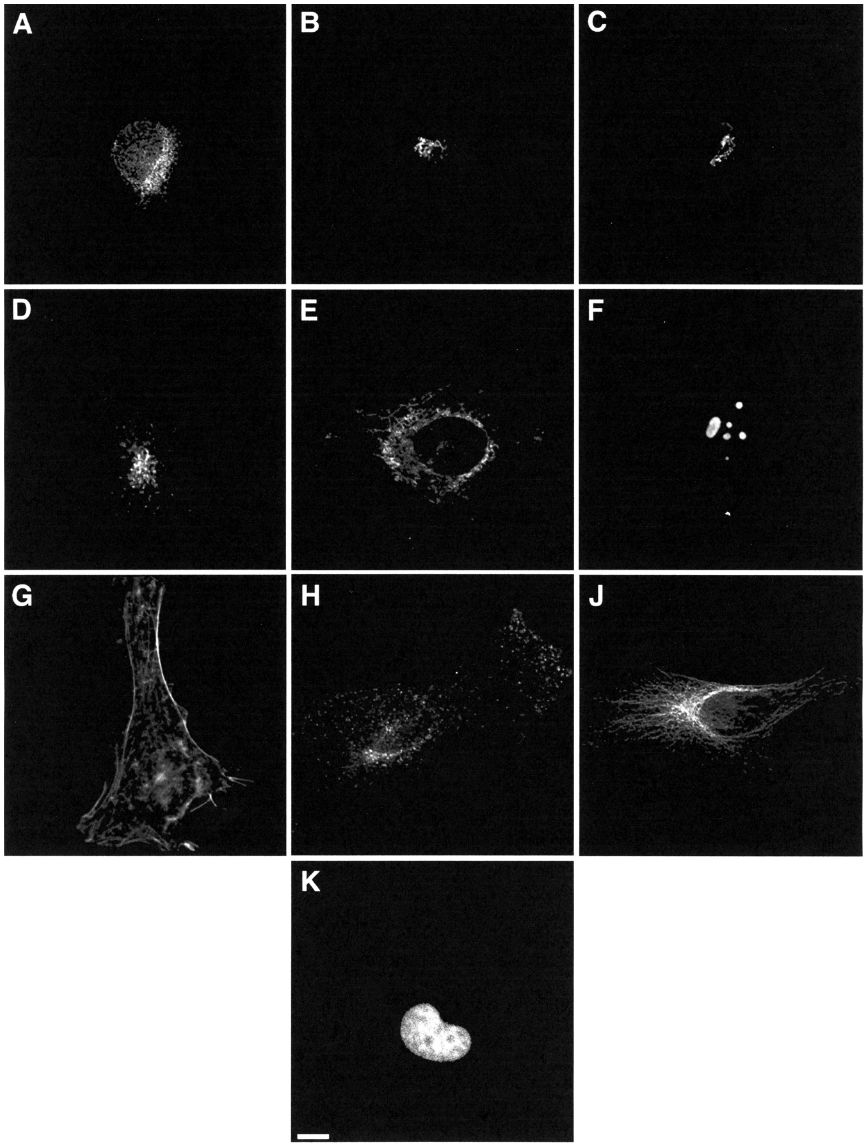

FIGURE 1. Representative images from the 2D HeLa cell data set described in the text. These images have had background fluorescence subtracted and have had all pixels below an automatically chosen threshold set to 0. Images are shown for HeLa cells labeled with antibodies against an ER protein **(A)**, 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. (Reprinted from ref. 3.)

orientation of organelles within each cell, make it difficult to use model-based approaches. We have therefore evaluated a number of types of numerical features for describing the patterns in cell images. These have been described in detail previously[3,7] and are briefly summarized here. Zernike moment features describe the overall pattern in a cell by measuring the degree to which the pattern matches a set of radially symmetric functions (the Zernike polynomials). Haralick texture features describe the frequency with which particular pixel values are found adjacent to other pixel values. Morphological features describe the properties of objects derived from thresholding the image (e.g., average object size). Edge features describe the distribution of edges (regions of sharply varying intensity) in an image (e.g., the fraction of fluorescence found along an edge). Last, hull features describe the pattern relative to the convex hull of the image, which connects the outermost set of above-threshold pixels.

We have used these features in various combinations to analyze cell images. To facilitate referring to a specific feature or combination of features, we have described a nomenclature for these Subcellular Location Features, or SLF. Sets of SLF are referred to by a set name (e.g., SLF3), and individual features are referred to by a set name and the number of the feature within that set (e.g., SLF3.7).

Obviously, the quality of classification results depends critically on the quality of the features used. For many classification approaches, the presence of uninformative features (features that have similar values for all classes) or the presence of redundant features (features whose values are correlated with those of other features) can complicate the learning task sufficiently so that poorer results are obtained than would have been obtained with a smaller set of informative, nonredundant features. One approach to creating such a set is to only describe images using features known to meet this criterion, which is often very difficult. A second approach is to describe each image using many features (some of which may be redundant or noninformative) and then use a method that automatically identifies which features best distinguish the classes being analyzed. There are many such methods and we have evaluated a number of them in the context of subcellular pattern analysis.[8] The best results were obtained with Stepwise Discriminant Analysis (SDA) and we have defined some SLF sets as the results of applying SDA to a larger SLF set. A description of each feature and set can be found at http://murphylab.web.cmu.edu/services/SLF/.

Results for one of our automated classification systems are shown in TABLE 1 in the form of a *confusion matrix* that tabulates how often images of a known class (shown in the row headings) are placed by the classifier in each predicted class (shown in the column headings). As can be seen, the classifier can distinguish all classes (including the two Golgi proteins) with an accuracy over 70%.

An important finding of our initial work on the HeLa data set[3] was that an automated system could recognize subtle differences in protein patterns that are not readily distinguishable by visual examination. To confirm that these patterns were indeed difficult to distinguish visually, we tested the ability of a human observer to learn this task.[7] The results are shown in TABLE 2. While our automated classifiers can distinguish the Golgi proteins giantin and GPP130 with an average accuracy of 75% (TABLE 1), a human observer, even after training until no further improvement occurred, had an average accuracy of only 50% (which is what is expected for random guessing). Both the automated system and visual examination had a similar overall accuracy when the two proteins were combined and considered as a single Golgi class.

TABLE 1. Confusion matrix for classification of images from the 2D HeLa data set combined with a parallel DNA image

	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
Mitochon.	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**

NOTE: The SLF13 feature set was used with a BPNN with a single layer of 20 hidden units over 10 cross-validation trials. The number of images in each predicted class is shown as a percentage of the number of test images for each known class (averaged across the 10 cross-validation trials). The average correct classification rate was 88% (91% when the two Golgi proteins, giantin and GPP130, are considered as a single class). Data from ref. 7.

TABLE 2. Confusion matrix for visual classification of images from the 2D HeLa data set

	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
Mitochon.	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**

NOTE: The number of images in each predicted class is shown as a percentage of the number of images for each known class (the results are from the last round of testing, after the classification accuracy had reached a maximum). The average correct classification rate was 83% (92% when the two Golgi proteins, giantin and GPP130, are considered as a single class). Data from ref. 7.

The above results are for classifying single cells, but even better performance can be obtained if we assume that all of the cells on a given slide show the same pattern. With this assumption, we can form small sets of cells from the same set, classify each individually, and then choose as the prediction for the whole set whichever class had the most cells. We have shown that such a "plurality voting" scheme can use a classifier with an average accuracy of 83% to classify sets of ten cells with an average accuracy greater than 98%.[3]

The improvement over visual examination demonstrated above for our automated systems on 2D images might be expected to be even greater for the analysis of 3D images, given the difficulty of visualizing and remembering complex patterns in more than two dimensions. To test this hypothesis, we collected a data set of 3D HeLa images covering the same patterns in the 2D data set.[5] Using only morphological features, these images could be classified with an average accuracy of 91%. This was about 5% better than that obtained for 2D classification using just the central slice from each 3D image.[5]

COMPARISON OF CELL POPULATIONS

The work described above addresses the assignment of cell images to defined classes, such as organelles. An equally important problem frequently addressed by fluorescence microscopy is determining whether the pattern of a protein changes in response to some treatment (such as the addition of a drug). More generally, this problem can be described as determining whether two sets of images represent statistically different patterns. Since the SLF contain sufficient information about subcellular patterns to allow those patterns to be accurately classified, it is reasonable to expect that they can be used to measure changes in those patterns as well. We have thus developed a system (called SImEC for Statistical Imaging Experiment Comparator) that performs rigorous statistical comparison of image sets.[4]

SImEC begins by converting image sets into a matrix in which each row represents a cell image and the columns contain the values of the chosen SLF set. The statistical question is then whether it is likely, at a given confidence level, that the matrices for the two sets could have resulted from images drawn from the same set. This hypothesis can be tested using the Hotelling T^2 test, which yields an F statistic with two degrees of freedom: the number of features and the combined number of images in the two sets minus the number of features. If the F statistic for two sets is greater than the critical F value for those degrees of freedom at the chosen confidence level, the hypothesis that the sets are drawn from the same population can be rejected. TABLE 3 shows the F statistics for all pairwise comparisons between the ten classes in the 2D HeLa cell data set. Not unexpectedly given that all of these classes can be distinguished by a classifier, the results indicate that all ten classes are statistically different at the 95% confidence level. To test that the test does not falsely identify all sets as different, random subsets drawn from the same class were compared at the same confidence level. Over repeated trials, approximately 95% of the randomly drawn subsets were considered to be the same (as expected). The conclusion is that the SLF can be used to create a statically sound method for comparing subcellular protein distributions.

TABLE 3. Pairwise comparison of classes from the 2D HeLa data set using SImEC

Class	No. of images	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Phal	TfR
DNA	87									
ER	86	90.6								
Giantin	87	138.9	49.9							
GPP130	85	154.1	51.3	2.6						
LAMP2	84	92.3	22.6	11.7	11.6					
Mitochon.	73	179.2	11.0	56.1	61.6	17.4				
Nucleolin	73	91.3	60.3	18.7	17.1	20.0	67.0			
Actin	98	523.5	58.1	374.2	358.2	127.4	17.0	274.2		
TfR	91	101.3	8.6	19.1	17.5	3.1	9.0	30.3	26.4	
Tubulin	91	185.5	12.5	97.3	102.4	31.3	8.0	100.5	21.4	6.5

NOTE: The values shown are *F* values from the T^2 test for the comparison of each class with each other class. Larger *F* values indicate that the two classes are more dissimilar. To determine whether two classes differ at a particular confidence level, the *F* value is compared to the critical *F* value for that confidence level. The critical values of the *F* distribution for a 95% confidence level range from 1.42 to 1.45 for the comparisons shown here (the critical value depends on the total number of images in the comparison and the number of features being used). Since all *F* values shown in the table are greater than this, all classes can be considered to be distinguishable from each other with 95% confidence. Note that the lowest *F* values were observed for the comparisons of giantin with GPP130 and of transferrin receptor with LAMP2. The highest *F* values were seen for pairs that are very different, such as for the DNA distribution compared with any of the others. Data from ref. 4.

IMPLICATIONS FOR IMPROVED DETECTION OF CANCEROUS AND PRECANCEROUS TISSUE

It is becoming increasingly clear that what may be largely normal-appearing tissue in the vicinity of skin (and other) cancers may be precancerous to a sufficient degree that recurrence at that site is likely. For nonmelanoma skin cancers, the most common current approach to surgery is to remove tissue until pathology indicates that the margins of the removed tissue are clear. For melanoma, additional tissue is removed until a 1- to 3-cm margin beyond the tumor is created. A difficulty with these approaches is that current methods cannot always determine whether the margins are indeed fully normal tissue.

Fluorescent probe staining has the potential to provide a dramatic increase in sensitivity and accuracy over traditional pathology stains. Probes that may be useful include antibodies against proteins known to localize in specific organelles, antibodies against proteins implicated in oncogenesis, or dyes that stain specific organelles or biochemical processes. However, a significant current limitation in the use of fluorescence microscopy for pathology is that fluorescence microscope images are difficult to interpret because the structural context visible with traditional stains is absent. One possible solution to this problem is the use of automated image analysis methods such as those described here. The prior work has been carried out on model systems consisting of cultured cells grown on coverslips, and thus one task to be accomplished is to extend them to images of cells in intact tissues. The next step is to identify proteins whose subcellular patterns change at various stages during the development of malignancies in a particular tissue (if they exist). This knowledge can potentially be used to screen for abnormalities (e.g., in biopsies) and to assess the stage or risk for a given abnormality.

ACKNOWLEDGMENTS

I thank John Kirkwood for helpful discussions on potential applications of our work to skin cancer detection. The original research described above was supported in part by Research Grant No. RPG-95-099-03-MGO from the American Cancer Society; Grant No. 99-295 from the Rockefeller Brothers Fund Charles E. Culpeper Biomedical Pilot Initiative; NSF Grant Nos. BIR-9217091, MCB-8920118, and BIR-9256343; NIH Grant Nos. T32 GM08208 and R33 CA83219; and a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund.

REFERENCES

1. Boland, M.V., M.K. Markey & R.F. Murphy. 1998. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry **33:** 366–375.
2. Murphy, R.F., M.V. Boland & M. Velliste. 2000. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. Proc. Int. Conf. Intell. Syst. Mol. Biol. **8:** 251–259.
3. Boland, M.V. & R.F. Murphy. 2001. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics **17:** 1213–1223.
4. Roques, E.J.S. & R.F. Murphy. 2002. Objective evaluation of differences in protein subcellular distribution. Traffic **3:** 61–65.
5. Velliste, M. & R.F. Murphy. 2002. Automated determination of protein subcellular locations from 3D fluorescence microscope images. *In* 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002), pp. 867–870.
6. Murphy, R.F., M. Velliste & G. Porreca. 2002. Robust classification of subcellular location patterns in fluorescence microscope images. *In* 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP 12), pp. 67–76.
7. Murphy, R.F., M. Velliste & G. Porreca. 2003. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J. VLSI Sig. Proc. **35:** 311–321.
8. Huang, K., M. Velliste & R.F. Murphy. 2003. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. Proc. SPIE **4962:** 307–318.