

AUTOMATED DETERMINATION OF PROTEIN SUBCELLULAR LOCATIONS FROM 3D FLUORESCENCE MICROSCOPE IMAGES

Meel Velliste and Robert F. Murphy

Biomedical and Health Engineering Program and Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A., murphy@cmu.edu

ABSTRACT

Knowing the subcellular location of a protein is critical to a full understanding of its function, and automated, objective methods for assigning locations are needed as part of the characterization process for the thousands of proteins expressed in each cell type. Fluorescence microscopy is the most common method used for determining subcellular location, and we have previously described automated systems that can recognize all major subcellular structures in 2D fluorescence microscope images. Here we show that 2D pattern recognition accuracy is dependent on the choice of the vertical position of the 2D slice through the cell and that classification of protein location patterns in 3D images results in higher accuracy than in 2D. In particular, automated analysis of 3D images provides excellent distinction between two Golgi proteins whose patterns are indistinguishable by visual examination.

1. INTRODUCTION

The understanding of protein function in mammalian cells is critically dependent upon knowledge of protein subcellular location because the function of a protein is affected by the biochemical environment of the compartment in which it is located. Fluorescence microscopy is the method of choice for obtaining protein subcellular location information in large-scale protein discovery experiments that utilize various gene tagging methods [1-3]. In order to make sense of the image data generated by such approaches, automated interpretation methods are needed because protein location patterns are highly complex. Due to extensive cell-to-cell variability of subcellular patterns it is not possible to compare images pixel by pixel. Therefore automated analysis of protein subcellular location must rely on concise numerical descriptors of the patterns in the images. We have previously developed numerical features (termed SLF for Subcellular Location Features) computed from 2D fluorescence microscope images [4]. We have shown the SLF to accurately represent the complexity in such images by using them successfully for automated classification of

protein location patterns [4-6], statistical comparison of imaging experiments [7] and objective choice of representative images [8]. The classifiers developed were shown to be capable of distinguishing all major protein subcellular location patterns and to be able to distinguish between pairs of very similar patterns indistinguishable by eye.

It has become common in biological research to collect 3D images of cells using optical sectioning techniques such as laser scanning microscopy. Having shown the feasibility of using numerical features to describe and interpret protein location patterns in 2D images we asked whether the same approach could be applied to 3D images of cells and perhaps more importantly, whether there would be any advantage to doing so. One might expect that 2D images would not capture sufficient information about protein location in some cell types (referred to by biologists as *polarized* cells) that exhibit a specific orientation in three dimensions. For example in epithelial cells (e.g. those forming the boundary between the gastrointestinal tract and the bloodstream), the apical surface typically has a different composition from that on the basal and lateral surfaces. One would anticipate that in this kind of a cell type it may be necessary to use 3D pattern recognition methods because 2D images can only capture information from one of these surfaces. Unpolarized cells on the other hand are sufficiently flat that a 2D image (i.e. a single optical section through the cell) can capture most, but perhaps not all of the protein location information. However, even for unpolarized cells, 3D images may contain additional information beyond that in 2D images. For example, even though F-actin in HeLa cells is found throughout the cell, it is often more concentrated above the nucleus than below it. The opposite is true for tubulin, which is preferentially located near the bottom of the cell.

2. RESULTS

2.1. 3D Image Datasets

As a first step toward evaluating the feasibility of 3D subcellular pattern recognition, a database of 3D images of ten different subcellular patterns with 50 to 52 images per class was created. To maintain the maximum

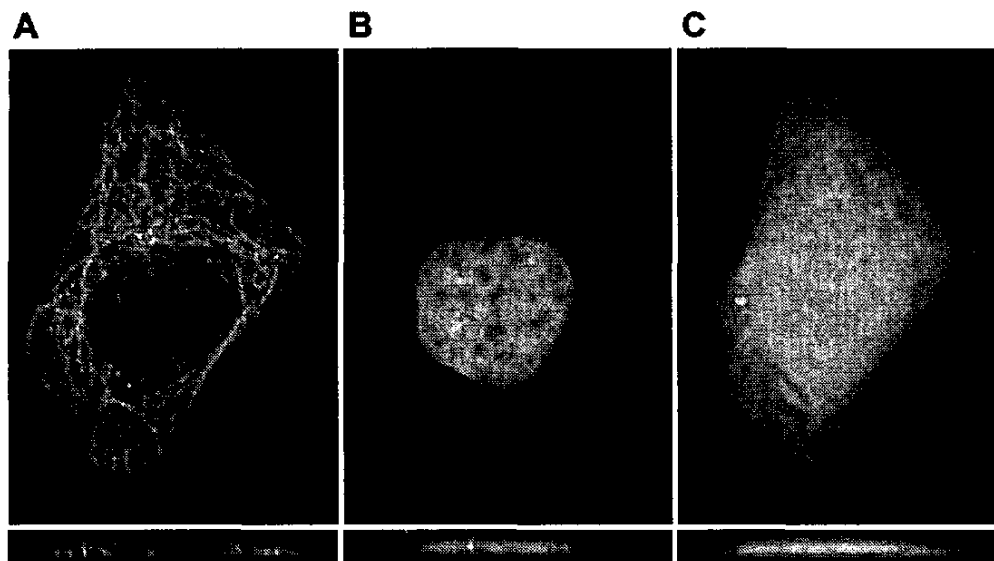


Figure 1. An example image from the 3D data set collected with a laser scanning confocal microscope. Each image in the data set consisted of three channels corresponding to a specific protein label (A) which is tubulin in this example, a DNA label (B), and an all-protein label (C). The images shown represent horizontal (top) and vertical (bottom) slices through the middle of the cell, chosen from the full 3D images so as to intersect the center of fluorescence (COF) of the specific-protein channel.

comparability to previous 2D work [4], the same ten classes of probes and the same cell line (HeLa) were used. Images were acquired using a three-laser confocal laser scanning microscope capable of imaging three independent channels of fluorescence. Each specimen was labeled with three different fluorescent probes (Figure 1) that bind to: a specific organelle or structure (e.g., tubulin for microtubules), DNA, and all cellular proteins. The DNA probe was used to provide a common frame of reference for protein distribution in the cell, allowing some numeric features to be calculated relative to the DNA image. The total protein distribution was used for automated segmentation of individual cells in fields containing more than one cell. The segmentation was performed using a watershed algorithm and using the DNA channel to define seed regions (because there is exactly one nucleus per cell). After discarding partial cell images at the edges of the images, the segmentation resulted in generating between 50 and 58 individual cell images per class. As with the 2D HeLa cell images described earlier [4,6] this image set was designed to include most major organelles and to include closely related patterns (Golgi proteins giantin and gpp130) to test the sensitivity of the features.

2.2. Features for 3D Classification

The SLF used in previous 2D work [4,6] consist of three major subsets of features: those based on image texture,

those resulting from decomposition using polynomial moments, and those derived from morphological and geometric image analysis. The last of these (SLF2) had been found to be the most useful single subset. Therefore the initial effort to extend the SLF features to work with 3D images was begun by implementing a 3D version of SLF2. These features were based on the division of the fluorescence intensity distribution in each image into objects. In 2D, an object had been defined as a contiguous group of above-threshold pixels in an 8-connected environment, where diagonally touching pixels as well as horizontally and vertically touching pixels were considered connected. In 3D, an analogous definition of object was used: a group of contiguous, above-threshold voxels in a 26-connected environment. First, 14 of the SLF2 features (SLF1.1-1.8 and SLF2.17-2.22) were converted to their 3D equivalents by simply replacing area with volume and 2D Euclidean distance with 3D Euclidean distance in the definition of those features (note that SLF2.17-2.22 require a parallel DNA image). Second, in order to exploit the inherently greater amount of information present in 3D images as compared to 2D images, some extra features were defined. The most obvious advantage of using 3D images is being able to analyze the vertical distribution of the protein. The vertical (basal or apical) position of a protein within a cell is often functionally important while its horizontal ("left" or "right") location does not have any significance. This is especially true of polarized epithelial cells, for instance,

True Class	Output of Classifier									
	DN	ER	Gia	gpp	LA	Mit	Nuc	Act	TfR	Tub
DNA	99	0	0	0	0	0	0	0	0	0
ER	0	89	0	0	0	0	0	2	2	7
Giantin	0	0	90	3	2	4	1	0	0	0
gpp130	0	0	5	81	9	0	0	0	4	0
LAMP2	0	0	1	4	90	2	0	1	2	0
Mitoch.	0	1	1	0	0	96	0	1	1	0
Nucleolin	1	0	0	0	0	0	98	0	0	0
Actin	0	2	0	0	0	2	0	92	3	0
TfR	0	1	0	0	4	3	0	2	85	5
Tubulin	0	5	0	0	0	0	0	0	4	91

Table 1. Classification results for 3D confocal images using the 28 3D-SLF9 features. The BPNN classifiers achieved an overall accuracy of 91% across all 10 classes.

where a protein localizing to the apical membrane can have a totally different function from a protein localizing to the basal membrane. To a lesser extent this is also true for unpolarized cells. The 3D-SLF2 features so far defined contain features that measure the positioning of the protein distribution relative to the nucleus or relative to the center of fluorescence (COF) of the protein distribution itself. Yet none of these features distinguished vertical distance from horizontal distance or apical location from basal. Therefore 14 extra features were created by defining two new features for each of the seven features that were based on a measure of distance (3D-SLF2.6, 2.7, 2.8, 2.17, 2.18, 2.19 and 2.20). This was done by separating out the vertical (z-dimension) and horizontal (combined x- and y-dimensions) components of distance in the definition of each of those features. The resulting set of 28 3D features was termed 3D-SLF9.

2.3. Classification with the New 3D Features

The 28 3D-SLF9 features were computed for the ten classes of 3D images, and single-hidden-layer back-propagation neural networks (BPNN) with 20 hidden nodes and 10 output nodes were trained with these features. To estimate classification accuracy on unseen images 50 cross-validation trials were used. For each trial, from the 50 to 58 images in each class 35 were randomly chosen for the *training set*, 5 for the *test set* and the remainder used for the *stop set*. The networks were trained until the error on the *stop set* reached a minimum. The overall accuracy on the *test set* images, averaged over all trials, was 91% (Table 1). Note that the two Golgi proteins, Giantin and gpp130 can be distinguished with high accuracy. This 3D classification result is better than the 83-84% classification accuracy previously achieved on 2D images [4,6]. However, those 2D images were collected with a different type of microscope at a different resolution, the fluorescent labeling was done using a different protocol, and the features used for classification

True Class	Output of Classifier									
	DN	ER	Gia	gpp	LA	Mit	Nuc	Act	TfR	Tub
DNA	100	0	0	0	0	0	0	0	0	0
ER	0	85	0	0	0	0	0	1	2	12
Giantin	0	0	80	8	7	3	2	0	0	0
gpp130	0	0	6	81	5	0	7	0	0	0
LAMP2	0	0	4	2	86	1	0	0	6	0
Mitoch.	0	0	0	0	2	94	0	2	2	0
Nucleolin	1	0	0	0	0	0	98	0	0	0
Actin	0	1	0	0	0	3	0	91	3	2
TfR	0	1	0	0	9	8	0	8	69	4
Tubulin	0	14	0	0	0	2	0	0	6	78

Table 2. Classification results for 2D optical sections chosen from the 3D stacks so as to include the COF of the specific-protein channel. With the 14 SLF2 features, the BPNN classifiers achieved an overall accuracy of 86% across all 10 classes.

were different. This means that in trying to evaluate whether 3D classification has any advantage over 2D classification one cannot simply compare the results to the previous 2D experiments.

Instead, since 3D images consist of stacks of 2D images, it is possible to make a comparable set of 2D images by selecting appropriate horizontal slices from the 3D images. A potential problem with this approach is that 2D classification accuracy may be dependent on the choice of the vertical position of the slice. This has been addressed previously by choosing the "most informative" slice defined as the one containing the largest amount of above-threshold pixels [9]. Here a systematic approach was taken to determine which slice to use based on simulating how an investigator would choose the slice when using a microscope to acquire 2D images. It was hypothesized that the investigator would choose a slice containing the center of either the DNA distribution or the protein distribution. Six different 2D image sets were created with slices containing either the center of fluorescence (COF) of the nucleus or the COF of the specific-protein distribution, as well as two slices above and below each of the centers. Each of these methods might be a valid way to simulate how an investigator would choose the vertical position when using a microscope to acquire 2D images of cells. 2D slice classification was performed using the 14 SLF2 features that had been used as the basis for 3D-SLF9, and using the same training/testing procedures as used for 3D images. Of the methods tried, choosing the slice containing the COF of the specific-protein distribution was empirically found to give the best classification accuracy of 86% (Table 2). For comparison, the classification accuracy was 81% for the set consisting of images two slices above the COF of the specific-protein distribution and 84% for images two slices below the COF. When using the DNA distribution as a reference for choosing the slices, the accuracy was 80% for slices containing the COF of the

DNA distribution, 83% for images two slices above the COF and 85% for images two slices below the COF. This leads to the conclusion that 2D classification accuracy is dependent on the choice of the vertical position of the 2D slice through the specimen. Also, the results show that classification of full 3D images works significantly better than classification of 2D optical sections. After carefully reviewing and comparing the confusion matrices in Tables 1 and 2, it appears that the 3D classifier performed better than the 2D classifier for most classes. Notably, the 3D classification accuracies of Giantin, TfR and tubulin are improved by 10%, 16% and 13% respectively compared to the 2D results.

3. CONCLUSIONS

We have shown in previous work that protein subcellular locations can be determined automatically with reasonable accuracy from 2D fluorescence microscope images based on numeric descriptors. Danckaert et al. [9] have also described classifiers capable of determining protein subcellular location with reasonable accuracy from 2D optical section images selected from 3D confocal stacks. In this work we have shown that calculating features on the entire 3D image results in better classification accuracy. It should be noted that for most cases, 2D imaging is presently preferable from a practical perspective. This is due to the considerably longer time required to acquire 3D images and the fact that 2D classification gives reasonably accurate results. However, even though the acquisition time required for 3D images has until recently been a limiting factor, improvements in imaging technology such as the development of the spinning disc confocal microscope are likely to remove that limitation. One can therefore predict that 3D imaging will be the norm in the near future, especially since automated analysis of imaging experiments works better on 3D images.

Since the SLF features have been validated by using them to achieve good classification accuracy for subcellular location patterns, it is possible to use them as a basis for a systematics of protein subcellular location. By calculating measures of similarity between protein distributions, it will be possible to create for the first time a grouping of proteins into classes that share a single location pattern. It is also possible to use the SLF for other automated analyses of fluorescence microscope images, such as automated selection of representative images from a set [8] and rigorous statistical comparison of imaging experiments [7]. An important challenge for the future is to bring the methods for interpretation of protein location patterns that work well on individual cells to bear on tissue images.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant R33 CA83219 and by NSF Science and Technology Center grant MCB-8920118. We thank Dr. Simon Watkins for helpful discussions on and assistance with 3D imaging, Aaron C. Rising for help with collecting 3D images, and Kai Huang for helpful discussions.

REFERENCES

- [1] J. W. Jarvik, S. A. Adler, C. A. Telmer, V. Subramaniam, and A. J. Lopez, "CD-Tagging: A new approach to gene and protein discovery and analysis," *Biotechniques* 20, 896-904, 1996.
- [2] M. M. Rolls, P. A. Stein, S. S. Taylor, E. Ha, F. McKeon, and T. A. Rapoport, "A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein," *Journal of Cell Biology* 146, 29-44, 1999.
- [3] C. A. Telmer, P. B. Berget, B. Ballou, R. F. Murphy, and J. W. Jarvik, "Epitope Tagging Genomic DNA Using a CD-Tagging Tn10 Minitransposon," *Biotechniques* 32, 422-430, 2002.
- [4] M. V. Boland, and R. F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics* 17, 1213-1223, 2001.
- [5] M. V. Boland, M. K. Markey, and R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry* 33, 366-375, 1998.
- [6] R. F. Murphy, M. V. Boland, and M. Velliste, "Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images," *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* 251-259, 2000.
- [7] E. J. S. Roques, and R. F. Murphy, "Objective Evaluation of Differences in Protein Subcellular Distribution," *Traffic* 3, 61-65, 2002.
- [8] M. K. Markey, M. V. Boland, and R. F. Murphy, "Towards objective selection of representative microscope images," *Biophysical Journal* 76, 2230-2237, 1999.
- [9] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes, "Automated Recognition of Intracellular Organelles in Confocal Microscope Images," *Traffic* 3, 66-73, 2002.