# Automated Recognition System to Classify Subcellular Protein Localizations in Images of Different Cell Lines Acquired by Different Imaging Systems

YUH-SHOW TSAI,[1] I-FANG CHUNG,[2] JEREMY C. SIMPSON,[3] MEI-I LEE,[1] CHIA-CHENG HSIUNG,[1] TAI-YU CHIU,[4,5] LUNG-SEN KAO,[5,6,7] TE-CHENG CHIU,[7,8] CHIN-TENG LIN,[7,8] WEN-CHIEH LIN,[7,8] SHENG-FU LIANG,[9] AND CHUNG-CHIH LIN[4,5,7]*

[1]*Department of Biomedical Engineering, Chung Yuan Christian University, Jhongli, Taiwan, Republic of China*
[2]*Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan, Republic of China*
[3]*Department of Cell Biology/Biophysics, EMBL Heidelberg, 69117 Heidelberg, Germany*
[4]*Department of BioMedical Sciences, Chung Shan Medical University, Taichung, Taiwan, Republic of China*
[5]*Department of Life Sciences and Genome Sciences, National Yang-Ming University, Taipei, Taiwan, Republic of China*
[6]*Graduate Institute of Biochemistry, National Yang-Ming University, Taipei, Taiwan, Republic of China*
[7]*Brain Research Center, University System of Taiwan, Hsinchu, Taiwan, Republic of China*
[8]*Department of Computer Science, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China*
[9]*Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan, Republic of China*

*KEY WORDS* subcellular features; automated recognition; CHO cells; Vero cells; GFP; rejection rate

*ABSTRACT* Systemic analysis of subcellular protein localization (location proteomics) provides clues for understanding gene functions and physiological condition of the cells. However, recognition of cell images of subcellular structures highly depends on experience and becomes the rate-limiting step when classifying subcellular protein localization. Several research groups have extracted specific numerical features for the recognition of subcellular protein localization, but these recognition systems are restricted to images of single particular cell line acquired by one specific imaging system and not applied to recognize a range of cell image sources. In this study, we establish a single system for automated subcellular structure recognition to identify cell images from various sources. Two different sources of cell images, 317 Vero (http://gfp-cdna.embl.de) and 875 CHO cell images of subcellular structures, were used to train and test the system. When the system was trained by a single source of images, the recognition rate is high and specific to the trained source. The system trained by the CHO cell images gave high average recognition accuracy for CHO cells of 96%, but this was reduced to 46% with Vero images. When we trained the system using a mixture of CHO and Vero cell images, an average accuracy of recognition reached 86.6% for both CHO and Vero cell images. The system can reject images with low confidence and identify the cell images correctly recognized to avoid manual reconfirmation. In summary, we have established a single system that can recognize subcellular protein localizations from two different sources for location-proteomic studies. *Microsc. Res. Tech. 71:305–314, 2008.* 

## INTRODUCTION

The localization of a protein in the living cells is directly related to the protein's function(s) (Heo and Meyer, 2003; Mochizuki et al., 2001). Miss-localization of proteins has been correlated with several diseases (Ameen and Salas, 2000; Neufeld, 1991). The Human Genome Project has identified a very large number of ESTs and genes, and therefore the analysis of the protein localization (location proteomics) of these genes will be a tremendous job. The development of an automatic large scale analysis system that is able to work with large image datasets has thus become important (Glory and Murphy, 2007). There are two ways of obtaining a protein's subcellular localization; one is prediction based on the protein sequence and the other is experimental. Several research groups have developed approaches that predict subcellular protein localization (Chou and Shen, 2007; Eisenhaber and Bork, 1998; Nakai and Horton, 1999). These efforts are able to correctly classify between 60% and 80% of proteins whose locations are already known; but their major limitation is associated with the nature of available training data. Moreover, these predictions still need to be confirmed experimentally.

*Correspondence to: Chung-Chih Lin, Faculty of Life Sciences and Institute of Genome Sciences, 320 Laboratory, Nursing Building, National Yang-Ming University, 155, Li-Nong Street, Section 2, Peitou, Taipei, Taiwan, Republic of China. E-mail: cclin2@ym.edu.tw

Received 27 June 2007; accepted in revised form 6 November 2007

Contract grant sponsor: National Science Council; Contract grant number: NSC94-2311-B-010-008; Contract grant sponsor: Brain Research Center, University System of Taiwan, Ministry of Education (Aim for Top University Plan); Contract grant sponsor: Chung Shan Medical University; Contract grant number: CSMU91-OM-B-014.

DOI 10.1002/jemt.20555

Published online 10 December 2007 in Wiley InterScience (www.interscience.wiley.com).

Advances in GFP technology have allowed protein fluorescence to become a useful tool in the visualization of subcellular localization in living cells. Construction of expression vectors, transfection, and cell imaging have been automated, which allows the large-scale analysis of the subcellular localizations of GFP-tagged fusion proteins to be easily accomplished. Several groups have used cell imaging to determine subcellular localization and have established image-based protein localization databases that classify proteins into groups experimentally (Bannasch et al., 2004; Glory and Murphy, 2007; Habeler et al., 2002; Simpson et al., 2000). However, classification of fluorescence cell micrographs is still subjective, time-consuming, and experience-intensive. Therefore, these datasets are highly variable and do not provide unambiguous information on localization that can be entered into other databases (Glory and Murphy, 2007; Murphy et al., 2000).

There are a number of different ways to turn these results into objective and numerical descriptions of protein subcellular localization that are suitable for databases. In tissue, cells have particular arrangements, and the recognition of subcellular protein localization using a tissue image can be done by simply comparing the test image with the model image pixel-by-pixel in order to measure similarity between these two images (for example, the KIND mediator; www.npaci.edu/DICE/Neuro). However, such strategy cannot be applied to cell lines due to their heterogeneity in cell morphology and subcellular structural arrangement. Therefore, some investigators have calculated numerical features of the images to help recognition. Examples of such are Zernik's moment and Harlick's texture, which have been used to extract features from micrographs of five subcellular structures, including the Golgi apparatus, nuclei, lysosomes, the nuclear envelope, and microtubules from CHO cells; these results were then applied to the automated recognition of CHO subcellular structures (Boland et al., 1998).

Several recognition systems have been developed for different cell types, including Hela cells (Boland and Murphy, 2001) and MCF7 cells (Conrad et al., 2004). All the recognition systems described earlier are restricted to images of a single particular cell line acquired by one specific imaging system. Some evidence, however, does show that such systems may be overtrained, and that this results in them being specific to single particular cell type. For example, CHO-specific subcellular classifiers cannot be applied to the recognition of Hela cell images because of their higher subcellular morphology heterogeneity compared with CHO cells (Boland and Murphy, 2001). Thus, these recognition systems cannot be applied by other investigators who use a different cell type and a different cell imaging systems. In this paper, we used cell images of two different cell lines acquired by different imaging systems to establish a single automated recognition system that is able to recognize a range of structures from cell image sources that were acquired by different imaging systems.

## MATERIALS AND METHODS

### Plasmids and Fluorescence Dye

pEYFP-Actin, pECFP-Peroxi, pEYFP-ER, pEYFP-Tub, pEYFP-Golgi, and pDsRed-Mito were used for labeling of the subcellular proteins in CHO cells and purchased from DB Biosciences (Clontech, BD Biosciences, NJ). Hoechst 33342 stain was from Sigma (St. Louis) and was used to visualize nuclei. The plasmids used to label subcellular proteins of Vero cells have been described earlier (Simpson et al., 2000). All chemicals for cell culture were obtained from Gibco-Invitrogen.

### Cell Culture

CHO cells were grown on tissue culture plates in Mc5A supplemented with 10% fetal calf serum plus penicillin and streptomycin in a 10% $CO_2$ atmosphere at 37°C and split 1–10 every 2 days. Vero cells were cultured in MEM supplemented with 10% fetal calf serum plus penicillin and streptomycin in a 10% $CO_2$ atmosphere at 37°C.

### Transfection

For Vero cells, the day prior to transfection, the cells were plated into 35-mm glass-bottomed dishes (Mat Tek Corp., MA) at a density of 20%. On the day of transfection, 1 μg of each DNA was used with 3 μL of FuGENE6 (Roche, Mannheim, Germany) to transfect the cells according to the manufacturer's instructions. For CHO cells, the day prior to transfection, cells were plated into dishes at a density of 50%. On the day of transfection, 1 μg of each DNA was used with 3 μL of Lipofectamine 2000 to transfect the cells according to the manufacturer's instructions. Cells were trypsinized to become a suspension and plated into dishes containing a poly-ι-lysine-coated coverslip at a density of 10%. When cells were attached to the coverslip, they were prepared for imaging (Simpson et al., 2000).

### Cell Imaging

Vero cells were imaged at 16, 24, and 40 h after transfection in carbonate-free culture medium equilibrated with 10 mM HEPES on a Leica DM/IBRE microscope with a 63× NA 1.4PL Apo objective, using custom-designed CFP or YFP filters. Images were captured with a Hammatsu CCD camera (ORCA 1) using Openlab 2.0 software (Improvision, Coventry, UK). Images were analyzed using Adobe Photoshop 5.0 (Simpson et al., 2000).

CHO cells were imaged at 48 h after transfection in carbonate-free culture medium equilibrated with 10 mM HEPES on a Zeiss Axiovert 25CF microscope with a 100× NA 1.3 oil EC Plan-Neofluar® objective using a DAPI, a custom-designed CFP or an YFP filter. Images were captured with an Axiocam CCD camera (color) using Axiovision software (Zeiss, Jena, Germany).

### Feature Generation and Classification

The raw cell images are passed through a series of processes before feature extraction (Fig. 1). The image size is first normalized to 500 × 500 by bilinear rescaling. The R-G-B color spaces of the image are separated, and the brightest color space is used for gray level conversion. A nonlinear sigmoid function is applied to adjust the contrast of the gray level image (Wen et al., 2001). This nonlinear transformation limits each pixel's value from 0 to 1. A searching algorithm is then

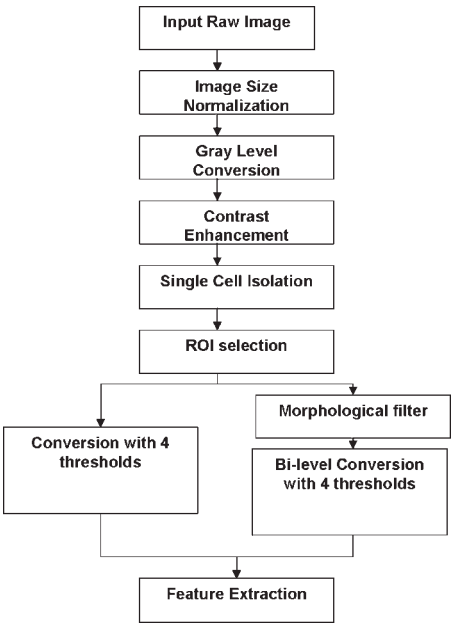


Fig. 1. Flowchart of feature extraction.

used to find and remove fragmented cells on edges of the image. Once there is only one cell preserved on the image, an automatic region-of-interest (ROI) selection procedure is exploited. Each row of pixel values is accumulated to delineate a vertical gray level profile. Likewise, each column of pixel values is accumulated to delineate a horizontal gray level profile. By setting an appropriate threshold for the profile, the boundary of a rectangle ROI can be defined (Fig. 2A). Top-hat and bottom-hat morphological filters are utilized to reduce the large and high gray level clusters and to enhance the edges of subcellular structures (Movafeghi et al., 2004; Fig. 2B). Figure 2B shows that the image after such morphological filtering and the processing yields better edge contrast and unambiguous subcellular architecture.

In order to represent progressive visual perception, the images before the morphological filters are converted to bilevel images with 0.1, 0.3, 0.5, and 0.7—four thresholds. Also, after the applying morphological filters, the images are converted to bilevel images with 0.2, 0.4, 0.6, and 0.8 thresholds. These eight bileveled images were subjected to a feature-extraction process (Fig. 2C).

Both geometric and texture features were used to identify the subcellular vesicles. The geometric features consisting of

– The number of objects
– The size of the largest object
– The number of objects which are bigger than 1/2 of the largest object size
– The number of objects which are bigger than 1/10 of the largest object size
– The perimeter of the vesicle
– The average object size
– The difference between the largest and average object size
– The largest and smallest circularities
– The difference between the largest and smallest circularities
– Compactness
– Euler number
– The ratio of hole size and vesicle size
– The minimum and maximum radiuses
– The ratio of the maximum and minimum radiuses
– Eccentricity

The texture feature extraction is based on the gray level co-occurrence matrix (GLCM) method proposed by Harlick (1979). Twelve GLCMs with distances of 1, 2, and 10 and angles of 0°, 45°, 90°, and 135° are applied to the bileveled images. Then, various co-occurrence quantities including entropy, energy, contrast, homogeneity, and correlation can be evaluated from the co-occurrence matrix to produce the feature set. These features are defined as

$$\text{Entropy} = -\sum_i \sum_j P(i,j)\ \log P(i,j)$$

$$\text{Energy} = \sum_i \sum_j P^2(i,j)$$

$$\text{Contrast} = \sum_i \sum_j (i-j)^2 P(i,j)$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{P(i,j)}{1+|i-j|}$$

$$\text{Correlation} = \sum_i \sum_j \frac{i \times j \times P(i,j) - u_x u_y}{\sigma_x^2 \sigma_y^2}$$

Training of an accurate classification system was greatly corrupted by atypical or wrong images such as dead or abnormal cells. Therefore, the images used for training were inspected and selected based on the following individual subcellular features. First, that the nuclei were oval in shape. Second, that the fluorescence cell images of nucleoli showed labeling mostly in the nucleolus with only minor labeling in nucleus, with there being only one or two nucleoli present in each nucleus. Third, that the peroxisomes were round in shape and were fewer and bigger in younger cells. Fourth, that most ER was connected to the nuclear envelope with the network spreading throughout the entire cytosol. Fifth, that the mitochondria were in threads that reached from the perinuclear regions to edges of cells. Sixth, that the Golgi apparatus was in the form of punctuated shapes and surrounded the nucleus. Seventh, that the actin filaments consisted of straight fibers in the cytosol. Finally, that the microtubules were curved fibers with the microtubule organization extending from the cell center to the cytosol. Based on these criteria, 317 Vero cell (selected from 4,439 cell images in gfp-cdna) and 815 CHO cell images were used to build an automated classification system that could identify eight different classes of subcellular localizations. Typical images of the subcellular classes are shown in Figure 3. These cells were applied to the feature extraction process. In total, 155 geometric features and 500 texture features were derived, and only

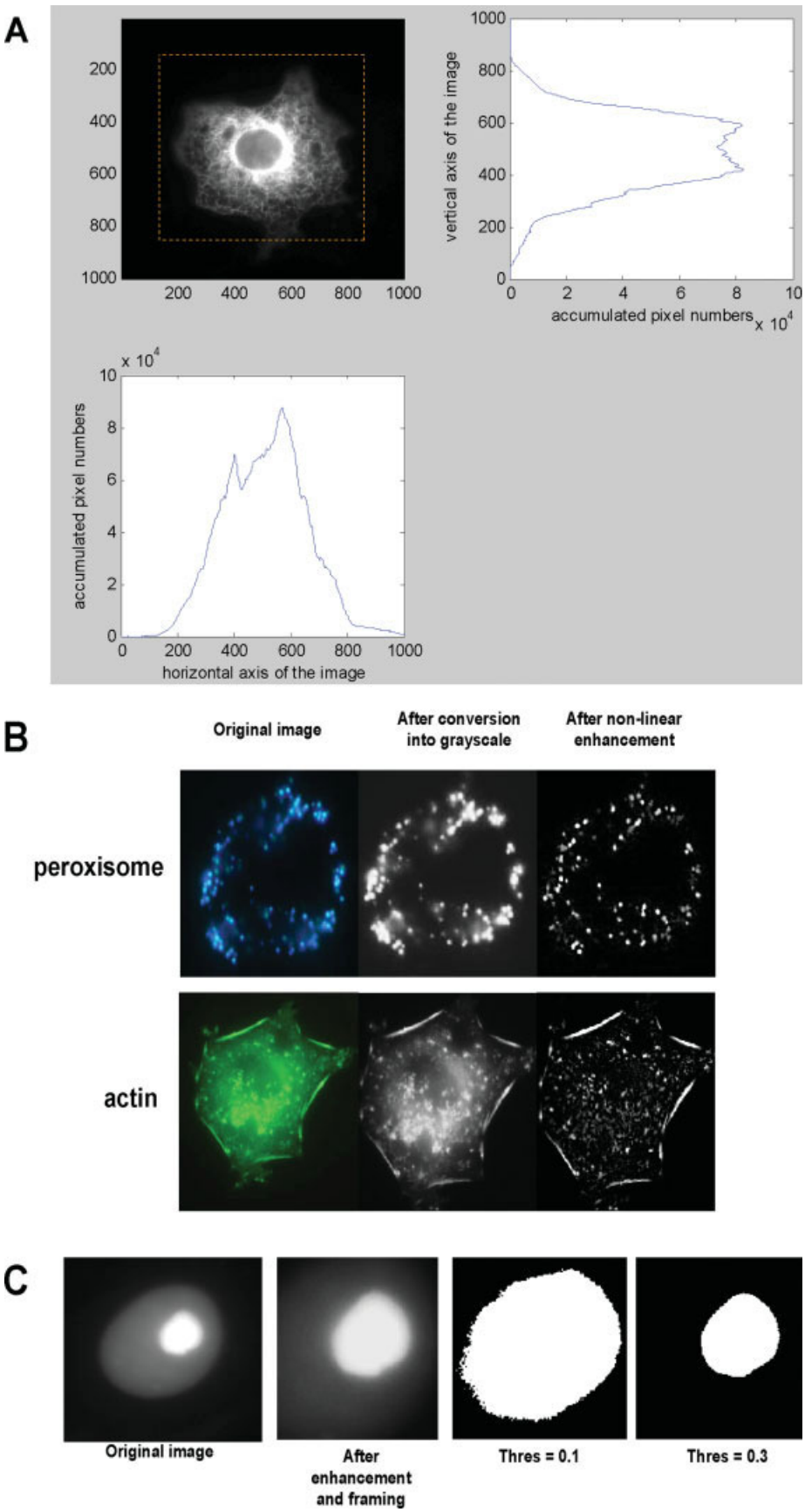


Fig. 2. Image processing for feature extraction. The cell images are framed to remove any useless area (**A**). The framed images were enhanced either linearly or by morphological filters, top-hat and bottom-hat (**B**). Enhanced images are bileveled at various thresholds (**C**). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

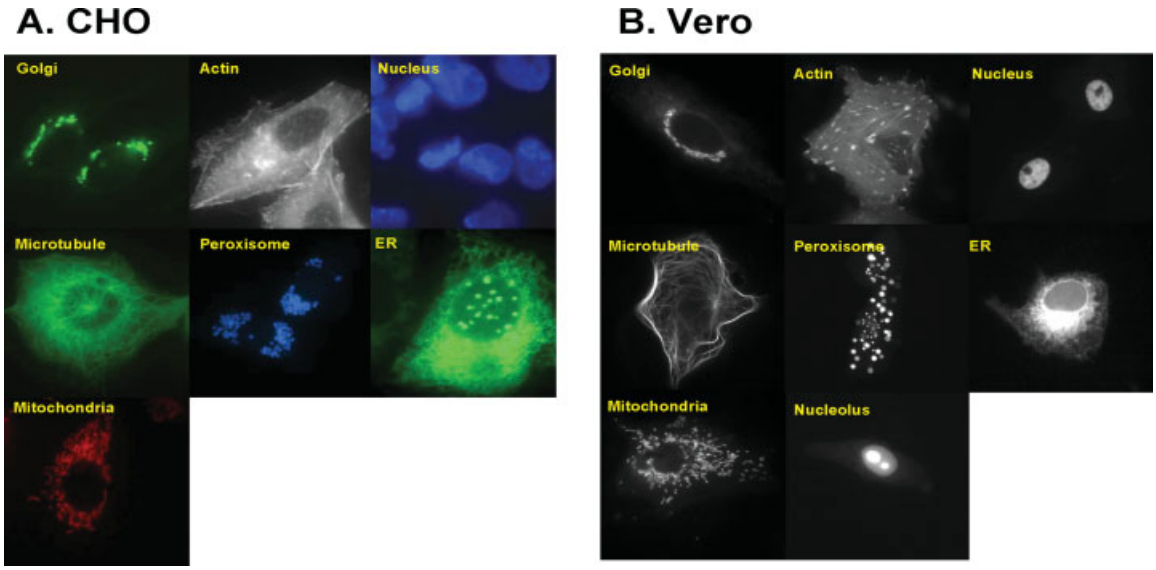


Fig. 3. Typical cell images of subcellular localizations used for training the system. The cells had expressed FP (fluorescence protein)-tagged marker proteins at subcellular structures for 2 days and the images were acquired by epi-fluorescence microscopy as described in the "Materials and Methods." Represented cell images of the subcellular structures are shown in **A** (CHO cell images) and **B** (Vero cell images from gfp-cdna). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

134 significant features were left for further protein subcellular localization after backward stepwise discriminant analysis.

## RESULTS

### Accuracy of the Recognition Systems Trained by Cell Images From a Single Source

The system trained by CHO cell images can recognize CHO subcellular structures with a high accuracy, 96% on average (Table 1A). However, using this system, the average recognition accuracy for Vero cell images is only 46% (Table 1B). It is possible that many subcellular features of the system are CHO-specific instead of subcellular structure-specific. Interestingly, the system is able to recognize Vero ER structure fairly well at 68%, possibly due to the fact that CHO ER structure tends to be more heterogeneous than that found in Vero cells.

The system trained by Vero cell images only is able to recognize Vero cell images quite well at 85% on average (Table 2A). The accuarcy is a little lower than for CHO cell images using the system trained by CHO cells. Possible reasons for this include the fact that the number of Vero cell images is smaller and that the complexity of the Vero subcellular structures is higher. In a similar manner, this system cannot correctly recognize CHO cell images and has an accuracy of only 50% (Table 2B). Some of CHO subcellular structures (actin, Golgi apparatus and nucleus) are recognized by this system fairly well (69–73%) and the morphology of these Vero subcellular structures are more heterogeneous than those of the CHO cells.

### Accuracy of the Recognition Systems Trained by Two Sources of Cell Images

Based on these findings, it would seem that the systems trained by a single source of cell images are specific to that cell type and that a mixture of images from various cell image sources may help the system to extract more general subcellular cell features and allow the recognition of a wider range of cell images from various sources. The 815 CHO and 317 Vero cell images were pooled and 9/10 of the images were used as the training group and the remaining images were used as the test group during system training. After cross validation, all the pooled images were used to test the performance of the system. The new system was able to recognize structures in Vero and CHO cells quite well at 84% and 89% on average, respectively,

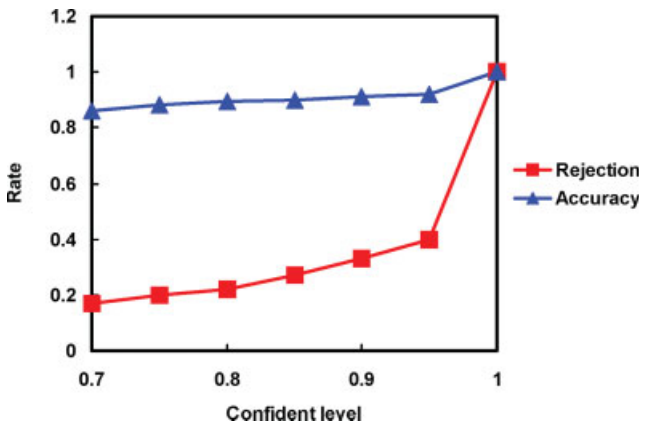


Fig. 4. Relationship between the accurate classification rate and the data rejection rate in GMM using a mixture of Vero and CHO cells. Features used in decision tree were further fed into the Gaussian mixture model (GMM) to measure the probabilities of an input pattern belonged to each class. Rejection rate (red) and accuracy of recognition (blue) at various confidence values is shown. This is plotted to set up a confidence level to filter out patterns that show a low probability of recognition into one of the eight classes. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

*TABLE 1. Recognition of Vero and CHO cell images by the system trained by CHO cell images*

| | True classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Output of decision tree | Actin | ER | Golgi | Peroxisome | Mitochondria | Microtubule | Nucleus | Nucleolus |
| A. Trained by 815 CHO cell images/tested CHO cell images | | | | | | | | |
| Actin | 303 | 6 | 0 | 1 | 0 | 3 | 0 | ND |
| ER | 6 | 157 | 0 | 0 | 3 | 3 | 0 | ND |
| Golgi | 2 | 0 | 121 | 2 | 1 | 0 | 0 | ND |
| Peroxisome | 1 | 1 | 1 | 63 | 1 | 0 | 0 | ND |
| Mitochondria | 2 | 0 | 1 | 0 | 49 | 0 | 0 | ND |
| Microtubule | 1 | 1 | 0 | 0 | 0 | 56 | 0 | ND |
| Nucleus | 0 | 0 | 0 | 0 | 0 | 0 | 30 | ND |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ND |
| Error | 12 | 8 | 2 | 3 | 5 | 6 | 0 | ND |
| Total images | 315 | 165 | 123 | 66 | 54 | 62 | 30 | ND |
| Rate (%) | 96.2 | 95.2 | 98.4 | 95.5 | 90.7 | 90.3 | 100.0 | ND |
| B. Trained by 815 CHO cell images/tested Vero cell images | | | | | | | | |
| Actin | 6 | 4 | 4 | 3 | 1 | 21 | 1 | ND |
| ER | 6 | 27 | 7 | 1 | 17 | 15 | 0 | ND |
| Golgi | 0 | 0 | 27 | 7 | 0 | 0 | 18 | ND |
| Peroxisome | 0 | 3 | 13 | 3 | 12 | 0 | 8 | ND |
| Mitochondria | 0 | 0 | 2 | 0 | 37 | 2 | 0 | ND |
| Microtubule | 5 | 6 | 4 | 0 | 1 | 13 | 7 | ND |
| Nucleus | 0 | 0 | 0 | 0 | 0 | 0 | 18 | ND |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 3 | ND |
| Error | 11 | 13 | 30 | 11 | 31 | 38 | 37 | ND |
| Total images | 17 | 40 | 57 | 14 | 68 | 51 | 55 | ND |
| Rate (%) | 35.3 | 67.5 | 47.4 | 21.4 | 54.4 | 25.5 | 32.7 | ND |

Overall, to train the system, 9/10 of the 815 CHO images were used as the training group and the remaining 10% images were used as the test group. After crossvalidation, all CHO and Vero cell images were used for testing the performance of the system. The accuracy of recognition for individual subcellular structures of CHO and Vero cells is shown in **A** and **B**. The column of a particular entry indicates the true classification of those images, while the row represents the class to which those images were assigned by the decision tree. Average recognition rate is calculated as the percentage of correctly recognized images.

*TABLE 2. Recognition of Vero and CHO cell images by the system trained by Vero cell images*

| | True classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Output of decision tree | Actin | ER | Golgi | Peroxisome | Mitochondria | Microtubule | Nucleus | Nucleolus |
| A. Trained by Vero cell images/tested Vero cell images | | | | | | | | |
| Actin | 15 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| ER | 1 | 33 | 0 | 0 | 1 | 3 | 0 | 0 |
| Golgi | 0 | 1 | 45 | 2 | 0 | 1 | 1 | 0 |
| Peroxisome | 0 | 0 | 0 | 11 | 0 | 0 | 2 | 0 |
| Mitochondria | 0 | 1 | 1 | 0 | 62 | 1 | 0 | 0 |
| Microtubule | 0 | 0 | 0 | 0 | 1 | 41 | 0 | 0 |
| Nucleus | 0 | 1 | 3 | 1 | 0 | 0 | 49 | 1 |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 14 |
| Error | 2 | 7 | 12 | 3 | 6 | 10 | 6 | 1 |
| Total images | 17 | 40 | 57 | 14 | 68 | 51 | 55 | 15 |
| Rate (%) | 88.2 | 82.5 | 78.9 | 78.6 | 91.2 | 80.4 | 89.1 | 93.3 |
| B. Trained by 317 Vero cell images/tested CHO cell images | | | | | | | | |
| Actin | 218 | 32 | 1 | 4 | 0 | 12 | 0 | ND |
| ER | 21 | 38 | 0 | 0 | 0 | 19 | 0 | ND |
| Golgi | 1 | 2 | 88 | 35 | 17 | 4 | 0 | ND |
| Peroxisome | 8 | 13 | 5 | 3 | 0 | 0 | 3 | ND |
| Mitochondria | 20 | 23 | 0 | 17 | 30 | 11 | 0 | ND |
| Microtubule | 29 | 12 | 0 | 0 | 4 | 10 | 0 | ND |
| Nucleus | 1 | 0 | 15 | 3 | 0 | 0 | 22 | ND |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ND |
| Error | 97 | 127 | 35 | 63 | 24 | 52 | 8 | ND |
| Total images | 315 | 165 | 123 | 66 | 54 | 62 | 30 | ND |
| Rate (%) | 69.2 | 23.0 | 71.5 | 4.5 | 55.6 | 16.1 | 73.3 | ND |

Overall, to train the system, 9/10 of 317 Vero images were used as the training group and the remaining images were used as the test group. After crossvalidation, all CHO and Vero cell images were used for testing the performance of the system. The accuracy of recognition for individual subcellular structures of Vero and CHO cells is shown in **A** and **B**.

only slightly lower than when the system is trained by single sources of cell images and the system is tested on that specific cell type (Table 3). The system can recognize mitochondria with very high accuracy in both cell types. When recognizing peroxisomes and mitochondria in Vero cells, this system has a higher accuracy rate than the system using Vero cells alone. In contrast, the accuracy rate of this system when recognizing CHO peroxisomes is much lower at 39% compared to the CHO cell system. When we used 1,488 Vero cell images from gfp-cdna, which were classified in eight subcellular classes, as a test, the accuracy was reduced to 61% on average (Table 4). Images of actin and the nucleolus have a lower accuracy, possibly due to too few Vero cell images (namely 17 and 15) being used for training the system. The subcellular struc-

*TABLE 3. Recognition of Vero and CHO cell images by the system trained by Vero and CHO cell images*

| | True classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Output of decision tree | Actin | ER | Golgi | Peroxisome | Mitochondria | Microtubule | Nucleus | Nucleolus |
| A. Trained by 815 CHO cell images and 317 Vero cell images/tested CHO cell images | | | | | | | | |
| Actin | 304 | 9 | 2 | 0 | 0 | 1 | 2 | ND |
| ER | 6 | 144 | 3 | 2 | 6 | 3 | 1 | ND |
| Golgi | 1 | 3 | 115 | 14 | 0 | 0 | 0 | ND |
| Peroxisome | 0 | 0 | 3 | 37 | 0 | 0 | 0 | ND |
| Mitochondria | 0 | 2 | 0 | 11 | 41 | 2 | 0 | ND |
| Microtubule | 3 | 7 | 0 | 2 | 7 | 56 | 0 | ND |
| Nucleus | 1 | 0 | 0 | 0 | 0 | 0 | 27 | ND |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ND |
| Error | 11 | 21 | 8 | 29 | 13 | 6 | 3 | ND |
| Total images | 315 | 165 | 123 | 66 | 54 | 62 | 30 | ND |
| Rate (%) | 96.5 | 87.3 | 93.5 | 56.1 | 75.9 | 90.3 | 90.0 | ND |
| B. Trained by 815 CHO cell images and 317 Vero cell images/tested Vero cell images | | | | | | | | |
| Actin | 12 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| ER | 2 | 31 | 6 | 1 | 3 | 3 | 4 | 0 |
| Golgi | 1 | 1 | 48 | 0 | 0 | 0 | 1 | 2 |
| Peroxisome | 0 | 0 | 0 | 13 | 0 | 0 | 5 | 1 |
| Mitochondria | 1 | 7 | 3 | 0 | 64 | 3 | 0 | 0 |
| Microtubule | 1 | 1 | 0 | 0 | 1 | 43 | 0 | 0 |
| Nucleus | 0 | 0 | 0 | 0 | 0 | 0 | 43 | 0 |
| Nucleolus | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 12 |
| Error | 5 | 9 | 9 | 1 | 4 | 8 | 12 | 4 |
| Total images | 17 | 40 | 57 | 14 | 68 | 51 | 55 | 15 |
| Rate (%) | 70.6 | 77.5 | 82.2 | 92.9 | 94.1 | 84.3 | 78.2 | 80.0 |

Overall, to train the system, a 9/10 mixture of 815 CHO and 317 Vero cell images were used as the training group and the remaining images were used as the test group. After crossvalidation, all CHO and Vero cell images were used for testing the performance of the system. The accuracy of recognition for individual subcellular structures of CHO and Vero cells are shown in **A** and **B**.

*TABLE 4. Recognition of 1,488 Vero cell images by the system trained by a mixture of Vero and CHO cell images*

| | True classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Output of decision tree | Actin | ER | Golgi | Microtubule | Mitochondria | Nucleolus | Nucleus | Peroxisome |
| Actin | 26 | 23 | 5 | 11 | 5 | 0 | 3 | 0 |
| ER | 25 | 245 | 23 | 7 | 26 | 0 | 8 | 7 |
| Golgi | 4 | 19 | 107 | 1 | 5 | 40 | 82 | 0 |
| Microtubule | 11 | 47 | 2 | 76 | 7 | 0 | 5 | 0 |
| Mitochondria | 3 | 34 | 9 | 3 | 131 | 0 | 7 | 4 |
| Nucleolus | 0 | 0 | 0 | 1 | 0 | 52 | 42 | 0 |
| Nucleus | 0 | 7 | 4 | 0 | 1 | 10 | 272 | 1 |
| Peroxisome | 3 | 5 | 8 | 0 | 4 | 10 | 27 | 30 |
| Error | 46 | 135 | 51 | 23 | 48 | 60 | 174 | 12 |
| Correct | 26 | 245 | 107 | 76 | 131 | 52 | 272 | 30 |
| Total | 72 | 380 | 158 | 99 | 179 | 112 | 446 | 42 |

In total, 1488 images of Vero cells made up of eight subcellular categories using gfp-cdna were tested by the system as described in Table 3. The accuracy of recognition for each category is summarized in this table.

tures, ER, Golgi apparatus, and the nucleus, that have a larger Vero cell image numbers for training still only gave an accuracy of 61–68%. In such classes, the test images are much greater in number by (3- to 10-fold and the subcellular features may not be sufficiently extracted from the training images.

### Rejection of Miss-Recognized Cell Images

It is necessary to establish a method to reject the images that are missed recognized so that no manual re-examination is necessary. Here we adopted the same 134 features used in decision tree and fed these into the Gaussian mixture model (GMM), which provided us with the probabilities of an input pattern belonged to each class. Hence, we were able to properly set up a confidence level to filter patterns with lower probabilities of recognizing a specific class. As shown in Figure 4, GMM was then used in the recognition of a mixture of Vero and CHO cell images to decide the proper confidence criterion by observing the relationships between the accurate classification rate and the data rejection rate. If a confidence level of 0.85 is chosen, then 25% of the input patterns are viewed as noisy data and removed; the average accuracy of our system now increases to 90% (Fig. 4). If a confidence level of 0.7 is chosen, 15% of the data is noisy and removed, giving an average accuracy of 87%. For our experiments, we chose a confidence level of 0.85, and when used with Vero and CHO cells, this gave an improved accuracy of 89% and 91% with 33% and 23% removal of noisy data, respectively. The recognition results for individual subcellular structure in Vero and CHO cells are shown in Table 5. In addition, we have also examined the rejected images. As expected, the characteristics of some images are not clear enough, so they may be easily classified into the wrong class. We need to fully review the image processing, including image segmentation and image enhancement of these miss-classified images, to find whether new solutions help

*TABLE 5. Recognition results of Vero and CHO cell images after rejection of images with low confidence of recognition*

| | Results | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Actin | ER | Golgi | Microtubule | Mitochondria | Nucleolus | Nucleus | Peroxisome |
| CHO cells | | | | | | | | |
| Accuracy rate (%) | 96.8 | 88.0 | 88.2 | 81.8 | 100 | No | 100 | 81.8 |
| Rejection rate (%) | 20.0 | 26.4 | 12.8 | 31.2 | 45.4 | No | 25.0 | 15.3 |
| Vero cells | | | | | | | | |
| Accuracy rate (%) | 75.0 | 94.1 | 71.4 | 88.8 | 87.8 | 85.7 | 93.9 | 80.0 |
| Rejection rate (%) | 46.6 | 26.8 | 63.1 | 62.5 | 23.2 | 22.2 | 26.1 | 28.5 |

The threshold of confidence value for rejection was set at 0.85 based on Figure 4. In total, 317 images of CHO cells and 1,488 images of Vero cells were applied to test for improvement in the recognition rate after rejection of images as described in Figure 4.

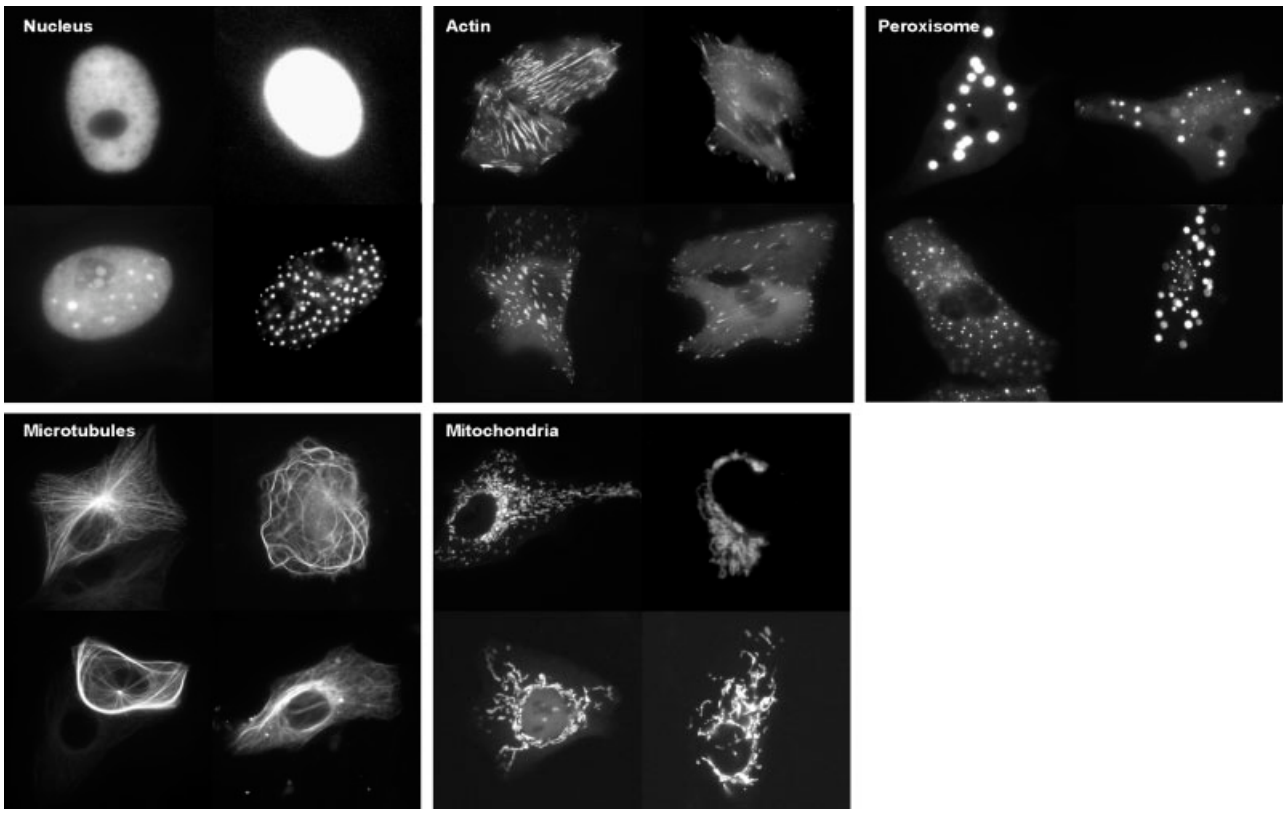


Fig. 5. Heterogeneity of subcellular structures of Vero cells. Heterogeneity of nucleus, actin, peroxisome, microtubule, and mitochondria is high in Vero cells, and four represented images of each subcellular structure are shown in this figure.

recognize these images correctly or just think these images as artifacts.

## DISCUSSION

Clearly, in this study, we have shown that it is possible for cell images of subcellular structures from various sources to be recognized by a single system at an accuracy of about 86%. Our system is able to reject miss-recognized images and identify correctly recognized images with an average accuracy of more than 90% correctly recognized; this will save time during re-examination.

The system has lower accuracy when recognizing 1,488 Vero cell images, even when they are from the same source as the Vero cell images used for training. There are a number of possible reasons for this low accuracy of recognition. First, subcellular classes with low recognition accuracy, such as actin, the nucleolus, the ER, the Golgi apparatus, and the nucleus, may have too few training images and this reduces the extraction of enough subcellular features for recognition. Second, images of some subcellular classes have variable morphological properties. For example, centromeric proteins are dotted in nucleus but proteins on chromatin fill the nucleus, but both localizations are categorized in nucleus group. A similar situation happens with Vero actin, microtubules, mitochondria, and peroxisomes, and this may result in low recognition specificity (Fig. 5). Third, some proteins have dynamic localizations, such as membrane proteins localizing to the ER, to vesicles, to Golgi apparatus, and to the plasma membrane during protein export. Some proteins may have several functions and localize in two or more subcellular compartments (Fig. 6A). In such

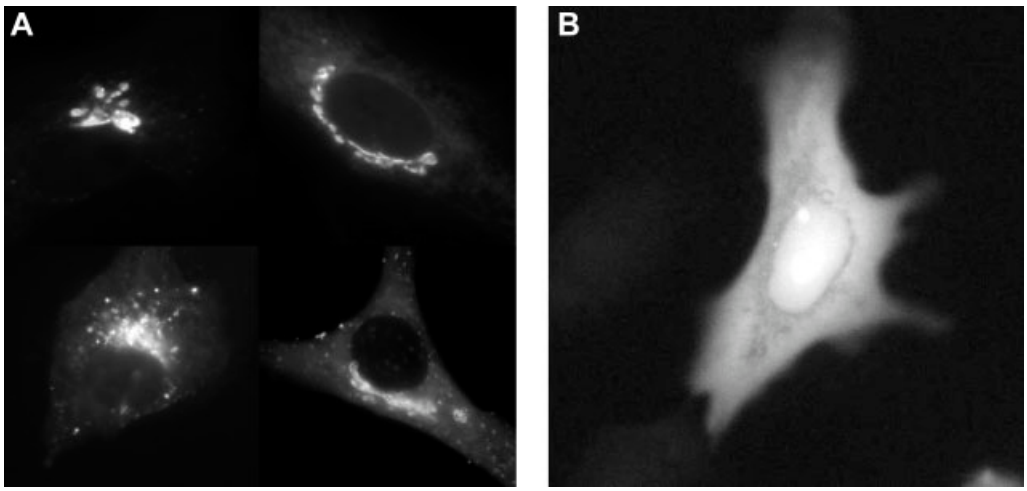


Fig. 6. Two major cases of miss-recognition. **A:** The lower two images have not only labeled Golgi apparatus, but the labeling also includes vesicles and the plasma membrane. **B:** This image has homogeneous fluorescence in nucleus and heterogeneous fluorescence in cytosol and this appears as similar to ER when the image is enhanced by the morphological filter.

cases, the images consist of two or more subcellular features, and this results in recognition being difficult. Fourth, some images contain two or more cells that cannot be segmented, and this will change their geometric properties reducing recognition accuracy. Fifth, every Vero image represents subcellular protein localization of one gene, but the properties for one particular subcellular protein localization within a cell population may be heterogeneous and the system may not have sufficient specific features for recognition to occur accurately. Sixth, some images will have similar subcellular features after processing, and this will result in miss-recognition. One example is where some nuclei show homogeneous strong fluorescence, but also has a low heterogeneous fluorescence background, which is easily recognized as ER. When we examined the path of decision tree of nuclei with these features after enhancement by the morphological filter, it was clear that the morphological filter recognizes the nuclear area with homogenous fluorescence as background and enhances the heterogeneous cytosolic fluorescence as a network of structures. In such circumstances, the enhanced nucleus images are very dark in nuclear region, and there is a strong network structure in cytosol that is very similar to ER (Fig. 6B).

We need to examine the earlier possibilities, and to do this we will increase the number of images including various proteins with similar localizations. We will then use other methods of feature selection to train the system to discover whether more images and other ways of selecting features may be able to extract more specific features for recognition. To increase the variety of images in every subcellular class, these images will be manually grouped into new subclasses to improve the system. To study cell images with several subcellular structures, the cells will be stained by several fluorescence probes or there will be overexpression of several subcellular structure specific fluorescence proteins; the result will be labeling of several subcellular compartments. Furthermore, subcellular morphologies will be merged in a single image to build cell images with several structures to know whether these cell images are difficult to be recognized. Since an automated cell image acquisition system cannot capture fields that contain only one cell, images containing several cells frequently occur. The efficiency of the segmentation method is very dependent on the images and some images are very difficult to be resolved into single cell. To help this, we will increase the subcellular features that are independent of size and number such as texture features, and this ought to improve the accuracy of recognition.

In this study, we have shown that recognition of the subcellular structures from various sources of cells is possible in this case by the use of two different sources of cells. This approach should be applicable to other investigators when they use images of subcellular structures to obtain more information about the subcellular localizations of novel proteins. Besides, the approach described here will be particularly useful if applied to high throughput screening of chemical compounds and drugs that affect subcellular localization of particular proteins involved in important cell processes. The use of two cell lines increases the reliability of such an assay and this system, particularly because of the rejection of unreliable images, will limit the need for human input into such a large-scale screen.

## ACKNOWLEDGMENTS

The authors thank all undergraduate students of classes 1999–2002 in Department of Life Sciences (current name, BioMedical Sciences), Chung-Shan Medical University, for providing some of the CHO cell images. We thank Dr. Rainer Pepperkok, Prof. Ralph Kirby, and Prof. Ueng-Cheng Yang for critical reviewing of the manuscript and valuable suggestions.

## REFERENCES

Ameen NA, Salas PJ. 2000. Microvillus inclusion disease: A genetic defect affecting apical membrane protein traffic in intestinal epithelium. Traffic 1:76–78.

Bannasch D, Mehrle A, Glatting K-H, Pepperkok R, Poustka A, Wiemann S. 2004. LIFEdb: A database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system. Nucleic Acid Res 32:D505–D508.

Boland MV, Murphy RF. 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.

Boland MV, Markey MK, Murphy RF. 1998. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 33:366–375.

Chou K-C, Shen H-B. 2007. Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678.

Conrad C, Erfle H, Warnat P, Daigle N, Lorch T, Ellenberg J, Pepperkok R, Eils R. 2004. Automatic identification of subcellular phenotypes on human cell arrays. Genome Res 14:1130–1136.

Eisenhaber F, Bork P. 1998. Subcellular localization of proteins based on sequence. Trends Cell Biol 8:169–170.

Glory E, Murphy R. 2007. Automated subcellular localization determination and high-throughput microscopy. Dev Cell 12:7–16.

Habeler G, Natter K, Thallinger GG, Crawford ME, Kohlwein SD, Trajanoski Z. 2002. YPL.db: The yeast protein localization database. Nucleic Acid Res 30:80–83.

Haralick RM. 1979. Statistical and structural approaches to texture. Proc IEEE 67:768–804.

Heo WD, Meyer T. 2003. Switch-of-function mutants based on morphology classification of Ras superfamily small GTPases. Cell 113:315–328.

Mochizuki N, Yamashita S, Kurokawa K, Ohba Y, Nagai T, Miyawaki A, Matsuda M. 2001. Spatio-temporal images of growth-factor-induced activation of Ras and Rap1. Nature 411:1065–1068.

Movafeghi A, Kargarnovin MH, Soltanian-Zadeh H, Taheri M, Ghasemi F, Rokrok B, Edaalati K, Kermani A, Rastkhah N. 2004. Quality improvement of digitized radiographs by filtering technique development based on morphological transformations, In: 2004 IEEE Nucleic Science Symposium and Medical Imaging Conference, Rome, Italy. pp. 579–582.

Murphy RF, Boland MV, Velliste M. 2000. Towards a systemics for protein subcellular localization patterns and automated analysis of fluorescence microscope images. Proc Int Conf Intell Syst Mol Biol (ISMB 2000) 8:251–259.

Nakai K, Horton P. 1999. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–35.

Neufeld EF. 1991. Lysosomal storage diseases. Annu Rev Biochem 60:257–280.

Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S. 2000. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 1:287–92.

Wen C-h, Lee J-j, Liao Y-c. 2001. Adaptive quartile sigmoid function operator for color image contrast enhancement. In: The Ninth Color Imaging Conference on Color Science and Engineering: Systems, Technologies, and Applications, Scottsdale, Arizona. pp. 280–285.