47

Automated Interpretation of Subcellular Location Patterns from Three-Dimensional Confocal Microscopy

Ting Zhao and Robert F. Murphy

INTRODUCTION

Confocal microscopy has been widely used for biological studies because of its unique ability to acquire images free of most out-of-focus light. This property is critical for the acquisition of high-quality three-dimensional (3D) images of cells. Automated image analysis and data mining methods can be used to extract rich information contained in such images. In this chapter, we focus on describing effective methods for automatically interpreting 3D images of protein subcellular location patterns.

Protein Subcellular Location

Proteomics, the large-scale study of proteins, is the next big step in decoding genomes and achieving a thorough understanding of how living systems work. An important aspect of protein behavior is subcellular localization, and the field of *location proteomics* is concerned with systematically analyzing the complex processes by which proteins are localized within cells (Chen *et al*., 2003). Context provided by the location of a protein can help biologists infer possible functions. For example, a protein located exclusively in the plasma membrane may serve as a transporter or ion channel, while a cytoplasmic protein may act as a catalyst. Recent projects have described systematic approaches to the determination of protein location (Kumar *et al*., 2002; Dreger, 2003). However, objectively and quantitatively describing subcellular location patterns in cells using computational methods has not been emphasized.

Most computational studies on protein subcellular location are focused on prediction of locations from the protein sequence (Chou and Elrod, 1999; Chou and Cai, 2002; Park and Kanehisa, 2003; Bhasin and Raghava, 2004). These prediction methods became popular because of the rapid increase in the number of available protein sequences. However, a significant problem with using any prediction system is that none of the methods can give perfect accuracy. This creates the risk that the unverified predicted results may be misleading when they are relied upon by biologists. Furthermore, the low resolution of current prediction schemes prevents them from distinguishing proteins in the same organelle. They can only classify proteins at the predefined organelle level, that is, only 12 categories were used in a recent study to describe the organelles and compartments in an animal or plant cell (Chou and Elrod, 1999). Not only can two proteins that reside in different subcompartments of the same organelle not be distinguished, but minor (or previously unrecognized) patterns cannot be recognized and proteins that show mixed patterns (e.g., residing in more than one organelle or moving from one to another) may not be recognized correctly.

Combining the rapid developments of microscope imaging techniques in biology with advances in image analysis and machine learning, we have over the past few years developed automated systems to analyze the subcellular distributions of proteins quantitatively and precisely at high resolution (Boland *et al*., 1997, 1998; Boland and Murphy, 2001; Huang *et al*., 2003; Murphy *et al*., 2003; Huang and Murphy, 2004b). Fluorescence microscopy is the most common method to experimentally determine subcellular location because it has much higher resolution than subcellular fractionation, and because, unlike electron microscopy, it permits rapid imaging of living cells.

Overview of 2D Dataset Analysis

From 2DCHO to 2DHeLa

Our work on automatic analysis of subcellular location patterns started with two-dimensional (2D) images, initially of five patterns in Chinese hamster ovary cells (the *2DCHO* dataset) and subsequently of 10 patterns in HeLa cells (the *2DHeLa* dataset). The quantitative description and classification of the *2DCHO* dataset opened up the area of systematic analysis of protein subcellular location (Boland *et al*., 1997, 1998). By using four primary antibodies, directed against the Golgi protein giantin, the lysosome protein LAMP2, the yeast nucleolar protein NOP4, tubulin, and a DNA stain, five classes of subcellular location patterns were created. After preprocessing steps including deconvolution, segmentation, and background subtraction, 49 Zernike moments and 13 Haralick texture features were calculated as features for each image. The Zernike moments represent the decomposition of an image into 2D polynomials and the Haralick features capture statistical properties of intensity distribution in neighboring pixels. Using these two sets of features, both global information and local information of an image will be captured. Ten of the 62 features were selected by stepwise discriminant analysis (SDA) and input into a single hidden-layer back-propagation neural network (BPNN). An average classification accuracy of 88% was obtained, demonstrating the initial feasibility of the application of image analysis and pattern recognition methods to subcellular location patterns.

Encouraged by the good results, a more thorough study was carried out on the *2DHeLa* dataset (Boland and Murphy, 2001). This covered most major organelles in animal cells (Fig. 47.1). In this dataset, HeLa cells were labeled for one of nine proteins and in parallel labeled with the DNA intercalating dye 4′,6-diamidino-2

Ting Zhao and Robert F. Murphy • Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

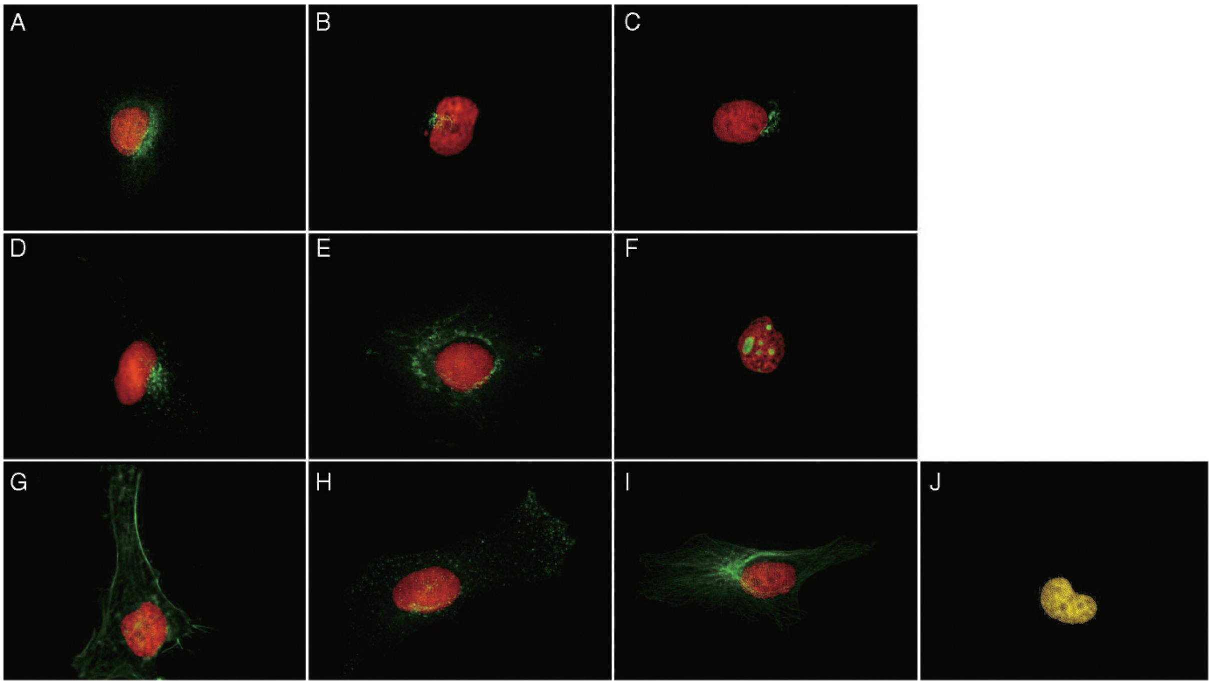

FIGURE 47.1. Typical images from the *2DHeLa* dataset (Boland and Murphy, 2001). The 10 classes are an ER protein (A), the Golgi protein giantin (B), the Golgi protein GPP130 (C), the lysosomal protein LAMP2 (D), a mitochondrial protein (E), the nucleolar protein nucleolin (F), the endosomal protein transferrin receptor (H), the cytoskeletal proteins filamentous actin (G) and tubulin (I), and DNA (J).

phenylindole dihydrochloride (DAPI). Seven of the proteins are located in organelles (endoplasmic reticulum, Golgi complex, lysosome, mitochondria, nucleolus, and endosome), with two of them in the same organelle, the Golgi complex. These two patterns were included to test whether automated methods could distinguish them. The other two proteins are found in cytoskeletal structures (filamentous actin and tubulin). A tenth class was created by considering the parallel DNA images as a class.

2D Subcellular Location Features

The Zernike moment features and Haralick texture features were found to be insufficient to describe all of the patterns in the *2DHeLa* dataset, and other types of measures were therefore implemented. We also implemented a standard nomenclature for referring to features and sets of features so that each result could be associated with the features used to create it (Boland and Murphy, 2001). Each set was referred to by the prefix SLF (for subcellular location features) followed by a number (e.g., SLF1) and each feature within a set was referred to by the set number and an index within it (e.g., SLF1.7).

The subcellular location features (SLF) can be divided into different subsets according to their properties (Table 47.1). SLF1 consists of eight morphological features, five edge-related features, and three convex hull features. These correspond roughly to attributes of patterns typically recognized by cell biologists. SLF2 is formed by adding six DNA-related features to SLF1. The DNA-related features provide the ability to distinguish similar patterns that differ in their proximity to the nucleus. For example, nucleolin can be easily distinguished from other proteins using these DNA-related features. SLF3 is the union of SLF1 with the 49 Zernike moment features and 13 Haralick texture features. SLF4 is the union of SLF3 and the six DNA related features.

TABLE 47.1. Brief Description of 2D Subcellular Location Features

Set	Size	SLF Index	Feature Description
SLF1	16	SLF1.1–1.8	Morphological features
		SLF1.9–1.13	Edge-related features
		SLF1.14–1.16	Convex hull features
SLF2	22	SLF2.1–2.16	SLF1.1–1.16
		SLF2.17–2.22	DNA-related features
SLF3	78	SLF3.1–3.16	SLF1.1–1.16
		SLF3.17–3.65	Zernike moment features
		SLF3.66–3.78	Haralick texture features
SLF4	84	SLF4.17–4.22	SLF2.17–2.22
		SLF4.1–4.16, 4.23–4.84	SLF3.23–3.78
SLF5	37	SLF5.1–5.37	Selected from SLF4 using SDA
SLF6	65	SLF6.1–6.65	SLF3.1–3.65
SLF7	84	SLF7.1–7.9	SLF3.1–3.9
		SLF7.10–7.13	Minor corrections to SLF3.10–3.13
		SLF7.14–7.65	SLF3.14–3.65
		SLF7.66–7.78	Modified Haralick texture features
		SLF7.79	The fraction of cellular fluorescence not included in objects
		SLF7.80–7.84	Skeleton feature
SLF8	32	SLF8.1–8.32	Selected from SLF7 using SDA
SLF12	8	SLF12.1–2.8	SLF8.1–8.8
SLF13	31	SLF13.1–13.31	Selected from SLF7 and SLF2.17–2.22 using SDA
SLF15	44	SLF15.1–15.44	Selected from SLF7 and 90 Gabor and wavelet features
SLF16	47	SLF16.1–16.47	Selected from SLF7, 90 Gabor and wavelet features and SLF2.17–2.22

It is often the case that features anticipated to be valuable for a given classification problem turn out to have little or no useful information. In this case, their presence can actually hinder the performance of a classifier. One approach to this problem is to select an informative subset of all available features, and we have found that the stepwise discriminant analysis (SDA) method performs very well for our purposes (Huang *et al.*, 2003). Thus, some SLF sets have been created by selecting informative features from one of the larger sets (e.g., SLF5 was selected from SLF4).

The 13 Haralick texture features are sensitive to image pixel resolution and number of gray levels. To permit the features to be compared for images at different resolutions, a set of improved texture features were defined by first downsampling each image to 1.15 µm/pixel and 256 gray levels. This resulted in a new feature set, SLF7, that consists of SLF3 (with the 13 improved texture features) and six new features that capture the fraction of cellular fluorescence not included in objects and object morphology as reflected in object skeletons.

To provide further information, a set of 90 features calculated by transforming an image into the frequency domain was added. These features were calculated based on the Gabor transformation and the wavelet transformation. Subsets of these features (either with or without the DNA features) were also selected (SLF15 and SLF16).

Classification Results for *2DHeLa* Dataset

The effectiveness of the SLF for describing subcellular patterns can be evaluated by training and testing classifiers. With good features, different patterns should be distinguished reasonably by a well-trained classifier. A feature set that gives better classification performance is more suitable for describing location patterns.

Since the introduction of the first SLF sets, we have explored additional feature sets and various classification approaches for the *2DHeLa* images. The most common classifier we have used was a BPNN with one hidden layer and 20 hidden nodes (the same architecture used for the *2DCHO* data). Table 47.2 shows that classification accuracy increased when more features were added (SLF2, SLF3, SLF4). However, fewer features can give higher accuracy if they are carefully selected (Huang *et al.*, 2003; Murphy *et al.*, 2003). SLF5 selected from SLF4 by SDA has fewer features than SLF4, but the classification accuracy is higher. Similarly, SLF8 and SLF13 gave better performance than SLF7.

Another way to increase the classification accuracies is to improve the classifiers. We have therefore compared the performance of eight different classification methods on the *2DHeLa* dataset (Huang and Murphy, 2004b). The results showed that ensemble methods increased the classification accuracy significantly. The best classification accuracy on SLF8 was achieved by Adaboost (88%), a method that keeps modifying the training set based on how well they can be classified by a base classifier. For SLF13, Mixture-of-Experts was the most suitable one among the eight classifiers (90%). This classifier can improve performance by separating the training set into many partitions. The eight classifiers can further be combined to organize optimal majority-voting ensembles. The highest accuracy on *2DHeLa* to date was achieved by such an ensemble on SLF16 (92%).

TABLE 47.2. Average Classification Accuracy of Different SLF Sets and Classifiers on the *2DHeLa* Dataset

Set	DNA	SDA from	Classifier	Accuracy
SLF2	Yes	—	BPNN	76%
SLF3	No	—	BPNN	79%
SLF4	Yes	—	BPNN	81%
SLF5	Yes	SLF4	BPNN	83%
SLF7	No	—	BPNN	74%
SLF8	No	SLF7	BPNN	86%
SLF8	No	SLF7	AdaBoost	88%
SLF13	Yes	SLF7	BPNN	88%
SLF13	Yes	SLF7	MOE	90%
SLF16	Yes	SLF7 + Gabor + wavelet	MVE	92%

The classifiers used included back-propagation neural network (BPNN), mixture-of-experts (MOE), and majority voting ensemble (MVE).

HIGH-RESOLUTION 3D DATASETS

Since the automated interpretation of the 2D images was shown to be successful, the next step was to extend the analysis to 3D images.

3DHeLa

The *3DHeLa* dataset was used to demonstrate the advantages of 3D images. Therefore all patterns used in *2DHeLa* are included in *3DHeLa* to make the results from two datasets comparable (Fig. 47.2).

HeLa cells were stained and imaged by a three-laser confocal scanning microscope. The dataset includes the same nine proteins as in *2DHeLa* and there are 50 to 52 3D images in each class. In addition to a DNA channel for each image, a channel measuring total protein was acquired at the same time (by adding reactive Cy5 dye to label all proteins non-specifically). This resulted in three-color 3D images. The benefits of acquiring the two additional channels are that (1) the two channels can be used to automatically isolate individual cells, (2) the DNA channel can be used for calculating additional features, and (3) a tenth and eleventh class could be created by adding images from the DNA channel (class *DNA*) and images from the total protein channel (class *Cytoplasmic*) to the nine classes of proteins.

For each 3D image, 14 to 24 2D slices were acquired at a *z*- (axial) interval of 0.203 µm. The in-plane intervals between neighboring voxels were 0.049 µm.

3D3T3

An additional source of 3D images is the CD-tagging project, which has created a demonstration database for a number of randomly tagged proteins in NIH 3T3 cells (http://cdtag.bio.cmu.edu/www/public/). By randomly inserting a green fluorescent protein (GFP)-encoding exon into the genome of a cell using a retroviral vector, a different protein can be tagged in many different cells. For the demonstration project, 90 different clones were isolated and the identity of the tagged gene was determined by real-time polymerase chain reaction (RT-PCR) sequencing and genome database searches. For each clone, a number of 3D images were acquired from living cells by spinning-disk confocal microscopy (no parallel channels for DNA and total protein were acquired). Eight to 33 images were collected for each clone, with each image consisting of 1024 × 1024 × 31 voxels and each voxel corresponding to a 0.11 µm × 0.11 µm × 0.5 µm region of the sample. (See Fig. 47.3.)

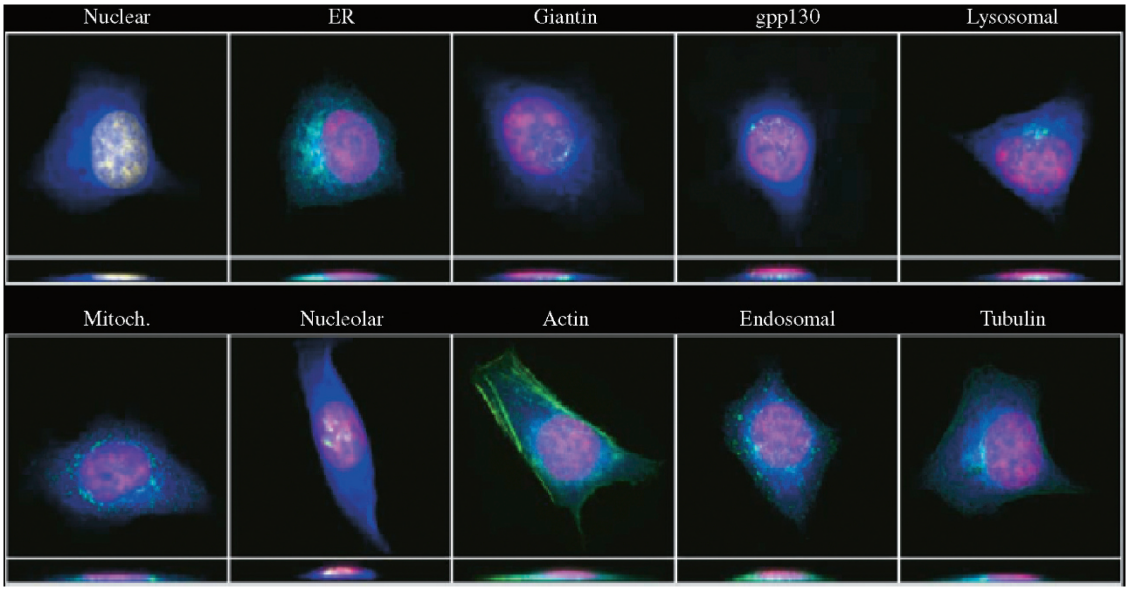

FIGURE 47.2. Typical three-color images from the *3DHeLa* dataset. Red, DNA channel; blue, total protein channel; green, protein channel. (Reprinted by permission of Carnegie Mellon University.)

Image Acquisition Considerations When Using Automated Analysis

Although the whole analysis procedure can be automated, the performance is influenced by how images are acquired. Some important points must be considered in image acquisition to get valid results.

1. In each image there should be at least one whole cell. The SLF are defined to capture information on whole, individual cells, and features from a partial cell might be very different from the features of the expected pattern. For 3D images, the slices should cover the top and bottom of the cell. Some microscopes provide axial range scanning to help decide the range of scanning. For those that don't, ensuring that the number and position of the slices are appropriate is critical.

2. If possible, cells being imaged should be well separated from each other to facilitate automated segmentation. As for a partial cell, the features of a region containing more than one cell are expected to be different from those of single cells.

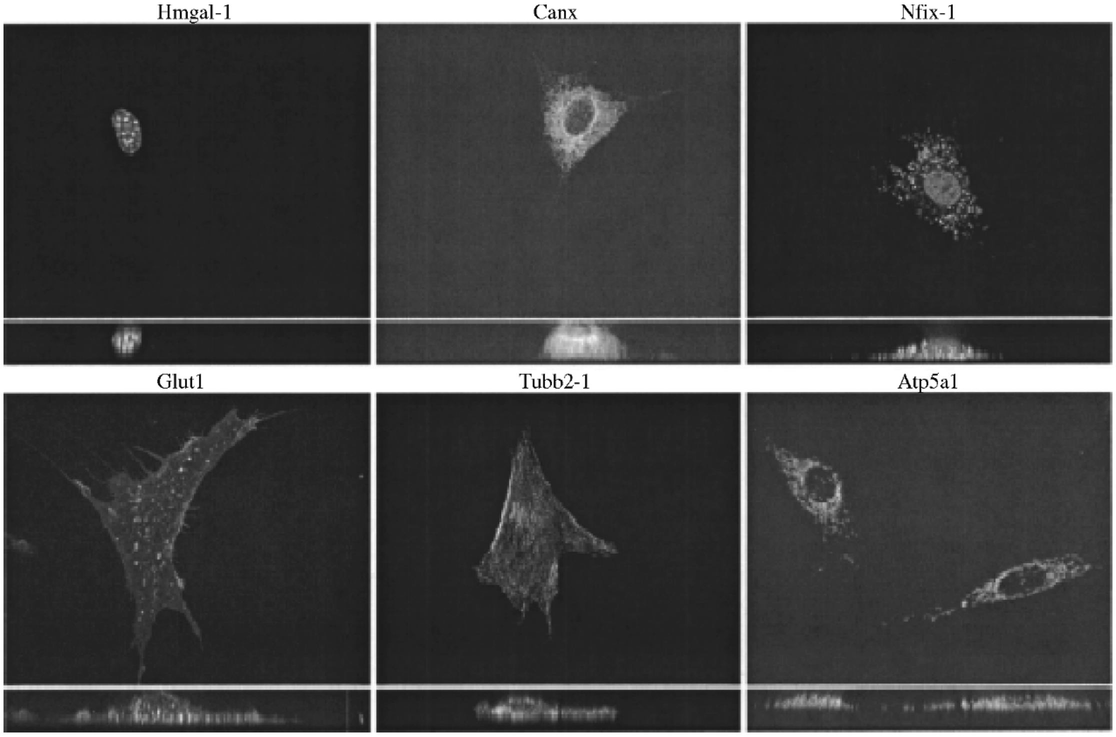

FIGURE 47.3. Example images from the *3D3T3* dataset. (Reprinted by permission of Carnegie Mellon University.)

3. Higher resolution is more desirable. An obvious reason of using higher resolution is that more information can be acquired. The Rayleigh criterion indicates that the resolution is decided by the transmitted wavelength and the numerical aperture of the objective that is used. Because the range of wavelengths is limited, the major consideration is to use an objective with a numerical aperture as large as possible. This is especially important for increasing the axial resolution because it is inversely proportional to the square of numerical aperture.

4. Sampling should be sufficient to capture as much information as possible. The two factors are the voxel dimensions and the number of gray levels in an image. The voxel size in the sample plane is determined by the total magnification and the spacing between pixels on the camera. An image with smaller physical size of each pixel shows finer structure of a location pattern and thus its feature values are more informative. However, the amount of information that can be obtained is limited. A common criterion is to use Nyquist sampling, or sampling at one half of the resolution from the Rayleigh criterion. For 520 nm light and a numerical aperture of 1.3, the Rayleigh limit is 0.244 µm. Additional magnification before the camera may be required to sample at half this value. Given the large dynamic range found in the distribution of many proteins, using a detector with greater than 8-bit image depth is also very valuable.

5. The number of images in each class should be sufficient to describe the variation of the pattern. The details of images in the same pattern are all different because of variations in size, shape, and cellular environment of each individual cell. The ability of numerical features to describe location patterns depends on how well the range of variations is covered. Because essentially all classification and clustering methods are statistical approaches, sufficient data are required to get significant results.

6. Imaging conditions should be kept as unchanging as possible. Most subcellular location features are sensitive to imaging conditions such as the objective used and exposure time. Although normalization steps can decrease such sensitivities, it is always more desirable to avoid such problems at the data acquisition step.

7. It is desirable to acquire parallel channels measuring DNA and total protein. These channels can facilitate isolating individual cells. They can also be used to develop location features because they carry information of cell morphology.

IMAGE PROCESSING AND ANALYSIS

The most critical step of automatic interpretation is the extraction of features from images. These features should be insensitive to cell rotation and translation. If images from more than one source (e.g., different microscopes) are to be analyzed together, the features should also be defined in such a way as to compensate for these differences. The procedure for moving from images to features is described in Figure 47.4.

Segmentation of Multi-Cell Images and Preprocessing

After acquiring images, preprocessing steps are necessary before feature calculation. The most important step is image segmentation, which is used to isolate individual cells from a multi-cell image. Without this step, features defined on individual cells cannot be extracted correctly. It is hard to work on the targeted protein image directly to isolate cells because we do not know how many cells there are in a particular image, and the locations of the cell boundaries relative to the targeted protein depend on which pattern the image contains. A segmentation method that is suitable for one pattern may fail for another. Fortunately, acquiring two more channels, DNA and total protein, make this problem much easier. Nuclei can easily be detected in the DNA channel by thresholding. Small, non-nuclear objects resulting from this thresholding can be easily removed by setting a second threshold on size or integrated intensity. Knowing how many cells are present and where their nuclei are greatly simplifies the problem.

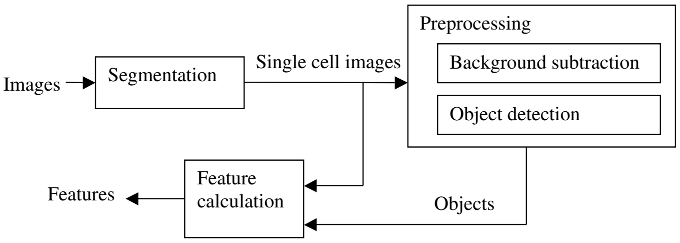

FIGURE 47.4. Feature calculation process.

With this information, a seeded watershed method (Lotufo and Falcao, 2000) can be applied. The nuclei are used as seeds and the total protein image is used to find cell boundaries. In our case, the seeded watershed algorithm was implemented by the *mmcwatershed* function from the SDC Morphology Toolbox for MATLAB (SDC Information Systems, Naperville, IL). The output from the function is a set of masks (binary images showing each separated region), each of which corresponds to one cell (and one original seed). It is possible that partial cells will be included, and therefore a criterion based on the distance to the image edge is used to remove partial cells.

By watershed segmentation, 50 to 58 individual cell images per class were obtained from the 50 to 52 original images per class acquired for the *3DHeLa* dataset. However, this segmentation method cannot be applied to the *3D3T3* dataset because there are no available images for the DNA or the total protein channel. Instead, manual segmentation was used.

After finding masks for segmentation, the following steps of preprocessing are the same for both *3DHeLa* and *3D3T3*:

1. Background fluorescence is subtracted from each image. This is done by subtracting the most common pixel value for each image from the values of all pixels and then setting any negative values in the subtracted image to 0. Some type of background correction is needed so that integrated fluorescence intensities properly reflect the fraction of cell fluorescence occupied by each object. The background correction is done before masking the image to isolate individual cells so that it can better reflect the global background (some masked regions may have no background pixels in them).

2. The segmentation masks are applied to each slice in the same 3D image.

3. An automated thresholding method (Ridler and Calvard, 1978) is used.

3D Subcellular Location Features

Subcellular location features (SLF) sets have been shown to be successful for describing 2D images, therefore, a set of 3D SLF were defined on 3D images by extending the 2D SLF. Each 3D image is composed of several 2D slices with the same size. The extension from 2D SLF to 3D SLF relies on the differences between a voxel and a pixel. First, a voxel is located in a 3D space and has three coordinates, two for the lateral plane and one for the vertical axis. Second, each voxel has 26 adjacent voxels while a pixel has only 8 neighbors. This is especially important for detecting an object.

Three categories of 2D SLF have been extended. They are morphological features, edge features, and texture features.

Morphological Features

These features compose the first set, SLF9, that was used for classifying 3D location patterns (Velliste and Murphy, 2002). SLF9.1–14 are morphological features extended from 2D SLF1.1–8 and SLF2.17–22. For this purpose, objects were defined as a contiguous set of above-threshold pixels (using 26-neighbor connectivity). The features were defined as:

SLF9.1: The number of fluorescent objects in the image.

SLF9.2: The Euler number of the image. This is the difference between number of objects and number of holes in the image.

SLF9.3: The average object size. The size of an object is defined as the number of voxels in the object.

SLF9.4: The standard deviation of the object size.

SLF9.5: The ratio of size of the largest object to the smallest.

Note: SLF9.3–9.5 describe the statistical distribution of object size in a cell.

SLF9.6: The average object distance to the center of fluorescence (COF).

SLF9.7: The standard deviation of object distances from the COF.

SLF9.8: The ratio of the largest to the smallest object to COF distance.

Note: SLF9.6–9.7 describe the spatial distribution of objects in a cell.

SLF9.9: The average object distance from the COF of the DNA image.

SLF9.10: The standard deviation of object distances from the DNA COF.

SLF9.11: The ratio of the largest to the smallest object to DNA COF distance.

Note: SLF9.9–9.11 describe the spatial distribution of objects relative to the nucleus in a cell.

SLF9.12: The distance between the protein COF and the DNA COF.

SLF9.13: The ratio of the volume occupied by protein to that occupied by DNA.

SLF9.14: The fraction of the protein fluorescence that colocalizes with DNA.

SLF9.15 to SLF9.28 were designed to describe the in-plane and axial distribution of the protein. This is especially useful for polarized cells. The membranes of these cells are divided into sides, apical, and basolateral domains. They contact different environments and some proteins must be localized to one of the two domains to carry out their function. For example, the physiologic properties of transporting epithelia are determined by the polarized distributions of ion transporters (Muth and Caplan, 2003). For nonpolarized cells, the vertical distribution is also significant because they can become polarized when attaching to a surface to grow.

SLF9.15: The average horizontal distance of objects to the protein COF.

SLF9.16: The standard deviation of object horizontal distances from the protein COF.

SLF9.17: The ratio of the largest to the smallest object to protein COF horizontal distance.

SLF9.18: The average vertical distance of objects to the protein COF.

SLF9.19: The standard deviation of object vertical distances from the protein COF.

SLF9.20: The ratio of the largest to the smallest object to protein COF vertical distance.

SLF9.21: The average object horizontal distance from the DNA COF.

SLF9.22: The standard deviation of object horizontal distances from the DNA COF.

SLF9.23 The ratio of the largest to the smallest object to DNA COF horizontal distance.

SLF9.24: The average object vertical distance from the DNA COF.

SLF9.25: The standard deviation of object vertical distances from the DNA COF.

SLF9.26: The ratio of the largest to the smallest object to DNA COF vertical distance.

SLF9.27: The horizontal distance between the protein COF and the DNA COF.

SLF9.28: The signed vertical distance between the protein COF and the DNA COF.

Edge Features

Edge features were originally developed for the *2DHeLa* dataset to distinguish proteins distributed along long and thin edges from other proteins. For convenience, two edge features were added to 3D SLF by a small modification. Edge detection was performed in each of the 2D slices instead of using a 3D edge-detection method (so that the different resolution of the lateral plane and axial axis does not have to be considered). The Canny method (Canny, 1986) was used for edge detection because it is less sensitive to noise than some other methods.

The two edge features were added to SLF9.1–9.8 and SLF9.15–9.20 to create SLF11:

3D-SLF11.15 The fraction of above-threshold pixels that are along an edge.

3D-SLF11.16 The fraction of fluorescence in above-threshold pixels that are along an edge.

Texture Features

The extension of texture features to 3D requires calculating a co-occurrence matrix in 13 directions because each pixel has 26 neighbors. The mean and range over all 13 directions were calculated for each statistical property of the co-occurrence matrix. This resulted in 26 texture features for each image:

SLF11.17/30 Average/range of angular second moment.
SLF11.18/31 Average/range of contrast.
SLF11.19/32 Average/range of correlation.
SLF11.20/33 Average/range of sum of squares of variance.
SLF11.21/34 Average/range of inverse difference moment.
SLF11.22/35 Average/range of sum average.
SLF11.23/36 Average/range of sum variance.
SLF11.24/37 Average/range of sum entropy.
SLF11.25/38 Average/range of entropy.
SLF11.26/39 Average/range of difference variance.
SLF11.27/40 Average/range of difference entropy.
SLF11.28/41 Average/range of info measure of correlation 1.
SLF11.29/42 Average/range of info measure of correlation 2.

AUTOMATED CLASSIFICATION OF LOCATION PATTERNS

The task of classification is to automatically recognize the pattern of an image. The SLF of images whose pattern is known can be used to train a classifier to recognize an image whose pattern is not known (a service called Subcellular Localization Image Classifier, or SLIC, is available on our Web site at http://murphylab.web.cmu.edu/services/PSLID/). The output of the classifier can only be one of the predefined patterns. So, if we want a classifier to recognize a new pattern, the classifier must be fed with a training set corresponding to the new pattern.

Classification of 3DHeLa Dataset

Feature Normalization

Normalization was performed for all features before feeding them into a classifier. This was done by translation and scaling so that each feature has the mean 0 and variance 1. Otherwise, some features will dominate others during calculation. For example, the number of objects in an image (SLF9.1) can be thousands, but the two edge features are all between 0 and 1. Although no additional information was generated after feature normalization, this step is helpful for training the classifier. Some classifiers like a neural network tend to look for borders in the subspace of dominant features even if these are not good for distinguishing classes. Because the features were divided into a training set and a test set, the normalization factors should be calculated on the training set and then the same factors are used to normalize the test set.

Feature Selection

As discussed above, a very useful procedure at the feature level is feature selection. As shown in 2D classification, more features do not necessarily mean higher classification accuracy. Besides requiring a longer time for training, a classifier may overfit with some noisy features. SDA can be used to deal with this problem by finding a feature subspace in which different classes are well separated while, at the same time, samples from the same class are close to each other. The measurement of this property is defined as the ratio of the within-group covariance matrix to the among-group covariance matrix. The SDA was implemented by the stepdisc function of SAS (SAS Institute, Cary, NC) with default parameter values. The feature set SLF10 is composed of 14 features selected from SLF9 by SDA.

Classification Results

As was done for the *2DHeLa* dataset, we have evaluated different feature sets and classifiers. At first SLF9 and a BPNN with one hidden layer and 20 hidden nodes was compared with classification results from SLF2. This resulted in 91% accuracy, which is higher than the counterpart of 2D classification (76%) and close to the best 2D results (92%). When an eleventh cytoplasmic pattern was added, the classification on the 11 classes was also 91%.

As expected, higher accuracy was obtained when SLF10, selected by SDA from SLF9, was used. This feature set gave 95% accuracy averaged over 10-fold cross-validation (Huang and Murphy, 2004b) using the same classifier. The best classification accuracy (96%, Table 47.3) on the feature set was achieved using an optimal majority-voting ensemble classifier, in which the eight base classifiers were the same as those used on *2DHeLa*.

Downsampled Images with Different Gray Scales

There are two motivations for exploring downsampling and gray-scale binning before calculating texture features. The first is that this is a step to normalize images from different sources. The second is that this may improve the texture features by focusing on the most appropriate spatial frequencies. Texture features calculated at three different pixel resolutions (0.2, 0.4, 1.0µm) and three different numbers of gray levels (16, 64, 256) have been explored (Chen and Murphy, 2004). To evaluate the contribution of texture features under different conditions, results from different features were compared. We found that higher resolution and more gray levels tend to give better performance (Table 47.4). Figure 47.5 shows the changes of classification accuracy with number of features. Each curve corresponds to a specific feature set. Surprisingly, one of the best classification accuracies was achieved by only seven features (Table 47.5), that were then defined as SLF17. The average accuracy was 98% (Table 47.6), which is the most accurate classification that has ever been obtained. This means that the 10 patterns can be well described in a 7D space, which even has fewer degrees of freedom than the number of the patterns.

TABLE 47.3. Confusion Matrix for Classification of 11 Classes of *3DHeLa* Images Using the Feature Set SLF10

	Output of the Classifier (%)										
True Class	Cyt	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
Cytoplasmic	**100**	0	0	0	0	0	0	0	0	0	0
DNA	0	**98**	0	0	0	0	0	2	0	0	0
ER	0	0	**97**	0	0	0	0	0	2	0	2
Giantin	0	0	0	**98**	0	2	0	0	0	0	0
GPP130	0	0	0	4	**96**	0	0	0	0	0	0
LAMP2	0	0	0	2	2	**96**	0	0	0	0	0
Mitochondria	0	0	0	4	0	0	**95**	0	2	0	0
Nucleolin	0	0	0	0	0	0	0	**100**	0	0	0
Actin	0	0	2	0	0	0	2	0	**95**	2	0
TfR	0	0	0	0	0	6	4	0	2	**85**	4
Tubulin	0	0	4	0	0	0	0	0	0	2	**94**

The results are averages over 10-fold cross-validation trials on an optimal majority-voting ensemble classifier. The overall average accuracy was 96%. (Data from Huang and Murphy, 2004b.)

TABLE 47.4. The Contents of Each Cell in the Table Are the Net Increase in Percentage Accuracy Over a Classifier Using No Texture Features

Pixel	Number of gray levels		
resolution	256	64	16
0.2 µm	7.7 (16)	5.5 (22)	4.7 (15)
0.4 µm	7.5 (7)	5.3 (18)	3.8 (34)
1.0 µm	3.8 (28)	2.6 (23)	0.1 (10)

The number included in each parenthesis is the number of features selected by SDA. [Data from Chen and Murphy (2004).]

TABLE 47.5. The Seven Features in SLF17 that Give 98% Classification Accuracy on the *3DHeLa* Dataset

SLF9.4	The standard deviation of object volumes
SLF11.16	The fraction of fluorescence in above threshold pixels that are along an edge
SLF11.21	The average inverse difference moment
SLF11.27	The average difference entropy
SLF11.28	The first average information measure of correlation
SLF11.40	The range of the first information measure of correlation
SLF11.42	The range of the second information measure of correlation

CLUSTERING OF LOCATION PATTERNS: LOCATION PROTEOMICS

One goal of proteomics is to study the relations between proteins and group them into meaningful categories. This has been done at the sequence level by sequencing all proteins and clustering them hierarchically according to their sequence similarity. For location proteomics, the clustering should be done by their localization similarity. Our work on automated interpretation of confocal microscope images of cellular proteins allows us to do this, given sets of images for large numbers of proteins. The approaches we have used are described below for the images in the *3D3T3* dataset. We refer to a cluster tree representing a set of subcellular location patterns as a subcellular location tree (SLT).

Exclusion of Outliers

To measure the similarity between the locations of two proteins, their images must be acquired under similar conditions. These conditions include not only the parameters of the imaging conditions, but also the state of the cells. To describe the typical location pattern of a protein, images of normal interphase cells are usually preferred. But biologists may want to randomly choose cells for imaging to avoid observer bias. So some unusual examples of a pattern were included in the *3D3T3* dataset. Cells in these images include dead cells, dying cells, cells just before mitosis, and cells just after cytokinesis. The influence of the outliers increases with their fraction in the whole dataset. Because of the small size of the *3D3T3* dataset (8–33 images per class), there is high risk that the outliers will bias the clustering results. Therefore, the outliers were removed by a stringent procedure discussed below.

Determination of Optimal Clustering

A number of algorithms and distance functions have been used for clustering. Because different feature sets and distance functions result in different clusters or trees, it is necessary to determine which one is the best. A simple way is to visually inspect the tree and determine how consistent it is with biological knowledge and the visual interpretation of the images. However, empirical evaluation is impractical while there are so many classes in the data (or when not all of the protein patterns are known!). Some measurement of goodness should be defined to automate the process. The basic criterion is having enough partitions to separate all distinguishable patterns while at the same time keeping indistinguishable patterns in the same group. The trade-off implies the existence of an optimal partitioning.

As a starting point, all of the *3D3T3* images were clustered by the unsupervised *k*-means method with varying cluster numbers. The optimal cluster number was decided using the Akaike Information Criterion (AIC), which balances the fitness and simplicity of the model represented by the clusters. The results showed that the optimal number of clusters was 17. They also showed that three of the 3T3 clones had their images spread over a number of clusters. We interpreted this as evidence that these clones were either not pure cell lines or that the tagged protein varied extensively in location (perhaps as a function of cell cycle). These 3 clones were therefore removed and all further studies were based on the 87 clones. These initial clustering results were also used to remove outliers: images that were not contained in the major cluster of each clone were removed from consideration.

When all data are used to build a cluster tree, the tree structure can be very dependent on which particular examples from a

TABLE 47.6. Confusion Matrix for Classification of 11 Classes of *3DHeLa* Images Using the Feature Set SLF17

	Output of the Classifier (%)									
True Classification	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	**98**	2	0	0	0	0	0	0	0	0
ER	0	**100**	0	0	0	0	0	0	0	0
Giantin	0	0	**100**	0	0	0	0	0	0	0
GPP130	0	0	0	**96**	4	0	0	0	0	0
LAMP2	0	0	0	4	**95**	0	0	0	0	2
Mitochondria	0	0	2	0	0	**96**	0	2	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	0	0	0	2	0	0	0	**96**	2
Tubulin	0	2	0	0	0	0	0	0	0	**98**

The results are averages over 10-fold cross-validation trials on a BPNN classifier with one hidden layer and 20 hidden nodes. The overall average accuracy was 98%. [Data from Chen and Murphy (2004).]

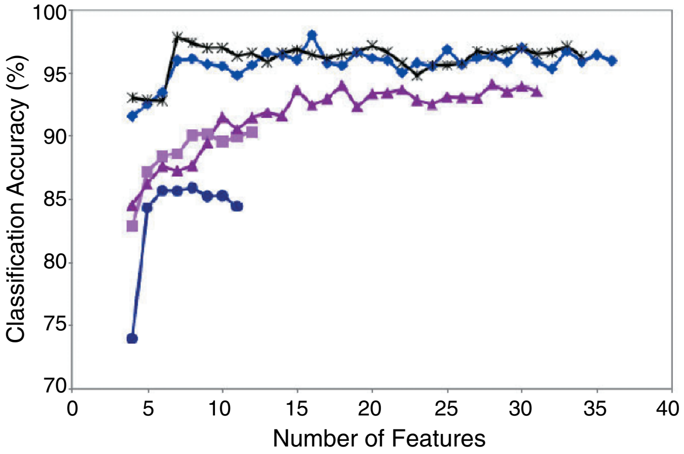

FIGURE 47.5. Classification accuracy over 10 major classes in *3DHeLa* dataset. Each curve represents one of five subsets of SLF11, morphological features only (*closed circle*), morphological and edge features (*closed square*), and morphological, edge, and texture features computed at 256 gray levels and different pixel resolutions: 0.2 µm (*closed diamond*), 0.4 µm (*), and 1.0 µm (*closed triangle*). [Data from Chen and Murphy (2004).]

cluster are present. Therefore, rather than building one tree, we have constructed many trees from random subsets of the images for each clone. The common properties of these trees were extracted to form a consensus tree (Chen and Murphy, 2005). In order to convert the tree into a specific number of clusters, we assumed that the number of clusters obtained by *k*-means/AIC analysis, 17, was correct and cut the consensus tree accordingly.

Many clustering methods use some distance function to measure which observations are similar to each other and should be grouped. As a criterion to choose an optimal distance function, we reason that a good distance function should give consistent results from different clustering methods (Chen and Murphy, 2005). Two other clustering methods independent of the distance function were also applied on the data. One is implemented based on the confusion matrix from a classifier to group easily confused clones together. The other was done by visual inspection. Finally, we had results from four clustering methods, consensus tree, *k*-means/AIC, confusion matrix based method (ConfMat), and visual inspection. Then we need a numerical value to measure the agreement between clusters from any two methods. This value was described by Cohen's κ statistic (Cook, 1998),

$$\kappa = \frac{\text{Observed agreement} - \text{expected agreement}}{1 - \text{expected agreement}}.$$

TABLE 47.7. Evaluation of Consistency of Clustering Methods Based on Different Distance Functions

	z-Scored Euclidean Distribution (κ)	Mahalanobis Distribution (κ)
k-means/AIC vs. consensus	1	0.5397
k-means/AIC vs. ConfMat	0.4171	0.3634
Consensus Vs. ConfMat	0.4171	0.1977
k-means/AIC vs. visual	0.2055	0.1854
Consensus vs. visual	0.2055	0.1156

Data from Chen and Murphy (2005).

Observed agreement is defined as the portion of protein pairs where the two clustering results agree and expected agreement is defined as the agreement between two random clusterings with the same distribution frequencies as the two clusterings being compared (these are obtained by simulation). Two distance functions, *z*-scored Euclidean and Mahalanobis, were tried for building the consensus tree and applying *k*-means/AIC. The results showed that Euclidean distance always results in better agreement (Table 47.7).

The results above allowed us to choose the free parameters to build a consensus tree that is optimal according to the criteria discussed above. Figure 47.6 shows the consensus tree built using these optimal parameters. Visual inspection suggests that this tree is consistent with available biological information. For example, it was found that all nucleolar proteins, including Rpl32, Unknown-25, and Unknown-32, are grouped together. The two groups representing nucleus only and nucleus–cytoplasm mixture were also well separated.

STATISTICAL COMPARISON OF LOCATION PATTERNS

Besides classification and clustering, another interesting task for biologists is to determine whether two patterns are different. A pair of patterns for comparison could be either patterns from different proteins or patterns for the same protein under different experimental conditions. Because the difference between two location patterns implies different functions of the protein, comparing two protein location patterns can help distinguish diseased cells from healthy ones, evaluate drug effects, and find similarities between two proteins.

Although we can tell something about whether two patterns are different or not from a confusion matrix or a cluster tree, it is hard to tell how confident we are in any conclusion. A better way to compare two location patterns is to use a hypothesis test, in which the null hypothesis is that the location patterns are the same. Because we can calculate a feature matrix for each pattern, the null hypothesis becomes that the two feature matrices are from the same population. The hypothesis test generates a value called a test statistic to measure the difference between two groups of location patterns. Usually the test statistic has a known statistical distribution under the null hypothesis. So we can calculate a *p* value, which is the probability that the test statistic is as much or more extreme than the observed one if the null hypothesis is true. Thus, the two location patterns are considered to be different if the *p* value calculated from their feature matrices is smaller than some chosen criterion, such as 0.05.

There are many ways to define a test statistic and then calculate the *p* value. A widely used method is the Hotelling T^2 test, which is the multivariate extension of the *t* test. This test results in a test statistic or *F* value measuring the difference between the means of the two feature matrices by considering the correlation of the features. For this test, the total number of images must be greater than the features. This is reasonable because sufficient data is needed to describe the correlation and variance of the features.

This test has been applied on the *2DHeLa* data (Roques and Murphy, 2002). It can be easily applied to the 3D images because after feature calculation both 2D images and 3D images are represented by a feature vector. The Hotelling T^2 test on *3DHeLa* showed that the *p* values for any pair is almost 0 (unpublished data), which means that all patterns are statistically different. The smallest *F* value (96.8; $p < 0.01$) was from the gpp130-giantin pair, which is consistent with the results of *2DHeLa*.

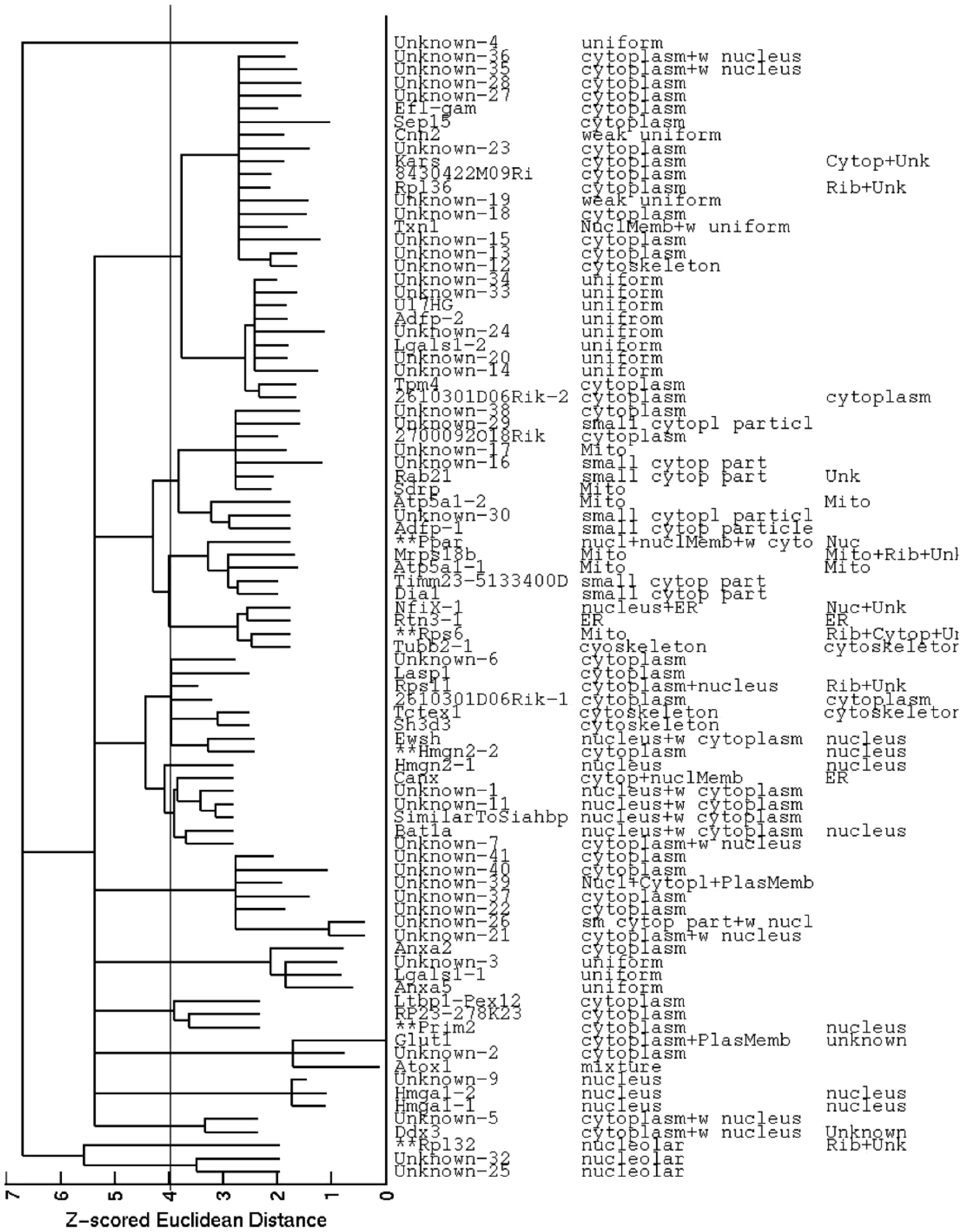

FIGURE 47.6. An optimal consensus tree for the 87 *3D3T3* clones. The text columns show protein names (if any), location description assigned by visual examination, and location description from protein data bases. The tree is also available in a Web page that can display representative images for any branch (http://murphylab.web.cmu.edu/services/PSLID/). From Chen and Murphy (2005).

A Web service for comparison of two location patterns, Statistical Image Experiment Comparator (SimEC) is also available through http://murphylab.web.cmu.edu/services/PSLID/.

IMAGE DATABASE SYSTEMS

The methods described in this chapter are all designed to be carried out in an automated fashion on large numbers of images. Indeed, all of them perform significantly better when the number of available images increases. They also rely on knowledge of the conditions under which an experiment was done and each image was acquired. It is therefore natural to consider an image database system as the most appropriate framework to utilize these tools. We have described such a system (PSLID, for Protein Subcellular Location Image Database) for storing images and their annotations with a focus on providing tools for analyzing subcellular patterns as an integral part of the database (Huang *et al*., 2002). Other image database systems have been described with a focus on cataloguing published images from both fluorescence and electron microscopy (Bioimage) (Carazo and Stelzer, 1999) or on supporting images from many fluorescence microscope systems and biological applications (OME) (Swedlow *et al*., 2003). A number of efforts to improve and expand the capabilities of databases for microscope images are ongoing, and it is anticipated and hoped that a consensus on standards for such systems will emerge to minimize duplication of effort (see Chapter 50, *this volume*).

FUTURE DIRECTIONS

In this chapter we have introduced how automated image analysis methods can be used to interpret 3D confocal microscope images. The core of the whole process is to convert an image to a numerical feature vector. The conversion allows us to apply any available multivariate analysis method to the data. As we have shown, we can perform classification, clustering and comparison on these images automatically. The whole procedure provides a high throughput tool for analyzing data from 3D confocal microscopy. Given these past successes, it is worth considering directions for future work in this area.

The images in the *3D3T3* dataset were segmented manually into single cell images because of the difficulty of automated segmentation without parallel DNA or total-protein channels. The problem becomes more difficult for tissue images in which cells are not well separated. Segmentation of tissue images has been accomplished by collecting parallel DNA and membrane protein images and using active contour methods (De Solorzano *et al*., 2001). An alternative that does not require the parallel images is to develop SLF that are insensitive to the number of cells in the image. We can calculate such features directly from multicell images without segmentation. This approach has shown to be successful for 2D images containing multiple cells (Huang and Murphy, 2004a).

The SLT shown in Figure 47.6 was obtained only for location patterns in interphase cells. To build such a tree, the images of cells under unusual states such as mitosis were removed as outliers. But this does not mean that these outliers do not contain useful information. When adequate images are acquired for each state, a bigger database containing location patterns for different cell states can be created. This is another advantage of determination methods over prediction.

Protein localization is a dynamic process. For example, some proteins are localized to ER or ER membrane after synthesis, and are then sorted to different compartments. The characteristics of such a sorting procedure cannot be described unless we do the analysis on the images of a 3D time series. We are currently extending our methods from 3D to 4D (space and time). Similarly, we can extract features describing movement of proteins and define a new set of SLF that will highlight these features.

Because we have shown that SLF is useful for both *3DHeLa* and *3D3T3*, one interesting question is whether the analysis would be valid if the datasets were mixed, for example, building a SLT for all the patterns in the two datasets. This suggests an even more challenging task: to develop a feature set insensitive to cell type, tagging method, or imaging method. This is an important goal given the number of different confocal microscopes being used and the number of different cell types being studied. Nonetheless, when applied under controlled conditions, the methods described here have been proven to be more objective, reproducible, and sensitive for characterizing the proteome of a particular cell type or comparing protein distribution patterns in the presence and absence of a drug than were visual examination methods.

REFERENCES

Bhasin, M., and Raghava, G.P.S., 2004, ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST, *Nucleic Acids Res.* 32:W414–W419.

Boland, M.V., and Murphy, R.F., 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, M.V., Markey, M.K., and Murphy, R.F., 1997, Classification of protein localization patterns obtained via fluorescence light microscopy, In: *19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, Chicago, pp. 594–597.

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. Machine Intell.* 8:679–698.

Carazo, J.M., and Stelzer, E.H.K., 1999, The BioImage database project: Organizing multidimensional biological images in an object-relational database, *J. Struct. Biol.* 126:97–102.

Chen, X., and Murphy, R.F., 2004, Robust classification of subcellular location patterns in high resolution 3D fluorescence microscope images, In: *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, San Francisco, pp. 1632–1635.

Chen, X., and Murphy, R.F., 2005, Objective clustering of proteins based on subcellular location patterns, *J. Biomed. Biotechnol.* 2:87–95.

Chen, X., Velliste, M., Weinstein, S., Jarvik, J.W., and Murphy, R.F., 2003, Location proteomics — Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins, *Proc. SPIE* 4962:298–306.

Chou, K., and Cai, Y., 2002, Using function domain composition and support vector machines for prediction of protein subcellular location, *J. Biol. Chem.* 277:45765–45769.

Chou, K.-C., and Elrod, D.W., 1999, Protein subcellular location prediction, *Protein Eng.* 12:107–118.

Cook, R., 1998, Kappa, In: *The Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds.), John Wiley and Sons, New York, pp. 2160–2166.

De Solorzano, C.O., Malladi, R., Lelievre, S.A., and Lockett, S.J., 2001, Segmentation of nuclei and cells using membrane related protein markers, *J. Microsc.* 201:404–415.

Dreger, M., 2003, Proteome analysis at the level of subcellular structures, *Eur. J. Biochem.* 270:589–599.

Huang, K., Lin, J., Gajnak, J.A., and Murphy, R.F., 2002, Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database, In: *2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002)*, IEEE, Chicago, pp. 325–328.

Huang, K., and Murphy, R.F., 2004a, Automated classification of subcellular patterns in multicell images without segmentation into single cells, In: *2004 IEEE International Symposium on Biomedical Imaging (ISBI-2004)*, IEEE, Chicago, pp. 1139–1142.

Huang, K., and Murphy, R.F., 2004b, Boosting accuracy of automated classification of fluorescence microscope images for location proteomics, *BMC Bioinformatics* 5:78.

Huang, K., Velliste, M., and Murphy, R.F., 2003, Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images, *Proc. SPIE* 4962:307–318.

Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., et al., 2002, Subcellular localization of the yeast proteome, *Genes Dev.* 16:707–719.

Lotufo, R., and Falcao, A., 2000, The ordered queue and the optimality of the watershed approaches, In: *Mathematical Morphology and Its Application to Image and Signal Processing*, (J. Goutsias, L. Vincent, and D.S. Bloomberg, eds.), Kluwer Academic Publishers, New York.

Murphy, R.F., Velliste, M., and Porreca, G., 2003, Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images, *J VLSI Sig. Proc.* 35:311–321.

Muth, T.R., and Caplan, M.J., 2003, Transport protein trafficking in polarized cells, *Annu. Rev. Cell Dev. Biol.* 19:333–366.

Park, K.J., and Kanehisa, M., 2003, Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs, *Bioinformatics* 19:1656–1663.

Ridler, T.W., and Calvard, S., 1978, Picture thresholding using an iterative selection method, *IEEE Trans. Syst. Man. Cybernet.* 8:630–632.

Roques, E.J.S., and Murphy, R.F., 2002, Objective evaluation of differences in protein subcellular distribution, *Traffic* 3:61–65.

Swedlow, J.R., Goldberg, I., Brauner, E., and Sorger, P.K., 2003, Informatics and quantitative analysis in biological imaging, *Science* 300:100–102.

Velliste, M., and Murphy, R.F., 2002, Automated determination of protein subcellular locations from 3D fluorescence microscope images, In: *2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002)*, IEEE, Bethesda, Maryland, pp. 867–870.