Automated Interpretation of Protein Subcellular Location Patterns

Xiang Chen* and Robert F. Murphy*,†
*Department of Biological Sciences, Center for Automated Learning and Discovery and Center for Bioimage Informatics
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
†Department of Biomedical Engineering, Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Proteomics is a major current focus of biomedical research, and location proteomics is the important branch of proteomics that systematically studies the subcellular distributions for all proteins expressed in a given cell type. Fluorescence microscopy of labeled proteins is currently the main methodology to obtain location information. Traditionally, microscope images are analyzed by visual inspection, which suffers from inefficiency and inconsistency. Automated and objective interpretation approaches are therefore needed for location proteomics. In this article, we briefly review recent advances in automated imaging interpretation tools, including supervised classification (which assigns location pattern labels to previously unseen images), unsupervised clustering (which groups proteins based on the similarity among their subcellular distributions), and additional statistical tools that can aid cell and molecular biologists who use microscopy in their work.

KEY WORDS: Location proteomics, Subcellular location features, Fluorescence microscopy, Cluster analysis, Protein distribution comparison, CD-tagging, Systems biology.

0074-7696/06 $35.00
DOI: 10.1016/S0074-7696(06)49004-5

I. Introduction

With the development of high-throughput analysis techniques, the approaches used for biological and biomedical research have been fundamentally changed. The completion of sequencing for dozens of genomes has revolutionized the way we think about acquiring biological data. We can now imagine projects to acquire comprehensive data in a reasonable time frame for a specific characteristic (sequence, structure, function, interaction, etc.) for a complete set of molecules (DNA, RNA, protein, lipids, etc.) in a given organism.

Although DNA contains all of the genetic information for an organism, it is largely protein expression that differentiates cell types within that organism. Therefore, the focus of biological research has shifted from genomics, which mainly studies sequence information, to proteomics, where the central goal is to characterize protein functionality.

Proteomics systematically characterizes different properties of all proteins in a given cell type or tissue, including their sequences, expression levels, structures, functions, regulations, interactions, and location patterns. Knowledge of the location pattern of a protein is necessary for a complete understanding of its function. There are a number of ways in which this is true. First, location pattern changes often correlate with activity changes. For example, it has been shown that activation of the *rgr* oncogene is partially associated with a change of its location pattern from the endomembrane network to the plasma membrane, which facilitates interaction with RAS and activates the RAS downstream pathways (Hernandez-Munoz *et al.*, 2003). Activity of p53 is also regulated by its location (O'Brate and Giannakakou, 2003). The protective effects of extracellular signal-regulated kinase 2 (ERK2) against apoptogenic stimuli are also dependent on the cellular location of ERK activation (Ajenjo *et al.*, 2004). Second, protein mislocalization is often associated with disease. For example, while lamin B receptors (LBR) are located in the inner nuclear membrane in normal cells, they are largely distributed in the cytoplasm in cells carrying a mutation that causes an autosomal-dominant form of Emery–Dreifuss muscular dystrophy (Reichart *et al.*, 2004). Third, the relationship between two proteins' locations can be used to support (or question) protein interaction results found in other experiments (such as those from yeast two-hybrid screening) since components of a single protein complex may be expected to have the same distribution pattern. We have coined the term location proteomics to describe the systematic study of protein location patterns (Chen *et al.*, 2003). It requires (1) methods to either experimentally determine or predict location patterns for all proteins, and (2) methods to systematically organize the set of possible locations where a protein can be found. This chapter reviews work

focused on automated experimental determination and discovery of protein location patterns.

II. Subcellular Location of Proteins

A. Description, Prediction, and Determination

1. Gene Ontology

Currently the most systematic approach to define and organize the set of all possible location patterns is the set of terms for "cellular component" in the Gene Ontology (GO) (Harris *et al.*, 2004). Cellular component is one of the three general categories defined in the GO database, describing locations at subcellular structure (such as plasma membrane) and macromolecular complex (such as the ubiquitin ligase complex) levels. The "cell" term (GO:0005623) in GO describes all components within and including plasma membrane as well as extracellular structures. The cellular component ontology is represented as a directed acyclic graph (DAG) in which location terms are grouped into higher order structures. A DAG differs from a tree-structured hierarchy in that a child can have multiple parents (i.e., multiple paths from the root to a specific node). For example, two GO codes are found for mouse Atp5a1 protein (ATP synthase, H^+ transporting, mitochondrial F1 complex, α subunit, isoform 1): GO:0005739 (mitochondrion) and GO:0005615 (extracellular space). A portion of the cellular component ontology leading to GO:0005739 (mitochondrion) is shown in Fig. 1.

2. Overview of Current Methodology for Protein Location Prediction and Determination

Genome sequencing projects have produced large numbers of putative amino acid sequences that are often largely unannotated. To learn protein functions in the context of their subcellular organelle, it is crucial either to experimentally determine their location pattern or to predict their subcellular location from a sequence.

a. Protein Subcellular Location Prediction In the past decade, many efforts have been made to develop location prediction methods using different approaches. One category of approaches is based on protein sequence similarity, such as LOCKey (Nair and Rost, 2002) and Proteome Analyst (Lu *et al.*, 2004). A protein sequence is searched against a set of experimentally annotated sequences, and text features are extracted from homologs and

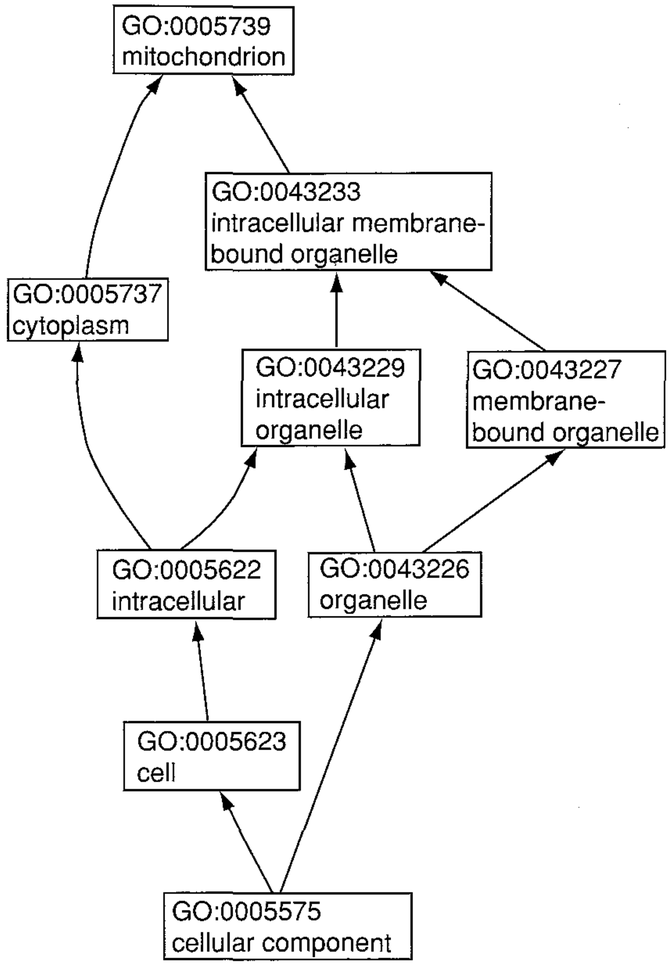

FIG. 1 Portion of cellular component ontology for GO:0005739 (mitochondrion).

used to predict the location pattern. However, this approach suffers when there is no significant match between the query protein and proteins in the training set. The second set of approaches is based on amino acid composition (or its variations), and includes NNPSL (Reinhardt and Hubbard, 1998) and SubLoc (Hua and Sun, 2001). Information about amino acid compositions and the correlation between amino acids is used as features for prediction. Although these approaches utilize general information of proteins, it is not clear whether they can capture enough information to distinguish the difference between some slightly different patterns (such as endosome and lysosome). A third category is based on motif finding, such as TargetP (Emanuelsson *et al.*, 2000). This approach tries to identify motifs that are important for localization (such as signal peptides) in the target sequence and use this information for prediction. The major limitations for this approach are (1) not all proteins in the same subcellular location have the same motif and (2) many genes in automatically annotated genomes have

unreliable 5′ sequences. Most current methods, such as PSORT (Nakai and Horton, 1999; Nakai and Kanehisa, 1992), LOC3D (Nair and Rost, 2003), and LOCtarget (Nair and Rost, 2004) use a mixture of these approaches to achieve better performance. Current research in this area focuses on (1) design of more discriminating sequence descriptors, (2) improvement of classification algorithms, and (3) incorporation of Gene Ontology information into the classification scheme.

Even with recent improvements, current prediction algorithms still suffer from two major limitations: (1) unsatisfactory prediction accuracy, especially for some organelles (such as mitochondria), and (2) limited coverage, both for the number of location patterns and for the number of taxonomic categories covered in prediction.

b. Protein Subcellular Location Determination Protein subcellular location prediction algorithms require a large database of proteins with experimentally determined locations. Without major improvement in automated determination of protein location patterns, the prediction methods will inevitably suffer from limited prediction accuracies and limited coverage. For example, many similar but still statistically different location patterns exhibited by different proteins in the same organelle or compartment cannot be distinguished by predictive schemes until there are sensitive and consistent methods to identify and systematically describe these slight differences in the first place.

Different approaches can be employed to determine protein location patterns. For example, differential centrifugation separation can be used to obtain subcellular fractions (Hardonk *et al.*, 1977) and proteins contained in each fraction can be identified by two-dimensional polyacrylamide gel electrophoresis, enzymatic digestion of separated protein spots, and mass spectrometry analysis. This approach is usually labor-intensive, and the resolution of differential centrifugation separation is limited. Recently, efforts have been made to create a high-throughput protocol for this approach (Jiang *et al.*, 2004).

Different microscopy technologies can also be employed for protein location study. For example, electron microscopy (Subramaniam and Milne, 2004), which provides ultrahigh resolution of the specimen, can be used to achieve precise localization information (Lujan, 2004). Because electron microscopy cannot be carried out on live cells, fluorescence microscopy has become the most commonly used technique to study protein distribution within a cell and the relationships between different proteins (i.e., colocalization of two proteins) (Brelie *et al.*, 2002).

Traditionally protein locations are determined from images by visual inspection. For large-scale, systematic location proteomics, however, visual inspection is no longer feasible since it is labor-intensive and can be highly

subjective. Unfortunately, while the determination of sequence, expression, and even structure has been automated and large-scale high-throughput projects have been initiated (Macbeath, 2002; Norin and Sundstrom, 2002), the automated determination of protein location patterns has just begun. Here we review current achievements for automated interpretation of protein subcellular location patterns from fluorescence microscopy images.

B. Fluorescent Labeling of Proteins

Modern fluorescence microscopy combines high-performance optical components with digital image acquisition components and computerized control to allow imaging of cells and tissues with a combination of temporal and spatial resolution and sensitivity achievable by no other method. Confocal approaches use spatial filtering to reduce or eliminate out-of-focus light and permit images of thin optical slices to be acquired. This provides confocal microscopy with the ability to collect a series of image sections in a thick specimen, and has led to the tremendous popularity of confocal microscopy (and its cousin, two-photon microscopy) in recent years.

While cells have some intrinsic fluorescence, little information can be obtained by fluorescence microscopy of unstained cells. The coupling of fluorophores and target proteins is therefore the key step in preparation for a fluorescence microscopy experiment. In general, there are two types of techniques used for this purpose.

The first method, commonly referred to as *immunofluorescence*, relies on delivering *external* fluorescent molecules into cells (Fujiwara and Pollard, 1976). Cells are first *fixed* by adding a substance (e.g., paraformaldehyde) that cross-links proteins in the cell, essentially immobilizing all cellular components. This prevents the contents of the cells from washing away when the cells are *permeabilized*; i.e., when a detergent is used to fully or partially dissolve the cell membrane. With the membrane barrier out of the way it is possible to introduce desired molecules into the cell—for example, antibodies conjugated to fluorescent dyes. An alternative to using antibodies is to use other substances known to bind to a particular protein. For example, the compound phalloidin binds to F-actin, the polymerized protein that forms part of the cytoskeleton. Therefore dye-conjugated phalloidin can be used to label the actin cytoskeleton. The use of such probes is not strictly *immuno*fluorescence, but, if the probes are specific and require permeabilization for entry, it is functionally identical. There are several limitations to immunofluorescent labeling, including dependence on the existence of specific antibodies or probes that are known to bind to the protein of interest and an inability to image live cells due to the need for fixation and permeabilization (which may also disrupt cellular structures). Vital fluorescent

probes, such as probes that equilibrate preferentially into an organelle, address the latter limitation but not the former.

The second method therefore is to have fluorescent molecules *internally* generated in the cells. DNA sequences coding for a naturally fluorescent protein (such as the green fluorescent protein, GFP) can be joined to either cDNA or genomic DNA to produce a fluorescent chimeric protein. If a genomic approach is used, either a specific gene or a random location in the genome can be tagged. There have been several examples of this approach (Habeler *et al.*, 2002; Jarvik *et al.*, 1996; Rolls *et al.*, 1999; Telmer *et al.*, 2002). Random tagging of proteins can also be done with the use of small epitopes (essentially short sections of a protein) instead of fluorescent proteins (Kumar *et al.*, 2000, 2002). In this case either a fluorescent antibody against the epitope tag (which requires fixation and permeabilization) or a cell-permeant fluorescent probe that binds to specific epitopes can be used (Griffin *et al.*, 1998). Epitope tags can be much smaller than fluorescent proteins and are less likely to disrupt the function of the protein to which they are attached.

When random-tagging experiments are repeated enough times, eventually most (or possibly all) proteins in a given cell type can be labeled. Combined with fluorescence microscopy, random tagging allows comprehensive libraries of images depicting the location patterns of proteins in a given cell type to be generated.

Assuming that a (large) collection of digital images has been acquired by one or more of the above methods, the next step is to automate extraction of information.

III. Subcellular Location Patterns of Proteins

A. Automated Classification

The first basic task to be carried out by automated interpretation is to assign a location pattern label to a previously unseen image. This task can be formalized as

> Given a set of images I_{known} and corresponding labels Y_{known}, learn a mapping function f between I_{known} and Y_{known}, such that when a new image $i \notin I_{\text{known}}$ is presented the label $y = f(i)$ that is assigned is optimal according to some criteria.

Such a system is termed a classifier.

Different approaches can be used for this task. Either raw images or characteristics (termed *features*) extracted from raw images can be used as input into a classifier. Most work in the area of subcellular analysis has

utilized the latter approach (Boland and Murphy, 2001; Boland *et al.*, 1998; Chen and Murphy, 2004; Conrad *et al.*, 2004; Huang and Murphy, 2004b; Murphy *et al.*, 2000; Steckling *et al.*, 2004), although a successful application of the first approach has been reported (Danckaert *et al.*, 2002). The reason for this preference is obvious: cells vary greatly in their size, shape, intensity, position, and orientation in fluorescent images, and as a result, raw pixel intensity values in general are not very useful in location pattern recognition. Consequently, the feature-based approach is the focus of this review.

The core of the feature-based approaches is the development of sets of numerical features to represent patterns that are not overly sensitive to changes in intensity, rotation, and position of a cell (Boland and Murphy, 2001; Murphy *et al.*, 2000). Several categories of features have been used for this purpose, including morphological, Zernike moment, Harlick texture, and wavelet features. A standard nomenclature for referring to specific features and sets of features has been introduced, in which the prefix SLF (for subcellular location feature) is followed by either the number of a set of features (e.g., SLF1) or a number and a subindex for a specific feature (e.g., SLF7.3). Tables I and II summarize all SLFs developed so far for two-dimensional (2D) and three-dimensional (3D) images, respectively.

Experience gained from the work described below indicates that calculation of a large number of features (dozens, even hundreds) from a number of different categories can benefit a classifier, since they may contain different types of information. However, this potential advantage comes at a cost: the automated classifier becomes more complicated when more input features are included. Computational learning theory (Mitchell, 1997) points out that the number of training samples required is linear with the complexity. Therefore, a large set of input features can be used only when sufficient training images are available. Unfortunately, even with the newly developed automated techniques, image collection can still be a slow process and often only limited training samples are available. In this situation, more features do not guarantee better performance. Therefore, reducing the size of the feature set by eliminating redundant or uninformative features is desirable in many machine learning applications. In addition to helping with classification problems, using fewer features is preferred for applications involving retrieval of images from databases using image content.

Different feature reduction methods have been tried for this problem, including both feature recombination and feature selection methods. Feature selection methods were observed to achieve better performance than feature recombination methods in the context of subcellular pattern analysis (Huang *et al.*, 2003). Among feature selection methods, stepwise discriminant analysis (SDA) and genetic algorithms have been observed to achieve the best overall classification accuracy, and SDA is preferred since it is much more

TABLE I
Subcellular Location Feature Sets for 2D Images

Feature set name	Number of features	Feature selection	Feature categories[a]									Reference
			M	E	C	S	D	Z	H	W	G	
SLF1	16	Unselected	8	6	2							Boland and Murphy (2001)
SLF2	22	Unselected	8	6	2		6					Boland and Murphy (2001)
SLF3	78	Unselected	8	6	2			49	13			Boland and Murphy (2001)
SLF4	84	Unselected	8	6	2		6	49	13			Boland and Murphy (2001)
SLF5	37	Selected from SLF4	8	2	1		3	11	12			Boland and Murphy (2001)
SLF6	65	Unselected	8	6	2			49				
SLF7	84	Unselected	9	6	2	5		49	13			Murphy *et al.* (2002)
SLF8	32	Selected from SLF7	8	3		3		7	11			Murphy *et al.* (2002)
SLF12	8	The first eight features from SLF8	2	1				2	3			Huang *et al.* (2003)
SLF13	31	Selected from SLF7 and six DNA features	6	3		3	3	7	9			Murphy *et al.* (2003)
SLF15	44	Selected	6	4	2	2		12	7	2	12	Huang and Murphy (2004b)
SLF16	47	Selected	5	4	2	2	3	12	7	4	11	Huang and Murphy (2004b)

[a]The feature categories are M, morphological features; E, edge features; C, convex hull features; S, object skeleton features; D, DNA features; Z, Zernike moment features; H, Haralick texture features; W, Daubechies D4 wavelet features; G, Gabor features.

TABLE II
Subcellular Location Feature Sets for 3D Images

Feature set name	Parallel DNA requirement	Number of features	Short description	Reference
3D-SLF9	Yes	28	Unselected morphological features	Velliste and Murphy (2002)
3D-SLF10	Yes	9	SDA selected features from 3D-SLF9	N/A
3D-SLF11	No	42	Unselected morphological, edge and Haralick texture features	Chen *et al.* (2003)
3D-SLF14	No	14	Subset of 3D-SLF9 which does not require DNA image	N/A
3D-SLF17	No	7	The first seven SDA selected features from 3D-SLF11 on 3D HeLa set[a]	Chen and Murphy (2004)
3D-SLF18	No	34	The first 34 SDA selected features from 3D-SLF11 on 3D 3T3 set[b]	Chen and Murphy (2005)
3D-SLF19	Yes	56	3D-SLF11 and 14 DNA features from 3D- SLF9	Nair *et al.* (2005)
3D-SLF20	Yes	52	SDA selected features from 3D- SLF19	Nair *et al.* (2005)

[a]The 3D Haralick texture features were calculated at 0.4 μm pixel resolution and 256 gray levels. This feature set achieved 98% overall classification on the 3D HeLa dataset.

[b]The 3D Haralick texture features were calculated at 0.5 μm pixel resolution and 64 gray levels. This feature set is used for clustering the 3D 3T3 dataset.

computationally efficient (i.e., it requires less computer time to find a good subset of features).

1. Classification Based on Cell Level Features

Studies of protein subcellular location patterns are most easily understood in the context of a single cell. Consequently, automated classification of protein location patterns was initially carried out on single cell images.

However, many acquired images contain multiple cells per field. To properly calculate features at the level of each cell, regions of each image corresponding to a single cell need to be defined. This *cell segmentation* can be done either manually (by drawing a polygon for each cell) (Boland and Murphy, 2001) or automatically (i.e., with balloon or watershed algorithms) (De Solorzano *et al.*, 2001; Nair *et al.*, 2005; Velliste and Murphy, 2002). The automated methods usually utilize either total protein staining (so that each cell appears as a contiguous region) or surface protein staining (so that the cell boundary is visible).

a. Classification of Major Subcellular Location Patterns in HeLa Cells Using 2D Images

To test the feasibility of using automated classification to determine protein subcellular location patterns, a 2D image dataset that covers all major location patterns was first generated in HeLa cells (Boland and Murphy, 2001). Nine proteins were fluorescently labeled, eight using antibodies against proteins in the endoplasmic reticulum (ER), Golgi (giantin and gpp130), lysosomes (LAMP2), endosomes (transferrin receptor), mitochondria, nucleoli (nucleolin), and microtubules (tubulin), and one using phalloidin that binds to microfilaments (actin). DNA was simultaneously labeled with a fluorescent probe (DAPI) that could be separately detected. Two Golgi proteins (giantin and gpp130) were intentionally included to test the distinguishing power of automated classifiers. These two proteins are almost indistinguishable by human visual inspection (Murphy *et al.*, 2003). The dataset contains 78–98 single cell images per class and 862 images in total. Figure 2 shows representative images for this dataset.

i. Optimizing Classification Accuracy A steady improvement in the accuracy of classification of this dataset has been achieved over the past few years. With the implementation of new features and the choice of more robust classifiers, the overall classification accuracy improved from 84% with a set of 37 features and a neural network classifier (Boland and Murphy, 2001) to the current best accuracy of 92.3% with 47 features and a majority-voting ensemble classifier (Huang and Murphy, 2004b). The performance of the current best classifier is presented in Table III in the form of a confusion matrix. The diagonal numbers of the confusion matrix show the accuracies for each individual class and the off-diagonal numbers represent the

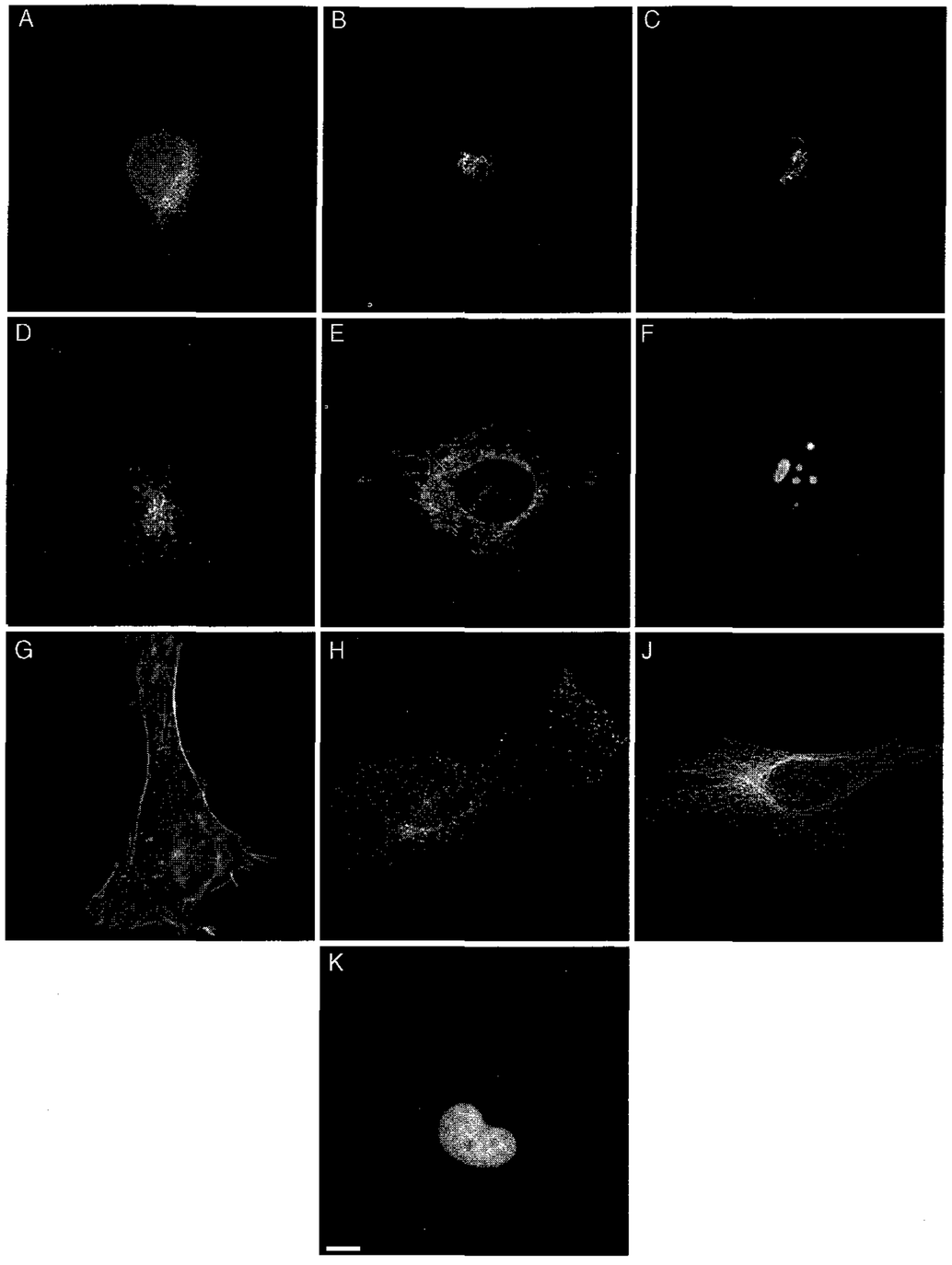

FIG. 2 Typical images from the 2D HeLa image dataset. Images are shown for cells labeled with antibodies against an ER protein (A), the Golgi protein giantin (B), the Golgi protein gpp130 (C), the lysosomal protein LAMP2 (D), a mitochondrial protein (E), the nucleolar protein nucleolin (F), the endosomal protein transferrin receptor (H), and the cytoskeletal protein tubulin (J). Filamentous actin was labeled with rhodamine-phalloidin (G) and DNA was labeled with DAPI (K). Scale bar = 10 μ. (From Boland and Murphy, 2001.)

TABLE III
Confusion Matrix for a Majority Voting Ensemble Classifier with SLF16 for the 2D HeLa Dataset[a]

	Output of the classifier									
True class	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	99	1	0	0	0	0	0	0	0	0
ER	0	97	0	0	0	2	0	0	0	1
Giantin	0	0	91	7	0	0	0	0	2	0
GPP130	0	0	14	82	0	0	2	0	1	0
LAMP2	0	0	1	0	88	1	0	0	10	0
Mitochondria	0	3	0	0	0	92	0	0	3	3
Nucleolin	0	0	0	0	0	0	99	0	1	0
Actin	0	0	0	0	0	0	0	100	0	0
TfR	0	1	0	0	12	2	0	1	81	2
Tubulin	1	2	0	0	0	1	0	0	1	95

[a]The values shown are the percentage of images from the class shown in the row heading that were classified as being in the class shown by the column heading. The overall accuracy is 92.3%. From Huang and Murphy (2004).

percentage of test samples of each "True" class (row heading) misclassified as the "Predicted" class (column heading). We can interpret the classifier's performance on each individual class from the confusion matrix. For example, the classifier shown in Table III correctly classified 88% of LAMP2 classes and mistakenly classified 1% of LAMP2 as giantin, 1% as mitochondria, and 10% as tubulin. Due to rounding errors, the sum for a row is not necessarily 100%.

ii. Tradeoff between Classification Accuracy and Computation Cost of the Feature Set The features (SLF16) used to achieve the best accuracy of 92.3% were SDA selected from a mixture of 180 features of various types (including wavelet features). Since wavelet feature calculation involves rotating each image to a common frame of reference and then decomposing it with different scale filters, the computation cost (CPU time per image) is significantly higher than for the other features (Huang and Murphy, 2004b). When the computation cost is a concern, a feature set (SLF13) with 31 features selected by SDA from all features except the wavelet features achieved a 90.7% classification accuracy with one-sixth of the average CPU time for feature calculation (Huang and Murphy, 2004b).

iii. Classification without Parallel DNA Images A parallel DNA channel provides information on nuclear location, which serves as a reference point for the cell center. Several features were designed to capture protein

distribution information relative to the nucleus and these features were included in the input feature sets above that give the best overall classification accuracy. However, not all imaging protocols require or permit the acquisition of the parallel DNA channel. A feature set with no DNA information is preferred for those cases.

A feature subset (SLF15) with 44 features SDA selected from a mixture of features that does not require DNA information achieved 91.5% overall classification accuracy (Huang and Murphy, 2004b), which indicates that while not having DNA features degrades classification performance slightly, satisfactory performance can still be achieved without them. An overall classification accuracy of 89.7% is achieved by a similar classifier trained on SDA-selected features (SLF8 with 32 features) when computation-costly wavelet features are removed (Huang and Murphy, 2004b). These results suggest that when using numerical features, the protein subcellular location pattern can be automatically determined in a reasonable time scale from fluorescence microscope images acquired using standard protocols that do not require collection of DNA images.

iv. Classification of Sets of Images When evaluating a protein's location, a cell biologist will usually not make a decision based on an image of a single cell. Instead, a set of images is viewed and a final conclusion is based on an overall impression. The same rationale can be used for automated classifiers. By considering sets of cells as small as 10 and choosing the label that has the most images assigned to it, an overall classification accuracy of 98% could be obtained with a classifier that had an average accuracy of only 83% on single cells (Boland and Murphy, 2001).

b. Classification of Two Proteins with Visually Similar Patterns in CSO-1 Cells Using 2D Images A similar feature-based approach was use to distinguish two visually similar proteins (Huntingtin and GIT) in CSO-1 cells (Steckling *et al.*, 2004). For this task, seven features, mostly defined in the previous section, were used to train a maximum likelihood classification scheme to separate the two proteins. In this experiment, 87% of the test images were correctly classified. This independent research confirms that the location pattern of previously unknown proteins could be automatically identified by feature-based approaches.

c. Classification of Images from Automated Microscopy An alternative to genomic tagging that is suitable for automated acquisition and analysis is the creation of a cell array in which each spot in the array contains a different GFP-tagged cDNA (Ziauddin and Sabatini, 2001). This approach has been combined with automated microscopy to demonstrate the feasibility of classifying images from automated acquisition (Conrad *et al.*, 2004). In this

study, images for 11 cDNAs showing different subcellular location patterns were captured automatically. An SVM classifier achieved an 82.2% overall classification with 25 SDA-selected features drawn from a large number of features (448) extracted from different feature categories (such as object, edge, texture, moments, wavelets). Achieving this accuracy required training of the system to recognize an "artifact" category; approximately half of the images were found to be in this category. Even with this step, the observed accuracies for two of the categories (ER and microtubules) were below 50%. The results are encouraging for the application of the methods described in this chapter to images obtained by automated microscopy.

d. Classification of Major Subcellular Location Patterns in HeLa Cells Using 3D Images With the development of optical sectioning techniques (i.e., confocal microscopy), it has become increasingly common to collect 3D images when visualizing protein distributions. A 3D image is simply a stack of 2D images taken at different focal planes. Compared to their 2D counterparts, 3D fluorescence images provide an opportunity to achieve better classification accuracy since we can expect that 3D images contain more information content than 2D images. Indistinguishable location patterns in a 2D slice could be potentially separated by their distribution along the third dimension. This is obvious for polarized cells where the protein composition of the apical surface is different from that of the basal and lateral surface. In this case, the protein distribution is different among three dimensions. Even for an unpolarized cell, the protein distribution on the third dimension still provides extra information that may help in distinguishing different patterns. For example, F-actin in HeLa cells is preferentially located above the nucleus while tubulin is more concentrated near the bottom of the cell.

A 3D HeLa cell image dataset was collected using the same labeling techniques as for the 2D HeLa dataset to determine whether an improved classification accuracy could be obtained (Velliste and Murphy, 2002). This dataset contains 50–52 single cell images per class and 502 imaged cells in total. Figure 3 shows the representative images of this dataset.

Most 3D features [morphological (Velliste and Murphy, 2002), edge (Chen *et al.*, 2003), and Haralick texture features (Chen *et al.*, 2003)] were natural extensions of their 2D versions. Some new features were also implemented to capture characteristics of protein fluorescence distribution along the z axis.

An overall classification accuracy of 98% was achieved using 3D-SLF17, an SDA-selected feature subset of seven features. Table IV shows the confusion matrix for this case. The results suggest that the classifier is near optimal, since cells were imaged randomly and inevitably a small fraction of "abnormal" cells, such as mitotic or dying cells, would be expected to be included. Protein location patterns in these cells are likely different from those

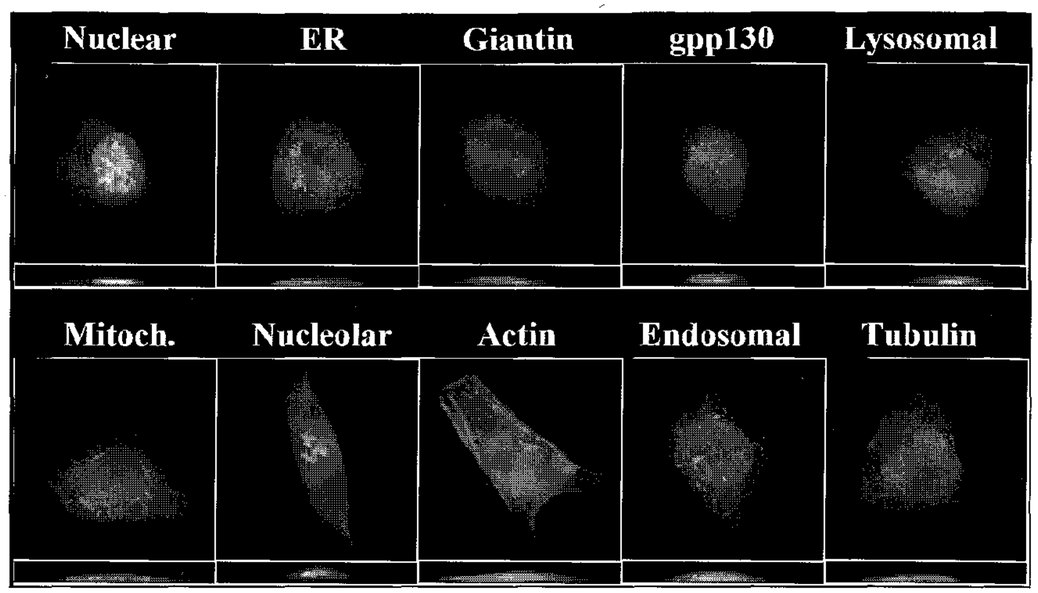

FIG. 3 Typical images from the 3D HeLa image dataset. Red, blue, and green colors represent DNA staining, total protein staining, and target protein fluorescence. Projections on the X–Y (top) and the X–Z (bottom) planes are shown. The proteins labeled are the same as those in the 2D HeLa image dataset (Fig. 2). (Reprinted by permission of Carnegie Mellon University.) (See also color insert.)

TABLE IV
Confusion Matrix for a Neural Network Classifier with 3D-SLF17 for the 3D HeLa Dataset[a]

	Output of the classifier									
True class	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

[a]The values shown are the percent of images from the class shown in the row heading that were classified as being in the class shown by the column heading. The overall accuracy is 98%. From Chen and Murphy (2004).

observed in normal cells and could account for the small classification errors. It is also worth noting that the 3D classifier is the first that is able to distinguish giantin and gpp130, two proteins known to be located in slightly different parts of the Golgi apparatus, with over 95% accuracy at the single cell level.

3D-SLF17 consists of at least one feature from each category (morphological, edge, and Haralick texture for 3D images), suggesting that different feature categories capture different information in the image and an optimal classifier should be trained from a combination of features from different categories. 3D-SLF17 can be calculated efficiently, as the time-consuming wavelet features are not included. Another important advantage of 3D-SLF17 is that it does not require a parallel DNA image. It confirms that for 3D images, even without the reference DNA information, an optimal classifier could be trained to learn major subcellular location patterns, a crucial step in extending the feature-based approach toward general usage for biological and biomedical researchers.

The already impressive performance can be improved further by classifying sets of images. Trained classifiers (using either neural networks or support vector machines) achieved above 99% classification accuracy with a set size of 3 and 99.9% with a set size of 5 (data not shown).

2. Classification Using Field Level Features

As discussed above, classifiers trained on SLFs are capable of automatically labeling protein subcellular location patterns in previously unseen images in a single cell setting. However, typically a field containing multiple cells is imaged. Partial cells on the boundary of imaged fields are also frequently observed. Since current SLFs are designed to capture protein spatial distributions in a single cell setting, their use on multiple cells (either intact or partial) would not necessarily give appropriate results. SLF1.1, the number of objects inside the cell (an object is defined as a set of connected above-threshold pixels in the image), is a good example. It is a very useful feature to distinguish the nuclear or nucleolar patterns (with one to a few objects) from the mitochondrial or lysosomal patterns (with hundreds of objects). However, this feature is meaningless if an image contains multiple cells (unless it is known that all images have the same number of cells!). As discussed above, a few automated segmentation algorithms have been described for identifying regions corresponding to a single cell in a multicell field. Unfortunately, they either require extra labeling (such as nucleus, total cellular protein, or plasma membrane protein labeling) or assume specific models (Sclaroff and Liu, 2001), which makes them difficult to be generalized to arbitrary protein fluorescence images. An alternative solution would be a set of features that is not sensitive to the number of cells in a field.

Such a feature set was created by selecting features that meet this requirement from previous SLFs. This approach was tested by building an SVM classifier to recognize images with multiple cells (Huang and Murphy, 2004a). So that the label of each cell in a multicell image could be known with certainty, synthetic multicell images were created by combining single cell images whose location patterns were known. A confusion matrix for this classifier is shown in Table V, and the overall classification accuracy of 94.8% was even higher than for the single cell counterpart. This "surprising" improvement in performance could be partially explained by so-called "majority voting" effects where the effect of abnormal cells in a class is diluted in multicell images so that the classifier makes the correct prediction based on the majority type of cells in a field.

Although experiments classifying 3D multicell images have not yet been carried out, higher accuracy is expected as well since most features in the best 3D feature set (3D-SLF17) are edge and texture features, which are insensitive to the number of cells in the field. The only morphological feature in SLF17 is 3D-SLF9.4, the standard deviation of object volumes, and since it does not depend on the number of cells, it is also a potential field level feature.

3. Classification of Mixed Patterns Using Object Level Features

Previous sections show that a classifier can be trained to distinguish a set of predefined patterns with high accuracy. However, protein localization is a complicated process, and proteins are not restricted to being in a single

TABLE V

Confusion Matrix for an SVM Classifier with Entire 2D Field Level Features for the Multicell Image Dataset[a]

	Output of the classifier									
True class	DNA	ER	Gia	GPP	LAM	Mit	Nuc	Act	TfR	Tub
DNA	100	0	0	0	0	0	0	0	0	0
ER	0	96	0	0	0	0	0	0	4	0
Giantin	0	0	100	0	0	0	0	0	0	0
GPP130	1	0	2	98	0	0	0	0	0	0
LAMP2	0	0	0	4	94	0	0	0	2	0
Mitochondria	0	4	0	2	0	96	0	0	2	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	4	0	0	2	4	4	0	82	4
Tubulin	0	4	0	0	2	4	0	0	8	82

[a]The overall accuracy was 94.8%. From Huang and Murphy (2004a).

organelle. For example, the human mannose 6-phosphate uncovering enzyme (UCE) cycles between the *trans*-Golgi network (TGN) and the plasma membrane (Rohrer and Kornfeld, 2001). In this case, the steady-state pattern of UCE is a mixture of two fundamental location patterns (TGN and plasma membrane) and the feature values for the UCE location pattern would be different from the typical values for either of those fundamental patterns. Consequently, a classifier trained to recognize the two fundamental patterns would fail to recognize this mixed pattern.

A straightforward solution would be to train a classifier to distinguish each possible combination of fundamental location patterns. However, the large number of possible combinations (exponential in the number of fundamental patterns) makes this approach prohibitive. The problem is even more complicated because different mixture ratios would be expected to yield different combined patterns. For example, a Y488A mutation in UCE slows down its traffic from the plasma membrane back to the TGN and consequently the mutant protein has a much higher plasma membrane distribution. The location patterns for wild-type and Y488A-mutant UCE are distinguishable by both human visual inspection and automated machine learning algorithms although they are both mixed from TGN and plasma membrane patterns.

A more suitable approach would be to apply a machine learning algorithm that is capable of recognizing mixtures composed of varying amounts of independent fundamental patterns. An initial approach to this problem has been presented (Zhao *et al.*, 2005). It is based on the principle that each fundamental pattern has a distinguishable (but stochastic) combination of object types. Object types were learned by cluster analysis and then classifiers trained to recognize each type. The distribution of object types within each fundamental pattern was also modeled. The decomposition of an arbitrary mixed image into its component patterns was then carried out by a two-step process. Each object in the image was assigned a type and then the mixture fractions that would most likely have led this distribution of object types were found.

For the training phase of this system, the individual objects in images of known fundamental patterns were first identified and a set of subcellular object features (SOF1, shown in Table VI) was calculated for each. These features for all objects in the dataset were then used to perform cluster analysis to determine the types of objects that exist in the dataset. A classifier was then trained to recognize each object type from its SOFs. Each fundamental pattern (whether for a training image or a test image) could then be represented by the number of objects of each type or the fraction of fluorescence in each object type (or both). This allows the object type vector for a mixture pattern to be considered as a linear combination of object type vectors of component fundamental patterns. The mixture fractions could

TABLE VI
Subcellular Object Features (SOF1) Used to Cluster Objects to Learn Object Types and to Train Classifiers to Recognize Them[a]

Feature ID	Description
SOF1.1	Number of pixels in object
SOF1.2	Distance between object center of fluorescence (COF) and DNA COF
SOF1.3	Fraction of object pixels overlapping with DNA
SOF1.4	A measure of eccentricity of the object
SOF1.5	Euler number of the object
SOF1.6	A measure of roundness of the object
SOF1.7	The length of the object's skeleton
SOF1.8	The ratio of skeleton length to the area of the convex hull of the skeleton
SOF1.9	The fraction of object pixels contained within the skeleton
SOF1.10	The fraction of object fluorescence contained within the skeleton
SOF1.11	The ratio of the number of branch points in the skeleton to the length of the skeleton

[a]From Zhao *et al.* (2005).

then be found by solving a set of linear equations or maximizing the likelihood of a multinomial model.

Using this approach on test images synthesized by creating random mixtures of up to eight subcellular patterns, an average of 83% accuracy in determining the mixture proportions was achieved (Zhao *et al.*, 2005). While further work on this important problem will clearly be required, these initial results suggest that decomposition of complex mixture patterns is feasible.

B. Objective Clustering

It has been shown in the previous section that a classifier trained on SLFs could assign a set of predefined location pattern labels to previously unseen images with high accuracy. However, it remains an open question as to how many location patterns exist in a specific proteome. Ideally, we would like to collect images of all proteins and learn the number of statistically distinguishable location patterns and the identity of each pattern in a proteome. This becomes a typical unsupervised learning problem, addressable by methods of cluster analysis.

Central to unsupervised clustering is the notion of the degree of similarity (or dissimilarity) between individual data points (cell images). A major

advantage of using SLFs to represent images is that the distance between two data points can be easily defined. This measures the dissimilarity between two points, e.g., a larger distance means two points are less similar to each other. Common distance functions include Euclidean distance [2-norm (Qian *et al.*, 2004)], Manhattan distance [1-norm (Batchelor, 1978)], Chebyshev distance [infinity-norm (Diday, 1974)], Mahanlanobis distance (Nadler and Smith, 1993), cosine angle distance (Qian *et al.*, 2004), and Hamming distance [suitable for binary features (Exoo, 2003)].

1. Dendrogram Analysis of the 2D HeLa Dataset

As an initial demonstration of the cluster analysis approach to subcellular patterns, hierarchical clustering was performed on the 10 class 2D HeLa dataset (Murphy *et al.*, 2002). Although many techniques that could improve the quality of the clustering were not used in this simple experiment, the resulting dendrogram (Fig. 4) still revealed that (1) similar location patterns were grouped first (the two Golgi proteins, and the lysosomal and endosomal proteins), and (2) it was consistent with the biological understanding of these organelle patterns. This first result suggested the utility of clustering approaches using SLFs to reveal the intrinsic relationships among subcellular patterns displayed in a dataset.

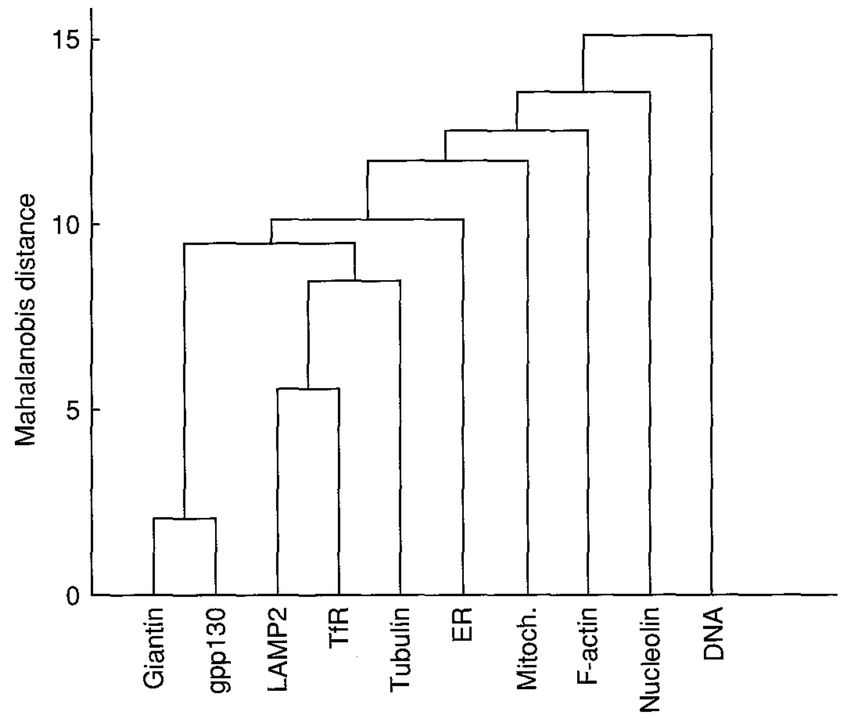

FIG. 4 Subcellular location tree created from the 2D HeLa dataset using SLF8 and hierarchical clustering with a Mahalanobis distance function. (From Murphy *et al.*, 2002. Copyright © 2002, IEEE.)

2. Clustering of the 3D 3T3 Dataset

One of the ultimate goals for location proteomics is to determine which proteins share the same location pattern. In other words, given a set of proteins, each with multiple image representations, we need to find a partitioning so that proteins in one partition share a single location pattern. Ideally, this approach would be applied to images of all proteins expressed in a given cell type. Methods are therefore needed to collect images of the distributions of large numbers of (or all) proteins.

As discussed in Section II.B , this can be done by creating GFP fusions for many different cDNAs and expressing them by carrying out individual transfections. The advantage is that the tagged gene is known in each transfection, but the disadvantage is that the protein is expressed under the control of an exogenous promoter that may lead to overexpression and mislocalization.

An alternative is to create random GFP fusions, make clonal lines, and then determine the tagged gene in each line. Although more work is required to identify the tagged gene, the advantage is that the protein is expressed in its normal genomic context with all transcriptional and posttranscriptional controls intact. One variation on this approach, CD-tagging, is particularly powerful because tags may be inserted at a number of sites on each protein (Jarvik *et al.*, 1996). The CD-tagging method employs an engineered retroviral vector to randomly insert into the target genome a CD-cassette, flanked by splicing acceptor and donor sequences. If the viral vector is inserted into a genomic intron, the CD-cassette will be transcribed as a guest exon. This approach was used in the laboratories of Drs. Jonathan Jarvik and Peter Berget to create a collection of 3T3 cell clones each of which expresses a different tagged protein (Jarvik *et al.*, 2002). The CD-cassette used for this collection contains a GFP coding sequence, thus permitting the tagged proteins to be visualized in live cells. To date, over 90 randomly tagged clones expressing GFP chimera proteins have been isolated. To acquire live cells images with a minimum of blur due to organelle movement, spinning disk confocal microscopy (Kozubek *et al.*, 2004; Nakano, 2002) was used to collect a large number of images of each clone (Chen *et al.*, 2003). The 3D 3T3 dataset obtained to date contains 8–33 single cell images per clone and 1554 images in total. Figure 5 shows representative images of example clones.

Three machine learning algorithms have been applied to the dataset, including k-means clustering of individual cells, hierarchical clustering of mean feature vectors for each clone, and an *ad hoc* clustering algorithm based on a confusion matrix from a classifier trying to identify each individual clone (Chen and Murphy, 2005). Both standardized (z-scored) Euclidean (where each feature is normalized to have unit variance) and Mahalanobis (where the correlation between features is taken into consideration) distance

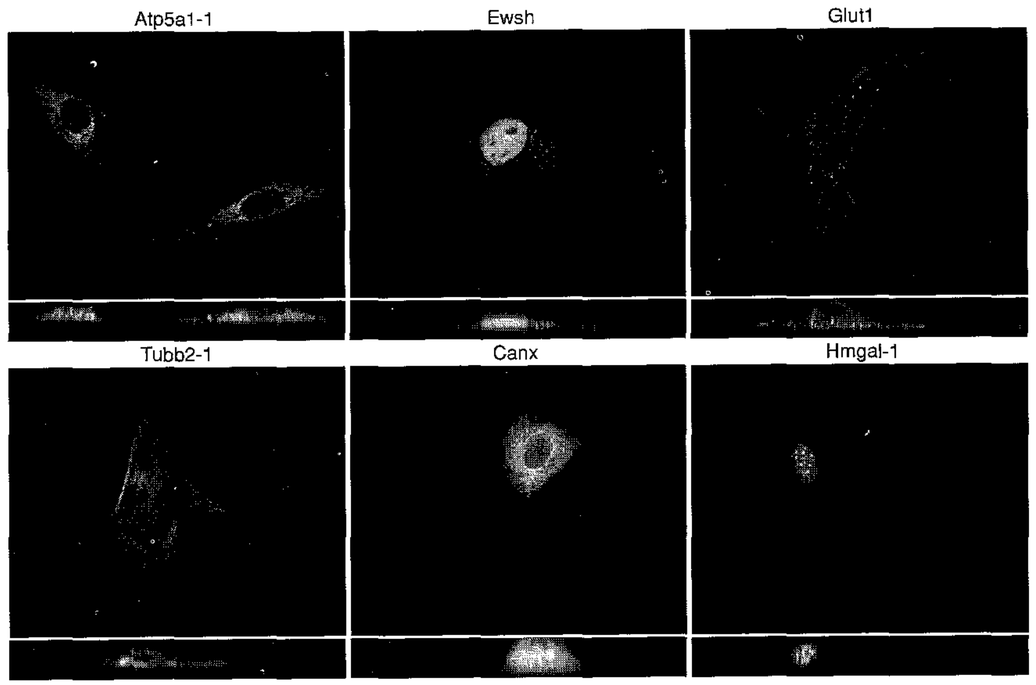

FIG. 5 Selected images from the 3D 3T3 image dataset. Tagged protein names are shown with a hyphen followed by a clone number if the same protein was tagged in more than one clone in the dataset. Projections on the X–Y (top) and the X–Z (bottom) planes are shown. (From Chen and Murphy, 2005.)

functions were evaluated. The performance of machine clustering algorithms was also compared to a grouping based on visual inspection.

A traditional difficulty in machine clustering approaches is to find the number of clusters in a dataset. For example, k-means approaches require the number of clusters to be directly specified in the program. For hierarchical clustering, it is also critical to know which branches represent identical classes. This can be achieved by finding a distance under which the separation is not statistically significant. The distance threshold identified also indirectly determines the number of clusters since data points connected at the threshold belong to the same cluster. In some specific problems, outside information can be used to determine the number of clusters. When this information is not available, a set of trials with different numbers of clusters can be performed and evaluated by some criterion that measures the goodness of the clustering results. There is no prior information suggesting the proper number of clusters for the 3D 3T3 dataset, although a reasonable assumption is that the cells from the same protein clone should belong to a single cluster. Therefore, the goodness of various clustering models was assessed by the Akaike information criterion (AIC) (Ichimura, 1997) for different numbers

of clusters (from two to the number of clones in the dataset) (Chen and Murphy, 2005).

Evaluation of the clustering results using different clustering algorithms and/or different distance functions is an equally difficult task since the actual partition of the dataset is unknown for any real problem setting. Nevertheless, it is reasonable to assume that a good distance function would achieve better agreement among different clustering algorithms. Cohen's κ statistic (Cook, 1998) can be used to measure the agreement between two partitionings. This statistic measures the portion of agreement beyond random and it has a value of 1 if two partitionings are in total agreement and an expected value of 0 if the two partitionings are mutually independent.

Table VII summarizes the comparison of the κ statistic using different clustering algorithms and distance functions. It clearly indicates that the standardized Euclidean distance function achieved better agreement compared to Mahalanobis distance. It also revealed that machine learning clustering algorithms (k-means/AIC, consensus analysis and clustering based on confusion matrix) achieved much higher agreement compared to pairs between a machine algorithm and clustering by visual inspection. The consistency displayed further suggested the value of automated clustering approaches in location proteomics.

When standardized Euclidean distance was used in the k-means/AIC algorithm to cluster the 3D 3T3 dataset, the optimal number of clusters was 30. However, 13 of these clusters contained only outlier images (outliers of a clone were those distributed to any cluster except the one with the most images for that clone). Therefore, these clusters were ignored, leaving 17

TABLE VII
Comparison of Clustering Methods and Distance Functions[a,b]

	Standardized Euclidean distance	Mahalanobis distance
	(κ)	(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

[a]The agreement between the sets of clusters resulting from the four clustering methods described in the text was measured using the κ test. A value of 1 represents perfect agreement between the clusters formed by the two methods. From Chen and Murphy (2005).

[b]The standard deviations of the statistic under the null hypothesis were estimated to range between 0.014 and 0.023 from multiple simulations.

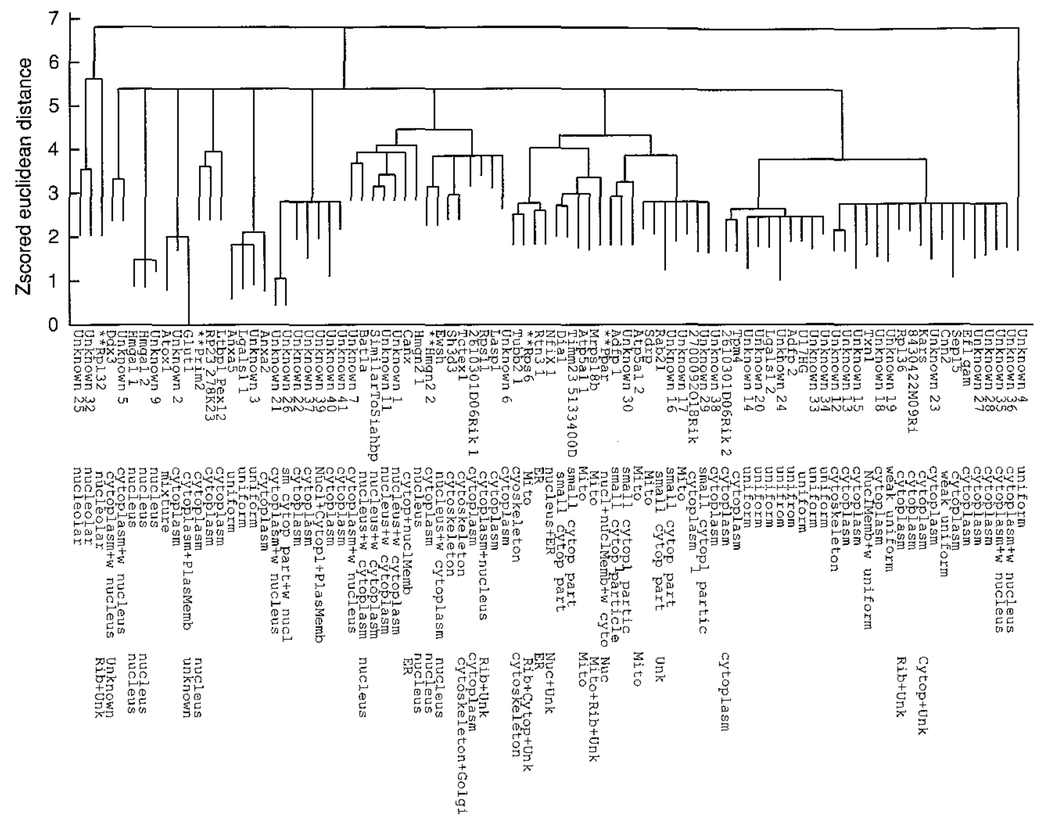

FIG. 6 A consensus subcellular location tree generated for the 3D 3T3 image dataset using 3D-SLF18 features. The columns to the right of the tree show the protein names (if known), descriptions of the subcellular location from visual inspections of the images, and subcellular location inferred from Gene Ontology (GO) annotations (if any). The sum of the lengths of horizontal edges connecting two proteins represents the distance between them in the feature space. Proteins for which the location described by human observation differs significantly from that inferred from GO annotations are marked (**). (From Chen and Murphy, 2005.)

clusters containing the major patterns in the dataset. The consensus tree from hierarchical clustering is displayed in Fig. 6 (this tree and the representative images for each protein clone are available in an interactive web interface: http://murphylab.web.cmu.edu/services/PSLID/tree.html). The same 17 clusters were identified in both the consensus clustering and AIC analysis.

A natural question is whether the clusters obtained from unsupervised learning could be used for supervised training of a classifier to recognize this new set of patterns. When this test was performed, an average accuracy of 87% was achieved. This classifier is able to recognize the largest number of subcellular patterns reported to date.

As at least a partial confirmation of the validity of the results obtained by clustering, we can observe that the tree shows visually similar location

patterns being grouped together. For example, the three nucleolar proteins in the dataset were combined in a distinct cluster. Also, two groups of nuclear proteins were found in the dataset. Close inspection of the images from both groups confirmed that there are differences between them, one with exclusive nucleus distribution and the other with weak cytoplasmic distribution as well (Chen *et al.*, 2003).

3. Cluster Analysis of Location Patterns in UCE Mutants

The clustering results above had the goal of clustering *proteins* with similar patterns. An analogous goal would be to cluster *mutant* versions of a single protein to determine which share the same location. Mutagenesis of suspected targeting sequences is often used to identify important residues in that sequence. To illustrate the value of the clustering approach in this context, we applied it to images of a set of mutants obtained by the laboratory of Dr. Jack Rohrer.

The trafficking of UCE, a key enzyme in lysosome biogenesis, has been discussed briefly above. UCE uncovers the mannose 6-phosphate recognition tag on lysosomal enzymes, a step necessary for mannose 6-phosphate receptors to recognize these enzymes and ensure their sorting to lysosomes. UCE is localized to the TGN in steady state and cycles between the TGN and plasma membrane. It requires targeting information both for exit from the TGN and for return from the plasma membrane. It has been determined that the Y^{488} residue is required for traffic from the plasma membrane back to the TGN since a Y^{488}-A mutant has impaired internalization from the plasma membrane. To locate the sequence that mediates exit from the TGN, the subcellular distributions of a set of mutants created by site-directed mutagenesis of a GFP-tagged UCE were studied by imaging and cluster analysis (Nair *et al.*, 2005).

In an initial experiment, adjacent pairs of residues suspected of being part of the targeting signal were mutagenized. When images of these mutants were clustered, the dendrogram (Fig. 7A, the dendrogram and representative images from each clone are available online at http://murphylab.web.cmu.edu/services/PSLID/HeLa_UCE/Figure4A.html) suggested that mutating the QE or MN residues of the cytoplasmic tail created an intermediate location pattern between wild-type UCE (which exhibited mostly TGN staining) and the Y^{488} mutant (which had a strong plasma membrane distribution). It was therefore postulated that one or more of the Q^{492}EMN residues could be responsible for the traffic from the TGN to the plasma membrane. Mutants with single residue mutation were consequently generated and studied together with the double mutants. The clusters formed by the single and double mutants are shown in Fig. 7B (available online at http://murphylab.web.cmu.edu/services/PSLID/HeLa_UCE/Figure4B.html). The

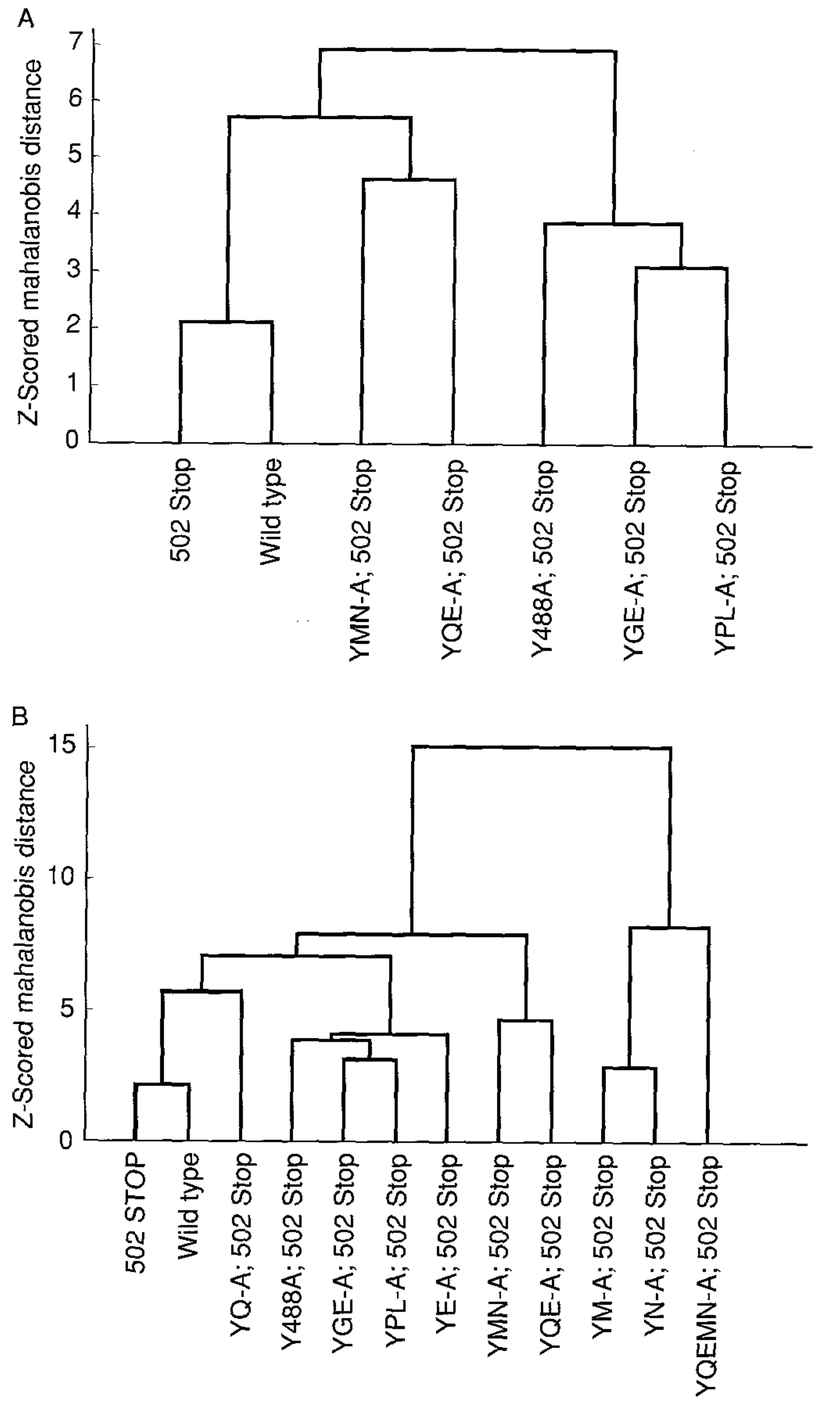

FIG. 7 Subcellular location tree using 3D-SLF20 and Mahalanobis distance showing grouping of various GFP constructs of the uncovering enzyme (UCE) by similarity in localization. (A) Results for the GFP-UCE wild-type, GFP-UCE 502 Stop (which truncates the cytoplasmic tail but does not affect targeting), GFP-UCE Y^{488}-A/502 Stop (which is not internalized from the plasma membrane), and mutant constructs with pairs of the Q^{492}EMNGEPL residues in the cytoplasmic tail of UCE converted to alanine. (B) Results for those mutants plus a mutant with all four of the Q^{492}EMN residues converted to alanine plus single mutants in those four residues. (From Nair *et al.*, 2005.)

largest cluster contained the Y^{488} mutant and three mutants that do not appear to affect TGN exit significantly. A second cluster contained wild-type UCE, a mutant with a truncated cytoplasmic tail that does not affect tracking, and a mutant in which both the Y^{488} and Q^{492} residues were replaced by alanines. The important conclusion is that the first mutation stops return to the TGN but the second mutation compensates for it by preventing exit from the TGN. However, additional conclusions not obvious on visual inspection could also be drawn from the remaining clusters. These are that mutations in the M, N, and E residues, while not blocking TGN exit, conferred a distinct phenotype on the enzyme. The results suggest that yet another step in the traffic process involves those residues.

This experiment indicated that coupled with prior biological knowledge and proper experimental design, an automated interpretation approach could yield new biological knowledge.

4. Clustering Drugs by Their Affects on Location Patterns

The work described above involved clustering of proteins (or mutant proteins) by their location patterns reflected in fluorescence microscope images. Clustering using features derived from analysis of fluorescence microscope images has also been described to group drugs by their mechanism of action on cultured cells (Perlman *et al.*, 2004). In this study, automated microscopy was used to collect 10 images for cells stained in parallel for 10 different marker proteins (plus DNA) upon treatment with one of 100 different drugs. The DNA fluorescence image was used to find nuclear regions, and then a 14-pixel-wide cytoplasmic annulus surrounding each nucleus was created. These regions were used to calculate a total of 93 descriptors from all of the marker images for a given drug. The descriptors mainly measured nuclear size, shape, and intensity, and the average intensities of the 10 marker proteins in the nuclear region and in the cytoplasmic annulus. For each drug, each descriptor for each marker was converted into a score reflecting how much the distribution of that descriptor changed from the control case. The drugs were then clustered using this vector of scores. Drugs with known mechanisms of action were observed to be clustered, validating the approach for determining mechanisms of action for unknown drugs.

C. Other Statistical Analyses

Results from classification and clustering analysis clearly show that SLFs capture the essential characteristics of protein fluorescence images. This suggests that they can also be used as a foundation for other kinds of

statistical analysis. Two approaches that address commonly encountered problems in cell biology studies are described below.

1. Objective Selection of Typical Images from an Image Set

Due to space limitations, only a small portion of the images in a dataset can be used for communicating results in papers or presentations. Selecting typical images from an image set is therefore a routine task for many researchers, especially for cell biologists. Similar to location pattern determination, visual inspection is the traditional and most commonly used approach for this task. Consequently, criteria for the selection are largely subjective and this approach suffers from inconsistency (i.e., for the same image set, two independent researchers are unlikely to choose the same image or set of images as representative).

The shortcomings of selection by visual inspection can be avoided by developing an automated system using SLFs that make objective selections of representative images based on statistical models. As in clustering, distance (or dissimilarity) is the core concept here. When each image is reduced to a feature vector, the distances from each vector to the mean feature vector of the image set can be used to reflect the dissimilarity between the corresponding image and the set as a whole. Therefore the inverse of this distance can be used to rank images in order of their typicality (Markey *et al.*, 1999). This approach was experimentally validated using 2D images of subcellular patterns in Chinese hamster ovary cells. Figure 8 shows the

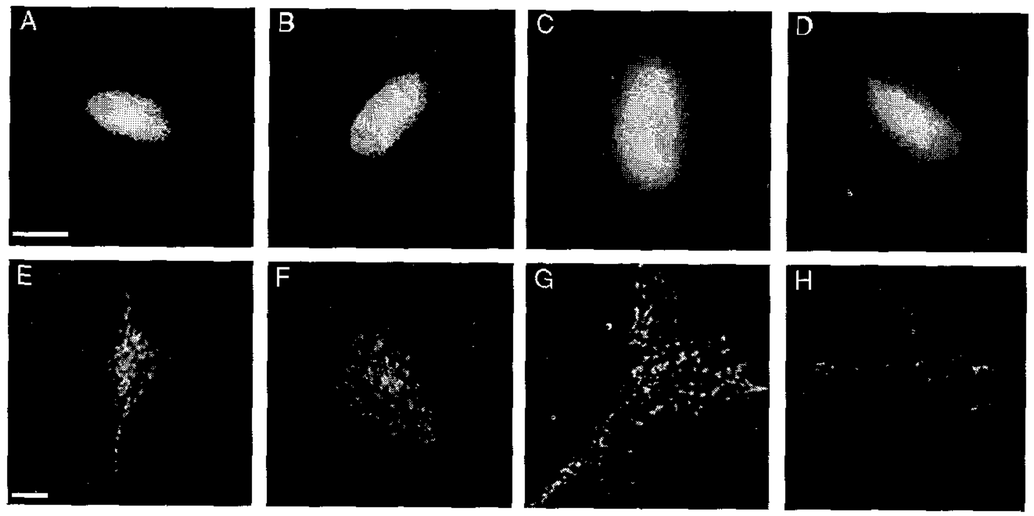

FIG. 8 The most (A, B, E, and F) and least (C, D, G, and H) typical DNA (A–D) and LAMP2 (E–H) images selected by the TypIC program from the 2D CHO dataset. Scale bar = 10 μm. (After Markey *et al.*, 1999.)

most and least representative images selected for DNA (A–D) and lysosomal protein (E–H) patterns. The most typical images (A and B) selected for the DNA pattern show a clear boundary between nucleus and cytoplasm while the least typical images (C and D) exhibit a weak but distinguishable punctuate pattern extending from the nucleus, which suggests either a poor fixation or perhaps that the nuclear membrane had been compromised before fixation. For the lysosomal protein pattern, the most typical images (E and F) showed some lysosomes concentrated in the perinuclear region and some distributed peripherally. In contrast, the least typical images showed a much more peripheral distribution (G), or an apparent reduction in the number of lysosomes (H).

A program for implementing this algorithm, TypIC (for *Typ*ical *I*mage *C*hooser), was originally implemented as a web service and is now included in the PSLID database service described below.

2. Objective Comparison of Two Image Sets

Another frequent question asked by biologists using fluorescence microscopy is whether two image sets represent different location patterns. For instance, there could be interest in whether expression of a mutated or engineered gene changes the location pattern of the targeted protein. Pharmaceutical researchers could be interested in determining whether a certain drug treatment significantly changes the location pattern of a protein of interest. As is the case for the task of selecting typical images, the conventional approach to comparing sets of patterns is visual inspection. The same limitation applies: two different researchers could possibly reach the opposite conclusion for the same two sets.

This problem can be avoided by using the automated and objective approach of statistical hypothesis tests, such as the Hotelling T^2 test, to compare two feature matrices (Roques and Murphy, 2002). When the Hotelling T^2 test (a multivariate version of the Student's t test) is employed, two image sets would be concluded to be different at a given confidence level if the resulting F value is greater than a critical F value calculated for that confidence level. By this method, all pairs of classes in the 2D HeLa dataset were shown to be statistically different at the 95% confidence level, which is consistent with the finding that a trained classifier could distinguish all 10 classes with relatively high accuracy. On the other hand, when two sets of images of giantin labeled with different antibodies were compared, the sets were found to be indistinguishable at that confidence level. This supports the utility of this approach for detecting meaningful differences without being overly sensitive to experimental variations.

A program for implementing this algorithm, SImEC (for *S*tatistical *Im*aging *E*xperiment *C*omparator) is also available in the PSLID system described below.

D. Protein Subcellular Location Image Database

As for other biological entities, there is a need for large scale online databases to organize biological images from different sources as well as to exchange and manage these images. A number of approaches to creating such databases have been described (Andrews *et al.*, 2002; Gonzalez-Couto *et al.*, 2001; Huang *et al.*, 2002). A critical requirement for such databases is integration of well-designed numerical features for describing each image into the database. The SLFs described above provide excellent discrimination between subcellular location patterns to create a database to capture large numbers of fluorescence microscope images depicting protein location (Huang *et al.*, 2002). This Protein Subcellular Location Image Database (PSLID) uses a publicly available database schema (Fluorescence Microscope Annotation Schema, http://murphylab.web.cmu.edu/services/FMAS), and permits query via text annotation of sample preparation and image collection.

PSLID is built using public domain database and web server software (Postgres and Apache). It contains a Java Server Page (JSP) web interface that permits the different tasks described in this chapter (i.e., classification, clustering, SImEC, and TypIC) as well as text and feature-based query to be carried out on large sets of images. The current PSLID database (containing the 2D and 3D images for HeLa and 3T3 cells described above) can be accessed at http://murphylab.web.cmu.edu/services/PSLID, and the open source software can also be downloaded to create additional local databases. One goal of the current work is focused on providing tools for federating such local databases to create a global but distributed database (Singh *et al.*, 2004).

IV. Concluding Remarks

Current advances in automated interpretation of protein subcellular location distributions were briefly discussed in this review. The studies summarized here have shown that protein subcellular location patterns can be interpreted automatically and objectively by feature-based approaches, and usually outperform visual inspection.

The core of the automated interpretation approaches is the development of features capturing essential characteristics of protein subcellular distributions while not being overly sensitive to experimental variations, such as the cell location, orientation, and absolute intensity. Even for unpolarized cells, 3D images have higher pattern information content than 2D images. Consequently, classifiers trained on 3D images achieved better performance.

Although the current methods are still far from perfect, they can be expected to form the foundation of future research in location proteomics.

For example, better performance could be achieved by continuously improving feature design (more complete coverage of information contained), better feature implementation (increased computational efficiency), and developing algorithms for using images with higher dimension (e.g., time series images).

Combined with advances in random tagging and high-throughput imaging techniques, automated interpretation tools can generate a complete view of the location patterns for most, if not all, proteins expressed in an arbitrary cell type. We are only at the threshold of discovering what new insights location proteomics can contribute to biological and biomedical research.

Acknowledgments

The work from our research group could not have been accomplished without the help of Drs. David Casasent, Jonathan Jarvik, Peter Berget, Simon Watkins, and our colleagues in the Center for Automated Learning and Discovery and the Center for Bioimage Informatics. That work was supported in part by NIH Grant R01 GM068845, NSF Grant EF-0331657, and Research Grant 017393 from the Commonwealth of Pennsylvania Tobacco Settlement Fund. X. Chen was supported by a Graduate Fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University founded by the Merck Company Foundation.

References

Ajenjo, N., Canon, E., Sanchez-Perez, I., Matallanas, D., Leon, J., Perona, R., and Crespo, P. (2004). Subcellular localization determines the protective effects of activated ERK2 against distinct apoptogenic stimuli in myeloid leukemia cells. *J. Biol. Chem.* **279,** 32813–32823.

Andrews, P., Harper, I., and Swedlow, J. (2002). To 5D and beyond: Quantitative fluorescence microscopy in the postgenomic era. *Traffic* **3,** 29–36.

Batchelor, B. G. (1978). "Pattern Recognition: Ideas in Practice." Plenum Press, New York. 71–72.

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. (1998). Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* **33,** 366–375.

Brelie, T. C., Wessendorf, M. W., and Sorenson, R. L. (2002). Multicolor laser scanning confocal immunofluorescence microscopy: Practical application and limitations. *Methods Cell Biol.* **70,** 165–244.

Chen, X., and Murphy, R. F. (2004). Robust classification of subcellular location patterns in high resolution 3D fluorescence microscope images. Proc. 26th Annual Intl. Conf. IEEE Eng. Med. Biol. Soc., pp. 1632–1635.

Chen, X., and Murphy, R. F. (2005). Objective clustering of proteins based on subcellular location patterns. *J. Biomed. Biotechnol.* **2005,** 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.

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

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

Danckaert, A., Gonzalez-Couto, E., Bollondi, L., Thompson, N., and Hayes, B. (2002). Automated recognition of intracellular organelles in confocal microscope images. *Traffic* **3,** 66–73.

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.

Diday, E. (1974). Recent progress in distance and similarity measures in pattern recognition. Proc. 2nd. Intl. Joint Conf. Pattern Recog., pp. 534–539.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016.

Exoo, G. (2003). A Euclidean Ramsey problem. *Discrete Comput. Geom.* **29,** 223–227.

Fujiwara, K., and Pollard, T. D. (1976). Fluorescent antibody localization of myosin in the cytoplasm, cleavage furrow, and mitotic spindle of human cells. *J. Cell Biol.* **71,** 848–875.

Gonzalez-Couto, E., Hayes, B., and Danckaert, A. (2001). The life sciences Global Image Database (GID). *Nucleic Acids Res.* **29,** 336–339.

Griffin, B. A., Adams, S. R., and Tsien, R. Y. (1998). Specific covalent labeling of recombinant protein molecules inside live cells. *Science* **281,** 269–272.

Habeler, G., Natter, K., Thallinger, G. G., Crawford, M. E., Kohlwein, S. D., and Trajanoski, Z. (2002). YPL.db: The Yeast Protein Localization database. *Nucleic Acids Res.* **30,** 80–83.

Hardonk, M. J., Dijkhuis, F. W., Haarsma, T. J., Koudstaal, J., and Huijbers, W. A. (1977). Application of enzyme histochemical methods to isolated subcellular fractions and to sucrose-Ficoll density gradients. A contribution to the comparison of histochemical and biochemical data. *Histochemistry* **53,** 165–181.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., De la Cruz, N., Tonatello, P., Jaiswal, P., Seigfried, T., and White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32,** D258–D261.

Hernandez-Munoz, I., Benet, M., Calero, M., Jimenez, M., Diaz, R., and Pellicer, A. (2003). rgr oncogene: Activation by elimination of translational controls and mislocalization. *Cancer Res.* **63,** 4188–4195.

Hua, S., and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17,** 721–728.

Huang, K., and Murphy, R. F. (2004a). Automated classification of subcellular patterns in multicell images without segmentation into single cells. Proc. 2002 IEEE Intl. Symp. Biomed. Imaging (ISBI 2004), 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., 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. Proc. 2002 IEEE Intl. Symp. Biomed. Imaging (ISBI 2002), pp. 325–328.

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.

Ichimura, N. (1997). Robust clustering based on a maximum-likelihood method for estimating a suitable number of clusters. *Syst. Comput. Jpn.* **28,** 10–23.

Jarvik, J. W., Adler, S. A., Telmer, C. A., Subramaniam, V., and Lopez, A. J. (1996). CD-tagging: A new approach to gene and protein discovery and analysis. *BioTechniques* **20,** 896–904.

Jarvik, J. W., Fisher, G. W., Shi, C., Hennen, L., Hauser, C., Adler, S., and Berget, P. B. (2002). *In vivo* functional proteomics: Mammalian genome annotation using CD-tagging. *BioTechniques* **33,** 852–867.

Jiang, X. S., Zhou, H., Zhang, L., Sheng, Q. H., Li, S. J., Hao, P., Li, Y. X., Xia, Q. C., Wu, J. R., and Zeng, R. (2004). A high-throughput approach for subcellular proteome: Identification of rat liver proteins using subcellular fractionation coupled with two-dimensional liquid chromatography tandem mass spectrometry and bioinformatic analysis. *Mol. Cell. Proteom.* **3,** 441–455.

Kozubek, M., Matula, P., Matula, P., and Kozubek, S. (2004). Automated acquisition and processing of multidimensional image data in confocal *in vivo* microscopy. *Microsc. Res. Tech.* **64,** 164–175.

Kumar, A., Cheung, K.-H., Ross-Macdonald, P., Coelho, P. S. R., Miller, P., and Snyder, M. (2000). TRIPLES: A database of gene function in Saccharomyces cerevisiae. *Nucleic Acids Res.* **28,** 81–84.

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

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D. S., Poulin, B., and Anvik, J. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* **20,** 547–556.

Lujan, R. (2004). Electron microscopic studies of receptor localization. *Methods Mol. Biol.* **259,** 123–136.

Macbeath, G. (2002). Protein microarrays and proteomics. *Nat. Genet.* **32,** 526–532.

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

Mitchell, T. M. (1997). "Machine Learning." WCB/McGraw-Hill, New York.

Murphy, R. F., Boland, M. V., and Velliste, M. (2000). Towards a systematics for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8,** 251–259.

Murphy, R. F., Velliste, M., and Porreca, G. (2002). Robust classification of subcellular location patterns in fluorescence microscope images. Proc. 2002 IEEE Intl. Workshop Neural Networks Sig. Proc. (NNSP '03), pp. 67–76.

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.

Nadler, M., and Smith, E. P. (1993). "Pattern Recognition Engineering." John Wiley & Sons, Inc., New York.

Nair, P., Schaub, B. E., Huang, K., Chen, X., Murphy, R. F., Griffith, J. M., Geuze, H. J., and Rohrer, J. (2005). Characterization of the TGN exit signal of the human mannose 6-phosphate uncovering enzyme. *J. Cell Sci.* **118,** 2949–2956.

Nair, R., and Rost, B. (2002). Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* **18,** S78–S86.

Nair, R., and Rost, B. (2003). LOC3D: Annotate sub-cellular localization for protein structures. *Nucleic Acids Res.* **31,** 3337–3340.

Nair, R., and Rost, B. (2004). LOCnet and LOCtarget: Sub-cellular localization for structural genomics tar. *Nucleic Acids Res.* **32,** W517–W521.

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

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14,** 897–911.

Nakano, A. (2002). Spinning-disk confocal microscopy—a cutting-edge tool for imaging of membrane traffic. *Cell Struct. Funct.* **27,** 349–355.

Norin, M., and Sundstrom, M. (2002). Structural proteomics: Developments in structure-to-function predictions. *Trends Biotechnol.* **20,** 79–84.

O'Brate, A., and Giannakakou, P. (2003). The importance of p53 location: Nuclear or cytoplasmic zip code? *Drug Resist. Updates* **6,** 313–322.

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., and Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science* **306,** 1194–1198.

Qian, G., Sural, S., Gu, Y., and Pramanik, S. (2004). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. Proc. 2004 ACM Symp. Applied Comput., pp. 1232–1237.

Reichart, B., Klafke, R., Dreger, C., Kruger, E., Motsch, I., Ewald, A., Schafer, J., Reichmann, H., Muller, C. R., and Dabauvalle, M.-C. (2004). Expression and localization of nuclear proteins in autosomal-dominant Emery-Dreifuss muscular dystrophy with LMNA R377H mutation. *BMC Cell Biol.* **5,** 12.

Reinhardt, A., and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* **26,** 2230–2236.

Rohrer, J., and Kornfeld, R. (2001). Lysosomal hydrolase mannose 6-phosphate uncovering enzyme resides in the trans-Golgi network. *Mol. Biol. Cell* **12,** 1623–1631.

Rolls, M. M., Stein, P. A., Taylor, S. S., Ha, E., McKeon, F., and Rapoport, T. A. (1999). A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J. Cell Biol.* **146,** 29–44.

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

Sclaroff, S., and Liu, L. (2001). Deformable shape detection and description via model-based region grouping. *IEEE Trans. Pattern Anal. Machine Intell.* **23,** 475–489.

Singh, A. K., Manjunath, B. S., and Murphy, R. F. (2004). Design of a distributed database for biomolecular images. *SIGMOD Rec.* **33,** 65–71.

Steckling, T., Hellwich, O., Wälter, S., and Wanker, E. (2004). Protein classification by analysis of confocal microscopic images of single cells. Proc. XXth ISPRS Congress, pp. 302–305.

Subramaniam, S., and Milne, J. L. (2004). Three-dimensional electron microscopy at molecular resolution. *Annu. Rev. Biophys. Biomol. Struct.* **33,** 141–155.

Telmer, C. A., Berget, P. B., Ballou, B., Murphy, R. F., and Jarvik, J. W. (2002). Epitope tagging genomic DNA using a CD-tagging Tn10 minitransposon. *BioTechniques* **32,** 422–430.

Velliste, M., and Murphy, R. F. (2002). Automated determination of protein subcellular locations from 3D fluorescence microscope images. Proc. 2002 IEEE Intl. Symp. Biomed, Imaging (ISBI 2002), pp. 867–870.

Zhao, T., Velliste, M., Boland, M. V., and Murphy, R. F. (2005). Object type recognition for automated analysis of protein subcellular location. *IEEE Trans. Image Process* **14,** 1351–1359.

Ziauddin, J., and Sabatini, D. M. (2001). Microarrays of cells expressing defined cDNAs. *Nature* **411,** 107–110.