Chapter 11

Location Proteomics: Systematic Determination of Protein Subcellular Location

Justin Newberg, Juchang Hua, and Robert F. Murphy

Summary

Proteomics seeks the systematic and comprehensive understanding of all aspects of proteins, and location proteomics is the relatively new subfield of proteomics concerned with the location of proteins within cells. This review provides a guide to the widening selection of methods for studying location proteomics and integrating the results into systems biology. Automated and objective methods for determining protein subcellular location have been described based on extracting numerical features from fluorescence microscope images and applying machine learning approaches to them. Systems to recognize all major protein subcellular location patterns in both two-dimensional and three-dimensional HeLa cell images with high accuracy (over 95% and 98%, respectively) have been built. The feasibility of objectively grouping proteins into subcellular location families, and in the process of discovering new subcellular patterns, has been demonstrated using cluster analysis of images from a library of randomly tagged protein clones. Generative models can be built to effectively capture and communicate the patterns in these families. While automated methods for high-resolution determination of subcellular location are now available, the task of applying these methods to all expressed proteins in many different cell types under many conditions represents a very significant challenge.

Key words: Location proteomics, Subcellular location trees, Subcellular location features, Fluorescence microscopy, Pattern recognition, Cluster analysis, Generative models, CD-tagging, Systems biology.

1. Introduction

A critical aspect of the analysis of a proteome is the collection of detailed information about the subcellular location of all of its proteins. Since subcellular location can change during the cell cycle and in response to internal (mutation) or external (drugs, hormones, metabolites) effectors, the acquisition of sufficient

Ivan V. Maly (ed.) *Methods in Molecular Biology, Systems Biology, vol. 500*

DOI: 10.1007/978-1-59745-525-1_11

information for even a single protein can be challenging. Two strategies are possible: experimental *determination* and computational *prediction*.

The former approach involves assigning class labels to data using automated learning methods. Depending on the application, classes can take on different meaning. Typically in location proteomic studies, various proteins or organelles define classes. If the classes of the data samples are known (in other words, if the data are class-labeled), then supervised learning approaches can be used, wherein classifiers are trained to distinguish between the classes, and new data can be automatically labeled as belonging to these classes. If data is not class labeled, then unsupervised learning approaches can be used, typically to group data by similarity and to identify important clusters in a dataset. In location proteomics, these clusters can correspond to important protein or organelle patterns.

A range of approaches to predicting location from sequence have been described, including detection of targeting motifs, analysis of amino acid composition, and modeling based on sequence homology *(1–6)*. What is clear is that all subcellular location prediction systems suffer from deficiencies in the training data: a limited number of proteins with known locations *and* insufficiently detailed descriptions for those that have been determined. This is because raw experimental data are converted into words that describe location, and both the process of assigning words and the limitations of the words themselves create loss of information. This is true even when standardized terms such as the Cellular Component terms from the Genome Ontology *(7)* are used. (Of course, many determinations of location are done by microscopy at low magnification and therefore the resolution of the imaging becomes the limiting factor.) There is thus an urgent need to collect new protein subcellular location data with high resolution. We first consider approaches using visual assignment of location.

Such efforts can be characterized along three dimensions: whether or not the approach used involves a selective *screen* for a particular location, whether or not the proteins to be analyzed are chosen *randomly*, and whether or not the resolution of the determinations is at or near the limit of optical microscopy. Tate et al. *(8)* used a gene trap approach to screen for proteins localized in the nucleus of mouse embryonic stem cells. Rolls et al. *(9)* used a cDNA library fused with GFP to screen for proteins with nuclear envelope distributions. Similarly, Misawa et al. *(10)* used a GFP-fusion cDNA library to identify 25 proteins showing specific intracellular localization. In contrast, Simpson et al. *(11)* used N- and C-terminal GFP fusion of cDNAs to assign locations to more than 100 novel proteins in monkey Vero cells, while Jarvik et al. *(12)* used random genomic tagging (CD-tagging) to create

more than 300 GFP-expressing cell clones and assign locations. Huh et al. *(13)* created a even larger library of 6,029 yeast strains with GFP-tagged ORFs (open reading frames) to characterize the localization of yeast proteins.

While the vast majority of studies of protein location using fluorescence microscopy have employed visual interpretation of the resulting images, there have been efforts to bring automation to this process *(14–23)*. These have been based on work over the past decade demonstrating not only that computational analysis can be used to recognize *known* subcellular location patterns *(24–30)* but also that the accuracies achieved are equal to, and in some cases better than, those of visual analysis *(17)*.

Images from many of these studies are publicly available. **Table 1** summarizes some of these and other studies and illustrates how they are different by design. In addition, Schubert et al. *(21)* have developed multiepitope ligand cartography, a robotically controlled immunofluorescence microscopy system that can capture as many as 100 distinct antibodies in the same

Table 1
Data collections relevant to location proteomics

Project	Species (cell type)	Number of proteins	Public access	Tagging method	2D/3D	Mag
Yeast GFP fusion localization database	Yeast	>4,000	yeastgfp.ucsf.edu	cDNA c-terminal GFP fusion	2D	100×
Human Protein Atlas	Human (>40 tissue types)	>6,000	proteinatlas.org	Immunohistochemical staining	2D	20×
CD-tagging database	Mouse 3T3	>100	cdtag.bio.cmu.edu	Internal GFP fusion	3D	60×
GFP-cDNA localization project	Human (HeLa) and monkey (Vero)	>1,000	gfp-cdna.embl.de	cDNA terminal GFP fusion	2D	63×
Protein subcellular location image database	Human (Hela) and mouse (3T3)	>100	pslid.cbi.cmu.edu	Immunofluorescence and genomic internal GFP fusion	2D/3D	100× + 60×
Cell centered database	Various	Various	ccdb.ucsd.edu	Various	2D/3D	60×–40,000×

image sample, but collections of images from this approach are not yet publicly available.

This review briefly covers the process of data collection for determination of subcellular location, followed by a more detailed discussion of a range of automated methods for analysis of the resulting images. The large scale application of these methods over the next few years will help to address the need for large sets of proteins with well-characterized locations, and this in turn will further aid development of future systems capable of modeling and predicting subcellular location.

2. Acquisition of Protein Subcellular Location Images

Perhaps the most common method for determining the subcellular location of a protein is to label the protein with a fluorescent probe and then to visualize the distribution of the protein within cells under a fluorescence microscope. We will limit our discussion to variations on this approach, and we will not consider alternatives such as cell fractionation followed by protein identification and quantitation. Such approaches have been described *(18, 20)* but are fundamentally limited by the resolution of the fractionation step.

A typical fluorescence microscope consists of a light source such as an arc lamp or laser. Light passes through an excitation filter that allows only a specific wavelength through. Next, a condenser focuses the light onto the sample. This excites fluorophores in the sample to emit higher wavelength light that passes through the objective and then an emission filter that removes any undesired wavelengths. Next, the emitted, filtered light hits the detector (a photomultiplier tube or CCD-camera) and is stored digitally as a grayscale image. Multiple filter sets and corresponding probes can be thus used to obtain multiple grayscale images, producing a multichannel image.

The various approaches to tagging a protein for fluorescence microscopy can basically be divided into those that tag native proteins with a fluorescent dye and those that modify the coding sequence of the protein to introduce a fluorescent group into the molecule (for review *see* **refs.** *9, 14)*.

Native proteins are most commonly tagged in situ using antibodies conjugated with a fluorescent dye, but fluorescent probes that can specifically bind to a protein, such as phalloidin binding to F-actin, are also used. However, these approaches cannot usually be applied to a living cell, since the cell membrane has to be made permeable for the probes to enter the cell; moreover, they also require antibodies or probes with appropriate specificity, which make them hard to apply on a proteome wide scale.

Significant efforts to apply immunolabeling at the proteomic level have been undertaken, notably by the Human Protein Atlas *(47)*.

Tagging of proteins by modifying their DNA sequence does not have the above disadvantages. This approach involves either modifying a coding sequence (cDNA) and then introducing this sequence into cells or modifying the genome sequence directly (in either a targeted or a random manner). One of the powerful random tagging techniques is CD tagging *(14)*. In this approach, the coding sequence of a green fluorescent protein (GFP) is inserted randomly into genomic DNA by a retroviral vector. Because the tagging happens to the genomic DNA, the modified protein keeps its original regulatory sequences and expression level. This is in contrast to cDNA modification, in which a constitutive, highly expressed promoter is usually used and thus the expression level of the protein is typically higher than normal. By repeatedly performing random tagging on cells of identical lineage, most of, or eventually all of the proteins within a given cell line can be tagged and have their subcellular locations determined.

3. Interpretation of Protein Subcellular Location Images

3.1. Subcellular Location Features

As mentioned earlier, systems for recognizing subcellular patterns in a number of cell types have been developed. The heart of each of these systems is a set of numerical features that quantitatively describe the subcellular location pattern in a fluorescence microscope image. These features, termed subcellular location features (SLFs), are designed to be insensitive to the position, rotation, and total intensity of a cell image *(29)*. The only requirement for the calculation of these SLFs is that each input image contain a single cell. This requirement can be met in multiple cell images by segmenting the images into single cell regions either manually or automatically, using approaches such as modified Voronoi tessellation *(28)*, watershed *(26, 27)*, levelset methods *(30)*, and graphical model methods *(29)*.

A specific nomenclature has been used to enable unambiguous references to the features used in a particular study. Sets of features are referred to using the prefix "SLF" followed by a set number. Individual features are referred to by the set name followed by a period and its index within the set. For example, SLF1 refers to the first set of features, and SLF1.2 refers to the second feature in this set. We briefly summarize the various types of SLFs below.

3.2. SLFs for 2D Images

Morphological features (SLF1.1–1.8). The high intensity blobs of pixels in fluorescence microscope images might be the first thing

a cell biologist looks at when trying to resolve subcellular location patterns. Morphological features mainly describe the characteristics of these blobs, or *objects*. An object is defined as a group of touching (connected) pixels that are above a threshold (the threshold is determined automatically). Eight morphological features have been defined *(15)* to describe the number, size, and relative position of the objects.

Edge features (SLF1.9–1.13). The edge features are calculated by first finding edges in the fluorescence image. These edges can be thought of as consisting of positions that have low intensity in one direction and high intensity in the opposite direction. The number of above-threshold pixels that are along an edge, the total fluorescence of the edge pixels, and measures of the homogeneity with which edges are aligned in the image are especially useful for characterizing proteins whose patterns are not easily divided into objects (such as cytoskeletal proteins). Proteins showing a radiating (star-like) distribution (such as tubulin) have low edge homogeneity, while those showing aligned fibers (such as actin) have higher edge homogeneity *(15)*.

Geometric features (SLF1.14–1.16). The starting point for these features is determination of the convex hull of the cell, which is defined as the smallest convex set which surrounds all above threshold pixels. Three features have been defined using the convex hull: the fraction of the area of the convex hull that is occupied by above threshold pixels, the roundness of the convex hull, and the eccentricity of the convex hull *(15)*.

DNA features (SLF2.17–2.22). The central landmark in eukaryotic cells is the nucleus, and thus having a parallel image of the DNA distribution of a cell is quite valuable. When this is present, a set of features can be calculated to measure quantities such as how far on average protein objects are from the nucleus, and how much overlap exists between the protein and DNA distributions *(15)*.

Haralick texture features (SLF3.66–3.78). For patterns that are not easily decomposed into objects using thresholding, measures of image *texture* are often very useful. Texture features are calculated as various statistics defined by Haralick *(24)* that summarize the relative frequency with which one gray level appears adjacent to another one. Adjacency can be defined in the horizontal, vertical, and two diagonal directions in two-dimensional (2D) images. The texture features are averaged over these four directions to achieve rotational invariance. These features were first introduced for classification of cell patterns in the initial demonstration of the feasibility of automated subcellular pattern analysis *(25)*.

Zernike moment features (SLF3.17–3.65). Like the convex hull and texture features, the rationale behind using these moment features is to capture general information about the dis-

tribution of a protein in a rotationally invariant way. Because the Zernike moments are defined on the unit circle, a cell image is first mapped to the unit circle using polar coordinates, where the center of a cell is the origin of the unit circle. Then, the similarity between the transformed image and the Zernike polynomials are calculated by conjugation. By using the absolute value of the resulting moments, the features become rotation invariant *(25)*.

Skeleton features (SLF7.80–7.84). The goal behind these features is to characterize the shape of the objects found by thresholding. This is done by first obtaining the skeleton of each object by a recursive erosion operation on the edge. Each skeleton is then described by features such as its length and degree of branching, and these are averaged over all objects to give features of the cell as a whole *(17)*.

Daubechies 4 wavelet features (SLF15.145–15.174). The principle behind wavelet decomposition is to measure the response of an image to a filter (a wavelet) applied in the horizontal, vertical, and diagonal directions. Wavelet decomposition can be performed recursively, with each pass measuring the response of the filter at a lower frequency *(31)*. Thus the average energy (sum of squared intensities) at each level of decomposition of an image using a wavelet function provides (among other things) information on the frequency (size) distribution of fluorescent objects but without the need for thresholding.

Gabor texture features (SLF15.85–15.144). These features are calculated by convolving an image with a 2D Gabor filter and calculating the mean and standard deviation of the resulting image *(32)*. By using different parameters to generate the Gabor filter, a total of sixty Gabor texture features can be calculated *(33)*.

3.3. Classification of 2D Images

In a series of studies, the SLFs described above have been applied to a set of 2D HeLa cells images showing the distribution of nine proteins and a parallel DNA-binding probe *(15)*. The nine proteins that were labeled by immunofluorescence are located in the endoplasmic reticulum (the protein p63), the Golgi complex (the proteins giantin and gpp130), lysosomes (LAMP2), endosomes (transferrin receptor), mitochondria, nucleoli (nucleolin), and cytoskeleton (beta-tubulin and F-actin). These protein classes which represent the major organelles in a cell were combined with a DNA-stained nucleus class selected from the parallel DNA images to form a 10-class subcellular location dataset. Example images are shown in **Fig. 1**. To evaluate the performance of an automated classifier, 90% of the images in each class were used to train that classifier and then its accuracy was obtained by testing it with the remaining 10% of the images. The process was then repeated nine additional times using different training and testing sets under the constraint that each image appears in a test set only once (this approach is termed tenfold cross-validation), and

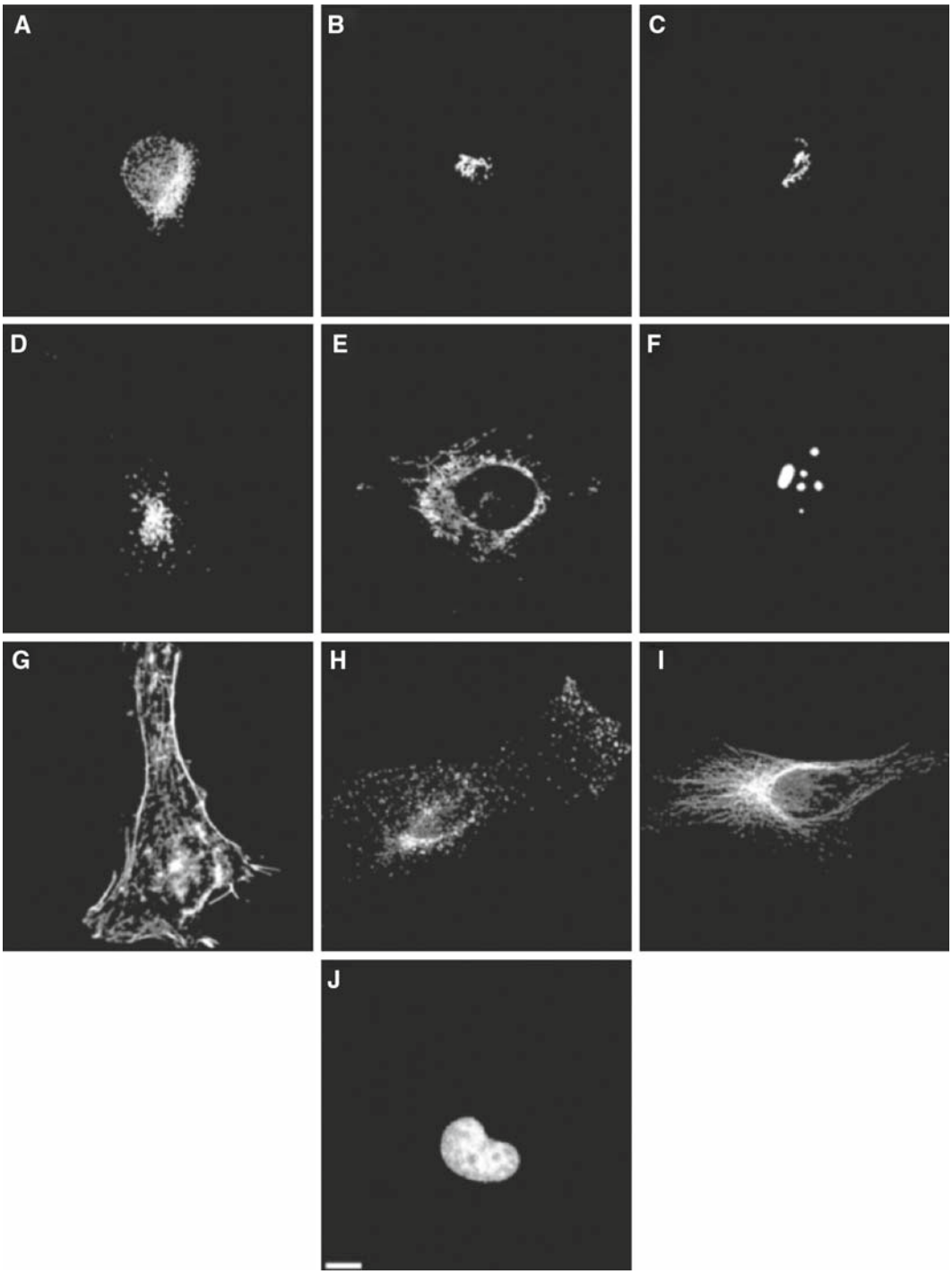

Fig. 1. Representative images of 2D HeLa dataset. These images have been preprocessed to remove background fluorescence and pixels below threshold. Images show the subcellular localization of (**A**) an ER protein, (**B**) Golgi protein giantin, (**C**) Golgi protein Gpp130, (**D**) lysosomal protein LAMP2, (**E**) a mitochondrial protein, (**F**) nucleolar protein nucleolin, (**G**) filamentous actin, (**H**) transferin receptor, (**I**) cytoskeleton protein tubulin, and (**J**) DNA. Scale bar = 10 μm. Reprinted from **ref.** *15* as allowed by Oxford University Press.

results from each repeat were averaged to get an overall classification accuracy. When the image set was first collected, an accuracy of 83% was obtained using a neural network classifier and a set of 37 SLFs *(15)*. Through the use of additional features and classifiers over the past few years, the accuracy on this dataset has risen to 92% for a majority-voting ensemble classifier using a set of 47 SLFs *(33)*. These results are shown in **Table 2**. Even better results have been obtained on this dataset using a multiresolution classification scheme that achieved an accuracy of 95% *(34)*. The automated systems are able to distinguish two Golgi proteins, GPP130 and giantin, which have been shown to be very hard to discriminate by visual inspection, as shown in **Table 3** *(17)*. A comparison of computer (**Table 2**) and human (**Table 3**) classifications is shown in **Fig. 2** *(35)*.

3.4. SLFs for 3D Images

Although most adherent cultured cells are very thin compared to their diameter in the plane of the substrate, a high resolution 2D image (which typically samples from only 0.5 to 1 μm in the axial direction) represents only a fraction of the compartments that are present in the three-dimensional (3D) cell. By taking 2D confocal microscope images at a series of depths within a cell, we can obtain a 3D image of a cell. Sampling in the axial direction is done typically every 0.5–2 μm, but depends on the microscope and the experimental design. Three types of 2D SLFs have

Table 2
Confusion matrix of 2D HeLa cell images using optimal majority-voting ensemble classifier with feature set SLF16

	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	**98.9**	1.2	0	0	0	0	0	0	0	0
ER	0	**96.5**	0	0	0	2.3	0	0	0	1.2
Gia	0	0	**90.8**	6.9	0	0	0	0	2.3	0
Gpp	0	0	14.1	**82.4**	0	0	2.4	0	1.2	0
Lam	0	0	1.2	0	**88.1**	1.2	0	0	9.5	0
Mit	0	2.7	0	0	0	**91.8**	0	0	2.7	2.7
Nuc	0	0	0	0	0	0	**98.8**	0	1.3	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	1.1	0	0	12.1	2.2	0	1.1	**81.3**	2.2
Tub	1.1	2.2	0	0	0	1.1	0	0	1.1	**94.5**

The overall accuracy was 92.3%. Data from **ref.** *33*

Table 3
Confusion matrix of human classification of images from 2D HeLa dataset

	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	**100**	0	0	0	0	0	0	0	0	0
ER	0	**90**	0	0	3	6	0	0	0	0
Gia	0	0	**56**	36	3	3	0	0	0	0
Gpp	0	0	53	**43**	0	0	0	0	3	0
Lam	0	0	6	0	**73**	0	0	0	20	0
Mit	0	3	0	0	0	**96**	0	0	0	0
Nuc	0	0	0	0	0	0	**100**	0	0	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	13	0	0	3	0	0	0	**83**	0
Tub	0	3	0	0	0	0	0	3	0	**93**

The overall accuracy was 83%. The major confusion came from the two Golgi protein, giantin and Gpp130, which were hard to distinguish by human inspection. Data from **ref.** *17*

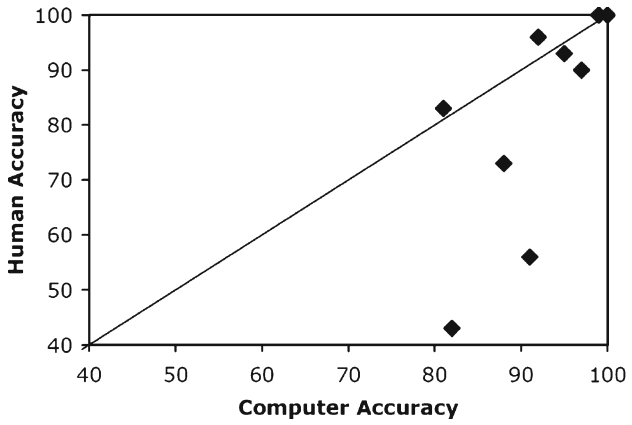

Fig. 2. Comparison of automated and visual classification of subcellular location patterns in 2D images of HeLa cells. Each *dark square* shows the classification accuracy of a specific pattern, while the *solid line* indicates equal performance between the two approaches. While six of the patterns are classified equally well by both, the computer performs significantly better on three of the patterns (two Golgi and one lysosomal). Reprinted from **ref.** *35* with permission (© 2004 IEEE).

been extended to three dimensions so that they can capture some information which is not available in 2D images. Results from automated classification using these 3D SLFs show improvement over the 2D SLF. Brief descriptions of these 3D features are presented below.

Morphological features (SLF9.1–9.28). The 3D morphological features are direct extensions of their 2D counterparts. Objects are found in 3D and size is replaced by volume. Moreover, distance features are decomposed into two components, one situated in the plane of the image and the other axially through the stack. Similar to the case for 2D images, a few 3D features (SLF9.9–9.14) can be defined relative to a parallel DNA image *(27)*.

Edge features (SLF11.15–11.16). The number of pixels along the edges and the total fluorescence of these pixels are calculated on every slice of the 3D images and then summed up. The fractions of these two values over the entire 3D image are used as 3D edge features *(16)*.

Haralick texture features (SLF11.17–11.42). The Haralick texture features can be extended to 3D images by considering all 13 directions in which a pixel can be considered adjacent to its neighbor pixels in 3D space (rather than the four directions in 2D space). The average value and the range of the 13 texture statistics over all 13 directions are used, yielding 26 features. Haralick texture features require a choice of image resolution and gray level bit depth to optimize the performance of recognizing patterns. Experiments revealed that 0.4 μm per pixel resolution and 256 (8 bit) gray levels were the best combination for recognizing subcellular patterns in the 3D HeLa image dataset described below *(36)*.

3.5. Classification of 3D Images

The 3D SLFs have been applied to a set of 3D HeLa images of the same nine proteins as in the 2D HeLa image collection *(27)*. A three-laser confocal microscope was used to record images of cells labeled simultaneously with three different probes (the images were collected in the Center for Biologic Imaging at the University of Pittsburgh with the kind assistance and support of Dr. Simon Watkins). In addition to probes for one of the nine targeted patterns, propidium iodide was used to stain DNA (after RNAse treatment), and a third probe was used to label total cell protein. The image of this third tag was used in combination with the DNA image to automatically segment images into single cell regions *(27)*.

The first evaluation of automated classification of this dataset used 28 morphological features, including 14 features which depend on the parallel DNA image. By using a neural network classifier, an overall accuracy of 91% was achieved *(27)*. To determine how well classification could be performed without using a

parallel DNA image, a new feature set SLF14 was created with 14 DNA-independent morphological images, two edge features, and 26 Haralick texture features. An overall accuracy of 98% was achieved using features selected from this set *(36)* as shown in **Table 4**. The results are nearly perfect, and the extension from 2D to 3D significantly increases the ability to distinguish the two Golgi proteins, Gpp130 and Giantin.

3.6. Clustering of Subcellular Location Images

The classification results described above have shown the ability of the SLFs to distinguish major subcellular location patterns with a classifier trained on class-labeled images. This is supervised learning, in which the protein or location classes are known at the outset. In contrast, unsupervised learning tries to find an optimal way of dividing *unlabeled* images into distinct groups or clusters. In location proteomics, clustering methods are used to find the major subcellular location pattern groups for all proteins across a proteome or large dataset. An optimal clustering on the location patterns of proteomes (finding subcellular location families) can offer a fundamental framework for assigning locations to proteins. Such a framework is useful for many reasons, one of which is because it can be used to automatically generate an ontology that effectively describes protein locations, and another of which is that each pattern (family) is tied to the images that defined it.

Table 4
Confusion matrix of 3D HeLa images using neural network classifier with seven features selected from SLF17

	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
Gia	0	0	**100**	0	0	0	0	0	0	0
Gpp	0	0	0	**96**	4	0	0	0	0	0
Lam	0	0	0	4	**95**	0	0	0	0	2
Mit	0	0	2	0	0	**96**	0	2	0	0
Nuc	0	0	0	0	0	0	**100**	0	0	0
Act	0	0	0	0	0	0	0	**100**	0	0
TfR	0	0	0	0	2	0	0	0	**96**	2
Tub	0	2	0	0	0	0	0	0	0	**98**

The overall accuracy was 98%. Data from **ref.** *36*

There are many different clustering algorithms and most of them require a similarity (or distance) function, which defines the way to calculate the similarity (or dissimilarity) of images in the feature space. Two well-known distance functions are Euclidean distance and Mahalanobis distance. Instead of unscaled Euclidean distances, which calculate the straight line distance in feature space between two images, standardized (or *z*-scored) Euclidean distances, which are Euclidean distances calculated after normalizing each feature to zero mean and unit variance, can be used. The Mahalanobis distance takes into account the correlation between features by scaling the distance with the covariance. Standardized Euclidean distance was shown to empirically produce the best agreement among different clustering algorithms applied to subcellular location images *(19)*.

k-Means clustering is a well-known centroid-based algorithm. Each data point is grouped into one of k clusters whose centroid is closest to it in the feature space. A centroid of a cluster is defined as the average feature vector of all the data points in that cluster. The starting centroids of the k clusters are randomly chosen from the data points or randomly generated. When a new data point is clustered into a certain cluster, the cluster centroid is updated accordingly. The process is repeated over all data points a few times until all the clusters converge.

To determine into how many groups the data should be clustered, an Akaike Information Content (AIC) score can be calculated for many values of k, the number of clusters. AIC measures the log-likelihood of the model penalized by the number of parameters of the model. A clustering result with small k and small variance of each cluster will have a relatively low AIC score, which means the clustering result is good. By varying k and comparing the AIC scores, an optimal k can be found *(37)*. Bayesian Information Criterion can be used in place of AIC.

Unlike the k-means clustering algorithm, hierarchical clustering does not depend on the choice of the number of clusters. Initially, each of the data points is a cluster. The distances of all of the clusters are calculated pairwise and the closest two clusters are joined together. This is repeated until all are joined. The result of hierarchical clustering shows how the clusters converge to fewer but larger clusters. A dendrogram is usually used to show the result of hierarchical clustering. A dendrogram generated from the SLFs of fluorescence microscope images has been termed a "subcellular location tree (SLT)" *(16)*. A SLT tells us how close the subcellular location pattern of one protein is to that of another protein. In order to increase the robustness of hierarchical clustering, consensus methods can be used *(19)*. In consensus clustering, a random half of images from each protein is used to build a hierarchical tree.

This is repeated and a consensus tree is built to show branches that are conserved *(38)*.

A third clustering approach is based on the confusion matrix generated by a classifier. This approach starts with training a classifier to discriminate all different proteins regardless of the possibility that some proteins may share the same location pattern. If two proteins actually do share a same location pattern, the classifier will not be able to tell them apart, which will then be shown in the confusion matrix as a large number in off-diagonal cells. By merging such confused proteins into a group, we can finally combine proteins which share a location pattern and obtain clusters which can be well separated by the classifier *(19)*.

As described before, the CD-tagging technique has been used to introduce an internal GFP domain in randomly targeted proteins in mouse 3T3 cells and to prepare a large library of subcellular location images *(12)*. 3D images have been collected for these clones using spinning disk confocal microscopy. The consensus clustering based on k-means algorithm divided 90 proteins into 17 groups, which represent the major location patterns distinguishable by the current 3D SLFs. A SLT was also generated on the same dataset. The proteins assigned to the same branch of the SLT often visually appear to display similar patterns. On the other hand, the proteins with distinct location patterns are well separated. This SLT (shown in **Fig. 3**) and the representative images of each leaf are available online at http://murphylab.web.cmu.edu/services/PSLID/ *(19)*. The whole process of building such a consensus SLT is automated and objective. The tree shows the major subcellular location patterns which are distinguishable in a collection of 90 different proteins in 3T3 cells as well as the hierarchical relations among these patterns. This clustering method is very promising to reveal the framework of protein subcellular location families when a complete image collection is available for all the proteins in a given cell type. Recently, images of 188 randomly tagged clones have been clustered into 35 distinct location clusters *(23)*. In addition to being used to group proteins by their location patterns, clustering of images has been used to group drugs by their effects upon subcellular patterns *(39)*.

3.7. Multiple Cell Image Analysis

Thus far we have discussed analysis of independent single cell regions. However, most fluorescence micrographs contain multiple cells per image field, and there is useful information in the spatial distribution of cells. Moreover, these cells may be expressing extracellular proteins of interest, and may be influencing each other (through things like cell division, hormonal signaling, or mechanical coupling).

There are various approaches to dealing with multicell images. The simplest are applied to images containing only one

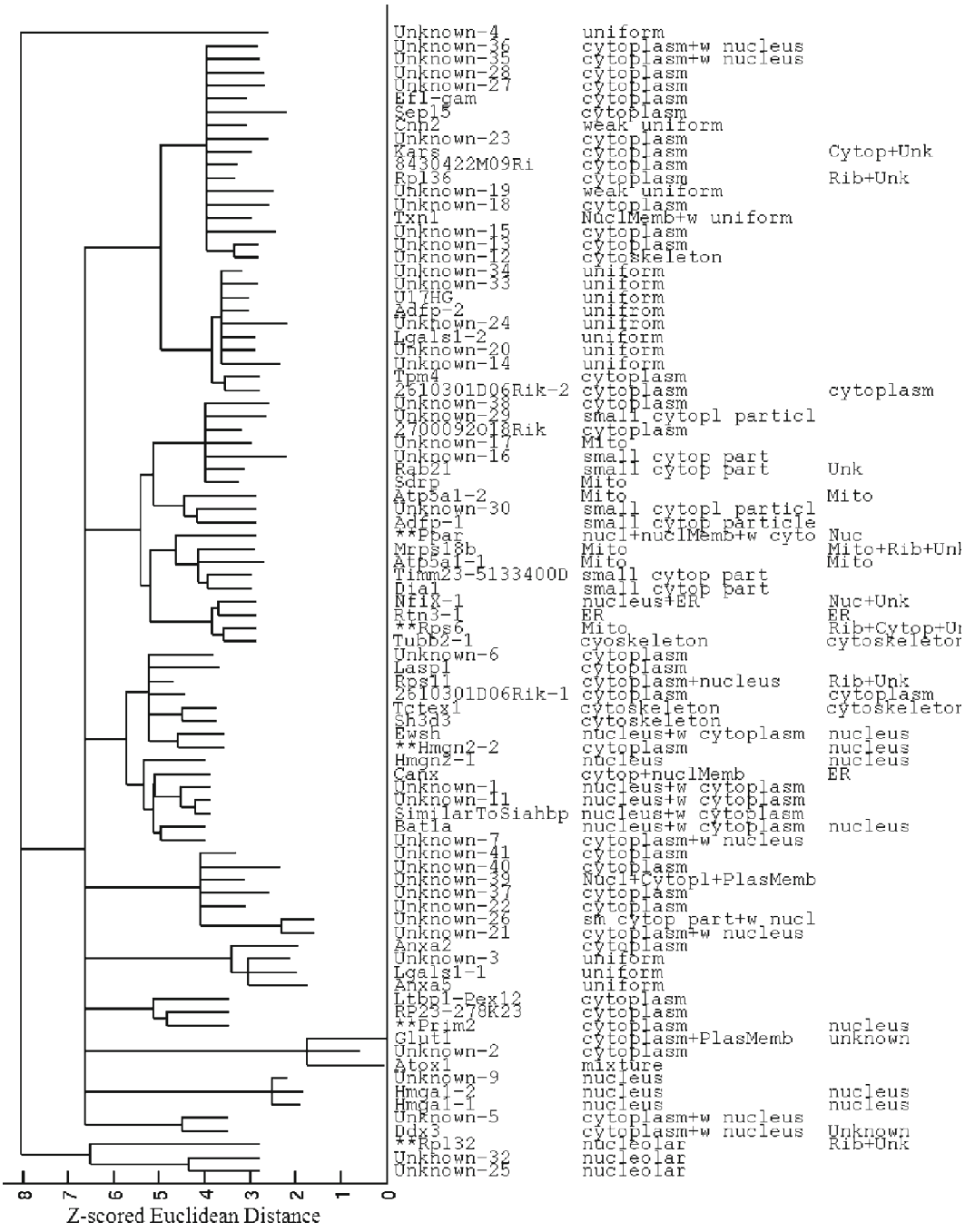

Fig. 3. A consensus subcellular location tree generated from the 3D 3T3. image dataset. The SLF11 feature set and standard (*z*-scored) Euclidean distance were used. The columns on the *right* of the tree show the protein name (if known), human observation of subcellular location, and subcellular location inferred from Gene Ontology (GO) annotations. Proteins marked with a *double asterisk* have significantly different locations between the description of human observation and the inference from GO annotation. Reprinted from **ref.** *19* under the terms of the Creative Commons License.

pattern in all of the cells. In one approach, field level features, which are independent of the number and rotation of cells in the image, are used to train classifiers. Huang and Murphy *(40)* showed that such features could be used to give a 95% accuracy. Their work was done using a modified version of the 2D HeLa images described above, where they used multiple single cell regions to synthesize multicell images containing anywhere from 2 to 6 cells. Following this work, Newberg and Murphy *(41)* showed that field level features combined with voting classification schemes can be used to effectively analyze protein patterns across human tissues. They trained a system that could distinguish between eight major organelle patterns with an 83% accuracy; this became 97% when only the most confident classification assignments were considered.

In another approach, information from surrounding cells is used to influence the classifier assignments for a local cell region in the image. This approach thus involves segmentation as a first step. If the image contains a homogeneous pattern (that is, all of the cells express the same protein pattern), simple voting methods can be used. These involve segmenting images, using SLFs in the classification of single cell regions, and then simply assigned the most common class label in the multicell image to all regions in that image. When multicell images contain more than one pattern (i.e., one group of cells expressing a tubulin pattern and another expressing a nuclear pattern), more complex voting schemes are needed. Chen and Murphy *(42)* showed that a graphical models approach can effectively deal with inhomogeneous data. This works by allowing close cell regions to have more influence than further away regions when deciding upon a class label for that region. Distances can be measured in both the physical space (where regions lie in an image) and feature space. Using synthetic multicell data (generated from the 2D HeLa image set), they were able to achieve greater than 90% accuracy in images containing up to four different types of patterns. This initial approach has been significantly improved and extended in subsequent work *(43, 44)*.

3.8. Object Type Recognition and Generative Models

The aforementioned methods consider protein subcellular location patterns at the level of each cell (or group of cells) and do not capture any information about the individual components of the cellular pattern. When they are applied to a new mixture pattern which combines the components from several different basic patterns (i.e., the location pattern of a protein which exists in different organelles or compartments), the cell level recognition methods tend to either generate a new location group (clustering) or simply be confused (classification). A more desirable result, however, might be a quantitative breakdown of how basic patterns compose the new mixed pattern *(45)*. To this

end, an object-based method was developed wherein object types are learned from several class-labeled images, and then they are used to recognize a new image pattern based on this pattern's object type composition. In this two-stage learning problem, first objects are extracted from known image classes and the object types are learned by clustering on object features, termed subcellular object features (SOFs). Note that objects in an image are defined as a group of connected pixels that are above some threshold. In the second stage, features, which describe the object type composition as well as the relative positions of these objects, are extracted from new mixture patterns. These in turn can be used to train classifiers to recognize the new patterns *(45)*.

This two-stage method has been applied to the previously described 2D HeLa dataset, which consists of ten different subcellular location classes. AIC-based k-means clustering on the extracted SOFs indicated that there were 19 unique object types in the images. Next, from each image sample, 11 SOFs and two SLFs were extracted for each of the 19 object types. A classifier was trained using a subset of these features to distinguish between the ten classes. Classification accuracy using cross-validation was 75%, and when the two Golgi apparatus proteins were merged, the accuracy increased to 82%. These results indicate that the SOFs and object types are informative for describing the protein patterns *(45)*.

The utility of these features and object types is that they can be used to characterize mixture patterns. Zhao et al. *(45)* demonstrated this using an unmixing approach to decompose mixture patterns into components of fundamental patterns. A linear regression method was first applied. It assumes that the features of a mixture pattern are linear combinations of the features of fundamental patterns. The coefficients (weights) of each fundamental pattern can be solved from linear equations. However, even in fundamental patterns, the fractions of each object type are not fixed. They vary from cell to cell. In a second unmixing approach, multinomial distributions were used to model the object type components of fundamental patterns and the fundamental pattern components of mixture patterns. The parameters of the model were then solved by the maximum likelihood method.

The object-type-based pattern recognition enables systems to recognize patterns composed of a mixture of components (object types) of the basic patterns. The learned object types can potentially be used to describe new subcellular location patterns or subtle protein location changes that might occur when cells are treated with drugs. More importantly, the recognition of the object types makes it possible to build generative models for protein location patterns. Zhao and Murphy *(46)* defined a method

that uses a three part model, with a nuclear, cell boundary, and protein component. Each component is learned separately, and the protein model uses object types at its core. In addition to capturing a subcellular pattern, the models also capture the variance of the pattern between images. Thus, these generative models can be used to create sets of images. The power of these generative models is that they, unlike conventional microscopy which only allows for a few proteins to be specifically imaged at a time, potentially allow for the creation of images that contain as many data channels as there are proteins in a proteome, and thus, these models are expected to become an essential tool for location proteomics and systems biology.

References

1. Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* 54, 277–344.
2. 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.
3. Guda, C., Fahy, E., and Subramaniam, S. (2004) MITOPRED: A genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* 20, 1785–1794.
4. Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D. S., Poulin, B., Anvik, J., Macdonell, C., and Eisner, R. (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 20, 547–556.
5. Chou, K. C., and Shen, H. B. (2006) HumPLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* 347, 150–157.
6. Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006) Prediction of protein subcellular localization. *Proteins* 64, 643–651.
7. 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., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261.
8. Tate, P., Lee, M., Tweedie, S., Skarnes, W. C., and Bickmore, W. A. (1998) Capturing novel mouse genes encoding chromosomal and other nuclear proteins. *J. Cell Sci.* 111, 2575–2585.
9. 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.
10. Misawa, K., Nosaka, T., Morita, S., Kaneko, A., Nakahata, T., Asano, S., and Kitamura, T. (2000) A method to identify cDNAs based on localization of green fluorescent protein fusion products. *Proc. Natl Acad. Sci. USA* 97, 3062–3066.
11. Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* 1, 287–292.
12. 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.
13. Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* 425, 686–691.

14. 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.
15. 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.
16. 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.
17. 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.
18. Jiang, X. S., Zhou, H., Zhang, L., Sheng, Q. H., Li, S. J., Li, L., 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. Proteomics* 3, 441–455.
19. Chen, X. and Murphy, R. F. (2005) Objective clustering of proteins based on subcellular location patterns. *J. Biomed. Biotechnol.* 2005, 87–95.
20. Drahos, K. L., Tran, H. C., Kiri, A. N., Lan, W., McRorie, D. K., and Horn, M. J. (2005) Comparison of Golgi apparatus and endoplasmic reticulum proteins from livers of juvenile and aged rats using a novel technique for separation and enrichment of organelles. *J. Biomol. Tech.* 16, 347–355.
21. Schubert, W., Bonnekoh, B., Pmmer, A. J., Philipsen, L., Bockelmann, R., Malykh, Y., Gollnick, H., Friedenberger, M., Bode, M., and Dress, A. W. M. (2006) Analyzing proteome topology and function by automated multi-dimensional fluorescence microscopy. *Nat. Biotechnol.* 24, 1270–1278.
22. Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., Raveh, T., Carpenter, A. E., Lahav, G., and Alon, U. (2006) Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nat. Methods* 3, 525–531.
23. Garcia Osuna, E., Hua, J., Bateman, N., Zhao, T., Berget, P., and Murphy, R. (2007) Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. *Ann. Biomed. Eng.* 35, 1081–1087.
24. Haralick, R., Shanmugam, K., and Dinstein, I (1973) Textural features for image classification. *IEEE Trans. Systems Man Cybernet.* SM S-3, 610–621.
25. 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.
26. Adiga, P. S. and Chaudhuri, B. B. (2000) Region based techniques for segmentation of volumetric histo-pathological images. *Comput. Methods Programs Biomed.* 61, 23–47.
27. Velliste, M. and Murphy, R.F. (2002) Automated determination of protein subcellular locations from 3D fluorescence microscope images. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging,* 867–870.
28. Jones, T.R., Carpenter, A.E., and Golland, P. (2005) Voronoi-based segmentation of cells on image manifolds. *ICCV Workshop on Computer Vision for Biomedical Image Applications,* 535–543.
29. Chen, S.-C., Zhao, T., Gordon , G.J., and Murphy, R.F. (2006) A novel graphical model approach to segmenting cell images. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology,* 1–8.
30. Coulot, L., Kirschner, H., Chebira, A., Moura, J.M.F., Kovacevic, J., Osuna, E.G., and Murphy, R.F. (2006) Topology preserving STACS segmentation of protein subcellular location images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging,* 566–569.
31. Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* 41, 909–996.
32. Daugman, J. D. (1988) Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoustics Speech Sig. Proc.* 36, 1169–1179.
33. Huang, K. and Murphy, R. F. (2004) Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinform.* 5, 78.
34. Chebira, A., Barbotin, Y., Jackson, C., Merryman, T., Srinivasa, G., Murphy, R. F., and Kovacevic, J. (2007) A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinform.* 8, 210.
35. Murphy, R.F. (2004) Automated interpretation of subcellular location patterns. 2004

IEEE International Symposium on Biomedical Imaging, 53–56.

36. Chen, X. and Murphy, R. F. (2004) Robust classification of subcellular location patterns in high resolution 3D fluorescence microscope images. *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 1632–1635.
37. Ichimura, N. (1997) Robust clustering based on a maximum-likelihood method for estimating a suitable number of clusters. *Syst. Comput. Jpn* 28, 10–23.
38. Thorley, J. L. and Page, R. M. (2000) RadCon: Phylogenetic tree comparison and consensus. *Bioinformatics* 16, 486–487.
39. 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.
40. Huang, K. and Murphy, R.F. (2004) Automated classification of subcellular patterns in multicell images without segmentation into single cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging,* 1139–1142.
41. Newberg, J. Y. and Murphy, R. F. (2008) A framework for the automated analysis of subcellular patterns in human protein atlas images. *J. Proteome Res.* 7, 2300–2308.
42. Chen, S.-C. and Murphy, R. F. (2006) A graphical model approach to automated classification of protein subcellular location patterns in multi-cell images. *BMC Bioinform.* 7, 90.
43. Chen, S.-C., Gordon, G., and Murphy, R.F. (2006) A novel approximate inference approach to automated classification of protein subcellular location patterns in multi-cell images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging,* 558–561.
44. Chen, S.-C., Gordon, G. J., and Murphy, R. F. (2008) Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns. *J. Mach. Learning Res.* 9, 651–682.
45. 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.
46. Zhao, T. and Murphy, R. F. (2007) Automated learning of generative models for subcellular location: Building blocks for systems biology. *Cytometry Part A***71A**, 978–990.
47. Uhlen et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics*, 4, 1920–1932.