# Automated Subcellular Location Determination and High-Throughput Microscopy

Estelle Glory[1] and Robert F. Murphy[1,*]
[1] Center for Bioimage Informatics, Molecular Biosensor and Imaging Center, and Departments of Biological Sciences, Biomedical Engineering and Machine Learning, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA
*Correspondence: murphy@cmu.edu
DOI 10.1016/j.devcel.2006.12.007

**Dramatic advances in methods for protein tagging and the development of fully automated microscopes enable collection of unprecedented volumes of image data on the subcellular location of proteins in live cells. Combining these approaches with machine learning methods promises to provide systematic, high-resolution pattern information on a proteome-wide basis.**

## Introduction

To understand the intricate pathways that regulate biological processes at the cellular level, we need to be able to capture data about the subcellular distributions of proteins and how these vary within cell populations. Automated analysis of fluorescence microscope images provides a powerful way of acquiring such information. The high specificity of fluorescent probes for labeling components of interest and the availability of advanced light microscopes permit high spatial and temporal resolution imaging of living cells. The determination of accurate protein location provides valuable information for understanding the molecular mechanisms that underlie the functions of cells (Ehrlich et al., 2002; Suzuki et al., 2002; Wu et al., 2003).

Knowledge of the localization of proteins within cellular compartments is critical to understanding their function for many reasons. Each compartment is defined by its own chemical and physical characteristics, such as the acidic pH in the lysosome, the viscoelasticity of the cytoskeleton, or the hydrophobicity of membrane. Thus, location can provide useful information for improving predictions of protein conformation. Besides, since organelles are the location of specialized functions in the cell, such as oxidative metabolism in mitochondria, transcription of ribosomal RNA in nucleoli, and maturation of newly synthesized proteins in the endoplasmic reticulum, the determination of subcellular location for a protein can yield hypotheses about the metabolism in which it is involved and the proteins with which it interacts. Changes in location over time are also critical to cell behavior. For example, in a signal transduction pathway, the transportation from the cytoplasm to the nucleus induced by the activation of the protein is characterized by the location of the protein before and after its activation, the activation moment, and its duration. Lastly, once the subcellular distribution of a protein is defined for healthy adult cells, comparison with diseased or developing cells can yield important insights that can lead to improved diagnostics and therapeutics.

Information on the subcellular location of proteins is increasingly being collected in parallel for large numbers of proteins (Hoja et al., 2000; Jarvik et al., 2002; Koroleva et al., 2005; Rolls et al., 1999; Simpson et al., 2000) or even for entire proteomes (Huh et al., 2003). As for many previous studies of individual proteins, the primary means of analyzing and annotating images depicting subcellular location in these large-scale studies has been visual examination. Over the past decade, however, the feasibility of using machine learning methods to automate the determination of subcellular location from fluorescence microscope images has been demonstrated convincingly (Boland et al., 1997, 1998; Boland and Murphy, 2001; Huang and Murphy, 2004b). In fact, these methods can perform better than visual examination (Murphy et al., 2003). Over the same time period, automated systems for performing cell-based assays were developed and used by pharmaceutical companies to screen for drugs with desired effects (Taylor et al., 2001; Zhou and Wong, 2006). These systems, variously referred to as high-content screening or high-throughput microscopy systems, are increasingly being used for basic research on biological pathways (Pepperkok and Ellenberg, 2006; Perlman et al., 2004; Price et al., 2002; Sigal et al., 2006; Starkuviene et al., 2004; Yarrow et al., 2005). This article reviews the methods currently available for automated, large-scale determination of the intracellular location of fluorescent-labeled molecules within cells.

## Approaches to Systematic Analysis of Subcellular Location

Proteins can display highly specialized locations within cells, such as being present in just the mitochondrial inner membrane, just the rims of a particular Golgi cisterna, or just specific regions of a chromosome. However, the number and resolution of locations considered in subcellular location classification varies greatly between studies. The simplest location studies are interested in three compartments (nucleus, cytoplasm, and extracellular environment), while more accurate studies have considered 7 to 22 different subcellular structures (often depending on the organism).

### *Knowledge Capture*

Systematic efforts to catalog the subcellular locations of proteins typically use either knowledge capture or data-driven approaches. The first seeks to collect and organize information on location that has been collected over many years and published in the archival literature. A critical starting point for these efforts is the development of a standard vocabulary to describe location. While many vocabularies have been used over the years, the creation of the Cellular Component ontology by the Gene Ontology (GO) Consortium (Harris et al., 2004) has had a major impact. This ontology describes locations at the levels of subcellular structures and macromolecular complexes. It has been widely used to create manually curated databases, such as SGD (http://www.yeastgenome.org) and Wormbase (http://www.wormbase.org), that contain assignments of specific GO terms to known proteins. While extremely valuable, this approach has a number of limitations. Perhaps foremost is the inherent limitation of using words to describe complex subcellular patterns. Another is the difficulty of capturing differences in location for a given protein, whether due to discrepancies between published results or changes in location due to a difference in conditions, cell cycle phase, or cell type. Other limitations include possible inconsistencies between annotators (assignment of different sets of terms for the same pattern or the same set of terms for different patterns) and the typical absence of traceable justifications for assignments.

Automated approaches have the potential to address some of these limitations. For example, automatic analysis of text documents has been used to associate a protein name with its location (Stapley et al., 2002). As a further step, systems such as SLIF (http://slif.cbi.cmu.edu) can associate fluorescence microscope images from figures in online journal articles with the protein name and location described in the captions (Murphy et al., 2001, 2004). Since this approach links directly to a published image, the assignment of subcellular patterns is not limited to the caption content and can be refined at any time by automated reinterpretation of the image (without needing to repeat the experiment).

### *Prediction*

In contrast to most knowledge-capture approaches, data-driven approaches seek to directly associate location assignments with raw data, typically using well-characterized automated tools. These approaches can be subdivided into two categories: computational prediction and experimental determination. Prediction approaches are typically based on the analysis of nucleic acid and amino acid sequences to seek sorting or signal peptides, sequence profiles, and similarities with known protein families (Chou and Shen, 2006; Lu et al., 2004; Yu et al., 2006). Naturally, prediction systems are inherently limited by their training data. They cannot properly assign proteins with previously unseen location patterns, are typically only trained to predict locations with low resolution (i.e., only major organelle classes), and are not able to predict differential localization (e.g., changes in location due to cell cycle phase or developmental state).

### *Determination*

Thus there is a strong need for additional experimental determination of protein subcellular location. Although both subcellular fractionation and electron microscopy have been used to analyze location, fluorescence microscopy currently is the most powerful approach since it can provide high-resolution images of protein distributions in living cells. A major bottleneck of this approach has been the slowness and subjectivity of human examination of the images to determine protein distribution patterns. Fortunately, automated image interpretation methods have been developed to recognize subcellular location patterns, initially by our group (Boland et al., 1997, 1998; Boland and Murphy, 2001; Murphy et al., 2000) and then by others (Conrad et al., 2004; Danckaert et al., 2002). These approaches combine objective feature extraction and machine learning algorithms. They have been shown to be as robust as a human annotator for recognizing major protein location patterns, and even more sensitive for discriminating subtle pattern variations not distinguishable by visual examination (Murphy et al., 2003). In the remainder of this review article we will describe methods for automatically acquiring and analyzing images of protein subcellular location.

## Image Acquisition Considerations for Automated Analysis

Comprehensive, systematic determination of subcellular location from fluorescence microscopy images requires automation and coordination of several steps, such as cell preparation, image acquisition, image preprocessing, quantitative feature extraction, pattern classification, database storage, and finally statistical analysis and data modeling. Automated microscopes originally designed for drug screening are rapidly evolving and becoming suitable for fundamental biological research. A number of factors affecting the systematic collection of images for later automated interpretation of subcellular location are described below.

### *Labeling*

A range of methods have been developed for visualizing protein distributions within cells using fluorescence microscopy, and the advantages and disadvantages of each approach have been discussed in detail in a recent review (Giepmans et al., 2006). Briefly, immunofluorescence methods have the advantage that they do not require modification of the target protein, but they have the disadvantages that they require specific antibodies, are subject to potential disruption of cell architecture during fixation, and cannot be used to observe dynamic patterns in living cells. Genetic methods involve creating a chimeric protein that includes all or part of a protein fused with either the coding sequence of a fluorescent protein (such as green fluorescent protein) or an amino acid sequence that can be specifically bound by an externally added reagent (such as the membrane permeable biarsenical reagent FlAsH). The genetic approaches can be further subdivided into those that tag cDNA sequences and those that tag genomic DNA sequences. The tag can either be added at one of the ends of the protein coding sequence or at an internal site. The genome approaches have the advantage that endogenous regulatory sequences are typically retained. A particularly powerful approach to genomic tagging is CD tagging, which can create internal tags at exon-exon boundaries (Jarvik et al., 1996). It has been combined with retrovirus-mediated random insertion to identify subcellular locations for previously uncharacterized proteins (Jarvik et al., 2002; Sigal et al., 2006).

### Magnification and Resolution

High-throughput microscopes are typically equipped with 10× or 20× magnification for applications limited to screening large populations of cells with coarse compartment distinctions into nucleus, cytoplasm, and extracellular. Higher magnification is needed in order to be able to distinguish intracellular structures, and some high-throughput microscopes can provide it. Since oil-immersion objectives are not well-suited to multiwell plate screening, most systems can use only air objectives (which limits the numerical aperture). The choice of objective for a high-throughput application is ultimately determined by the types of high-throughput systems available, the size of the cells and structures of interest, and the desired time of acquisition (acquiring multiple fields of a multiwell plate with a 63× objective can take many hours).

The specific microscope objective being used determines the spatial resolution that can be achieved using it. Spatial resolution is defined as the smallest separation between two point sources that permits them to be resolved. This is given by the Rayleigh limit, $1.22\lambda/2NA$ (where l is the wavelength of emitted light and NA is the numerical aperture of the objective). For 520 nm light and a 1.3 NA objective, this corresponds to 244 nm. When using digital imaging, one would ideally sample the image formed by the microscope at twice this resolution (this is referred to as Nyquist sampling). The pixel size of the camera often determines whether this can be achieved (the diameter or width of each pixel in the camera chip divided by the magnification gives the size of each pixel in the sample plane). Of course, any binning of camera pixels (i.e., summing of a two-by-two set of pixels to give one value) increases the pixel size in the sample plane. For a 100× objective and a camera with 2 micron-wide pixels, the pixel size (with no binning) in the sample plane would be 200 nm.

This discussion applies generally to fluorescence imaging using a digital camera, but there are a number of approaches that can provide better resolution (so-called super-resolution). These typically require that imaging be conducted under conditions where individual fluorophores can be resolved. A particularly exciting variation on this approach is photoactivated localization microscopy, which builds up distributions of molecules by repeated rounds of photoactivation and bleaching (Betzig et al., 2006).

To provide some insight into the pixel size required to distinguish subcellular patterns, the performance of automated classifiers has been compared for images originally collected with a 100× objective after downsampling to varying degrees to simulate using lower magnification (Murphy et al., 2003). The results demonstrated that the classifier was robust to these changes for identifying the major organelles, with the loss of only several percentage points when the spatial resolution was divided by 2 or 3. However, pairs of similar patterns (such as two Golgi proteins or an endosomal and a lysosomal protein) that could be distinguished at the highest resolution showed significantly lower classification accuracies at the lower magnifications. The results confirmed the notion that the choice of resolution has to be in accordance with the context of the study. For studies of the location of unknown proteins, the results provide a strong argument that the highest possible resolution should be used.

### Dimension

The new generation of fluorescent microscope systems allows the acquisition of images of higher dimension (time and space) than traditional 2D epifluorescence microscopes. Confocal, two-photon and spinning disc microscopes provide images with high spatial and/or time resolutions for studying 3D phenomena by live cell, multispectral, time-lapse imaging. The comparison of 2D images and 2D time series showed that subcellular patterns were more accurately identified when temporal information was involved (Hu et al., 2006). Since cells are not flat, similar conclusions were drawn for 3D images compared with 2D images (Huang and Murphy, 2004b; Velliste and Murphy, 2002). It remains to be determined for which applications 3D imaging is required and when, for example, large numbers of 2D images can suffice.

### Channels

Traditional determination of subcellular location has relied heavily on comparison with the patterns of previously characterized proteins, especially by labeling the known and unknown proteins with different fluorescent probes and acquiring parallel images (referred to as colocalization). This approach worked reasonably well for investigations of single proteins in which an iterative approach could be used to determining location (initial imaging to suggest candidate compartments, followed by double labeling with markers of those compartments). It is much less feasible when considering thousands of proteins. Fortunately, results from our group strongly suggest that if sufficient images are acquired, automated approaches can distinguish very similar patterns without using colocalization (Chen and Murphy, 2005; Murphy et al., 2003). Nonetheless, automated analysis is aided by the inclusion of some simple reference channels, such as a DNA probe to provide a frame of reference (Huang and Murphy, 2004b) or plasma membrane or total protein markers to facilitate cell segmentation (De Solorzano et al., 2001; Velliste and Murphy, 2002).

### Number of Images

The number of images that need to be acquired for each unknown protein depends on the minimum accuracy demanded and whether it is desired to be able to identify new patterns or just assign new proteins to known patterns. From previous experiments, anywhere from one to ten images can give nearly perfect accuracy for the latter task, but 50 to 100 images per protein are needed to allow discovery of new patterns and/or for training a new classifier to recognize them.

## Subcellular Location Feature Extraction

Having discussed image acquisition considerations, we now turn to the task of numerically describing location

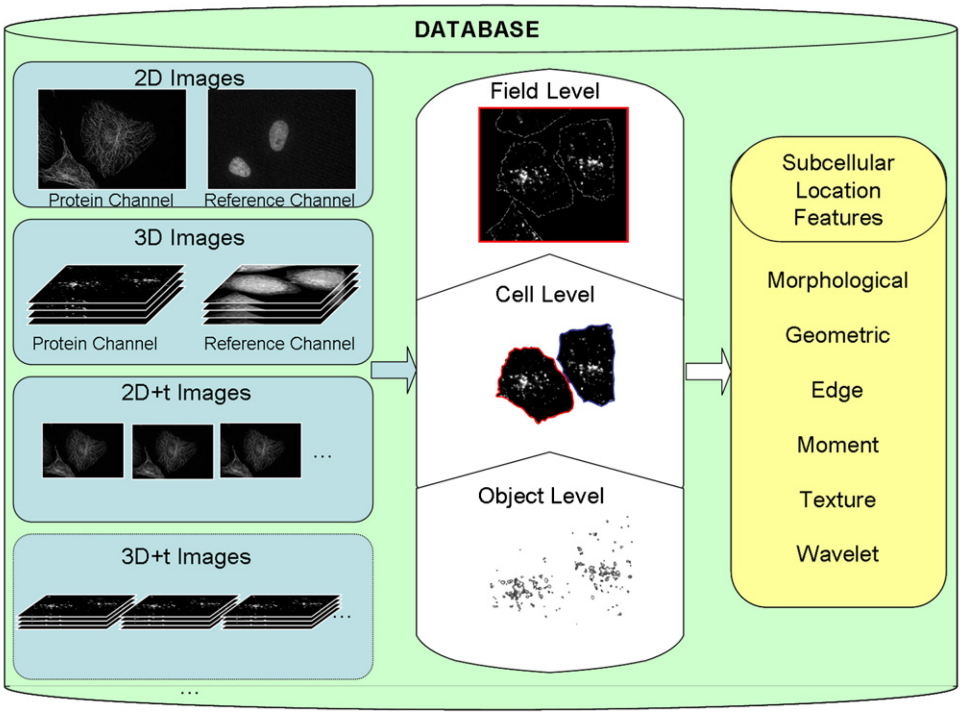

**Figure 1. The Image Database Depicted Contains Images with Related Biological Protocol, Acquisition Parameters, and Subcellular Location Features**

These numerical descriptors are computed at different semantic levels of the image content. The field-level features are calculated on the whole image, while cell- and object-level features require segmentation. The subcellular location features characterize the number, shape, gray-level distribution (texture, moments, and frequency), and relative size and position of the objects, in some cases relative to a reference channel. Some specific features are added for describing 3D and 2D+t stacks of images to improve the description by taking into account higher dimensions. As the number of dimensions increases by sampling in 3D or over time, more complex and informative features can be calculated.

patterns. The goal is to identify numerical features that capture the intrinsic properties shared by cellular organelles, while being insensitive to variations in cell shape, orientation, and position in images. Significant effort has been devoted to characterization of candidate features for this task, and the various feature types have been reviewed previously (Huang and Murphy, 2004c). Some of the descriptors are intuitive, such as the mean intensity or the shape of objects, while others are less intuitive, such as spatial frequency analysis or time variations. Computers can outperform human observations by capturing features not perceived by the eye (such as 3D distances, frequencies, and high-order statistical moments).

Microscope images of cultured cells can be described at different levels: an entire field of cells, individual whole cells, or individual fluorescent objects within cells (see Figure 1). Specific features can be defined as appropriate for each level, and lower-level descriptors can be aggregated to define upper-level descriptors (e.g., object features can be averaged to define cell-level features). The following sections present the various types of Subcellular Location Features that have been used to describe protein distribution patterns. Most of the features have 2D and 3D versions, which are computed as appropriate given the dimension of the available data. 3D measures take into account the fundamental difference between distance along the microscope axis (z) and distance in the focal plane (x,y) by separating the two components (Velliste and Murphy, 2002).

### *Subcellular Object Features*

An important step in describing subcellular patterns is defining individual fluorescently labeled objects within cells. The most frequent starting point for this task is finding a threshold to distinguish between negative (background) and positive pixels in the image. This can be done using methods (Otsu, 1979; Ridler and Calvard, 1978) that are suitable for fully automated processing. Objects are then defined as groups of contiguous above-threshold pixels. Each object can be described by a variety of features that reflect their size, shape, and position relative to a DNA reference channel, if available. These features can be used directly to recognize some subcellular patterns (Zhao et al., 2005) and can also be aggregated to form cell-level features (as described below).

### *Single-Cell Features*

The elementary entity studied to localize proteins within the cell is the cell itself. Analysis at this level therefore requires a segmentation step that divides each image into subregions containing individual cells on which the single-cell features are computed. This is a critical step of image analysis because its accuracy determines the accuracy of the resulting cell measurements, and, unfortunately, there is no universal method to tackle this problem in all possible conditions. The most common methods to segment fluorescent images are described below.

The most commonly used approach for generating single-cell regions is Voronoi segmentation. It involves finding the positions of nuclei in an image of a DNA probe and then creating polygonal regions that separate the nuclei. This approach is frequently used in high-content screens that do not require accurate cell boundaries. To get more accurate boundaries, additional information beyond the DNA image is needed. This can take the form of a parallel image in which the cell membrane or total protein is labeled. Nuclear positions can be found as above and then active contour methods (De Solorzano et al., 2001), the seeded watershed algorithm (Velliste and Murphy, 2002), or a modified Voronoi method (Jones et al., 2005), can be used. However, errors in finding nuclei can lead these algorithms to over- or undersegment an image (i.e., create regions containing partial or multiple cells).

This problem has led to the development of a number of approaches for improving cell segmentation by jointly considering nuclei finding and boundary finding. These include using parametric and geometric active contours (Coulot et al., 2006; Dufour et al., 2005; Zimmer et al., 2002), a combined filtering and watershed algorithm (Adiga et al., 2006), and graphical models (Chen et al., 2006). A drawback of these more complex segmentation approaches is that it can be difficult to find optimal values for free parameters in the algorithms. The methods are also quite computationally expensive for large data set applications.

Once individual cell regions have been identified, a range of features can be calculated to describe the fluorescence distribution within each region. Perhaps the most intuitive features are *morphological* features derived by calculating various quantities for the set of subcellular objects within each cell. For each object descriptor, cell-level morphological features can be calculated by finding means, variances, minima, and maxima. *Edge* features are derived by first finding those pixels in an image that are in regions where the intensity changes dramatically. Features that can then be calculated include how much of the total fluorescence is in the edge pixels and whether there is a preferred angle of the edges within a cell. The latter feature can be useful for distinguishing between patterns that show circular symmetry (such as microtubules radiating outward like a star) and those that don't (such as oriented stress fibers). *Texture* features are very powerful for distinguishing subcellular patterns. They are based on measurements of the frequency that any particular gray level is observed adjacent to any other gray level. A number of statistics can be calculated from these frequencies. The purpose of these statistics is to determine whether the overall pattern in a cell is more like a checkerboard than a solid or more like random speckling than a solid. Other types of more complex features include *geometric*, *moment*, and *wavelet* features. Detailed procedures for calculating all of these have been described (Huang and Murphy, 2004c).

#### *Field Features*

When automated microscope acquisition is used, images are taken independently of their content and generally contain cells truncated by the boundary of the image. Field features describe multicell fields without requiring cell segmentation, and they are expected to be insensitive to the number of cells in the field. Assuming that a field contains a homogeneous population of cells expressing a single labeled protein, most of the morphological features and all of the texture features used at the cell level can be used at the field level with good classification performance (Huang and Murphy, 2004a).

### Temporal Features

Dynamic cell population studies are becoming more and more important in understanding pathways and networks. Therefore, the addition of temporal parameters such as the change of size and shape of nuclei and the duration between the different stages are important indicators of the cell division cycle (Zhou and Wong, 2006). There is also extensive work on analyzing the behavior of specific labeled proteins (especially cytoskeletal and chromosomal proteins) by tracking individual objects in time series images (Meijering et al., 2006). This approach can yield exquisitely detailed models of how the target protein's distribution changes in space and time.

Tracking methods, however, require some description (model) of the type of object or structure to be tracked in a time series. This would be difficult to obtain when analyzing proteins on a proteome-wide basis (especially since many previously uncharacterized proteins would be present). An alternative is to use temporal versions of the texture features described above (Bouthemy and Fablet, 1998). Temporal textures measure overall patterns in the changes in pixel intensities in an image over time and do not require tracking of objects. This approach has been demonstrated to be able to distinguish protein patterns better than can be distinguished using features calculated from static images (Hu et al., 2006).

## Major Computational Questions in Subcellular Pattern Analysis

The subcellular location features described above are numerical descriptors that can characterize the distribution of proteins within cells. These features have been integrated in most of the advanced high-throughput image analysis systems (Carpenter et al., 2006; Conrad et al., 2004) and can be used for a number of different goals. However, among the hundreds of features, some are redundant while others are irrelevant to distinguishing a particular set of protein distribution patterns. Therefore, a significant increase in the quality of results can be achieved by selecting a specific set of discriminatory features before applying certain algorithms. A comparison between different feature reduction methods concluded that step-wise discriminant analysis is the most useful (Huang et al., 2003).

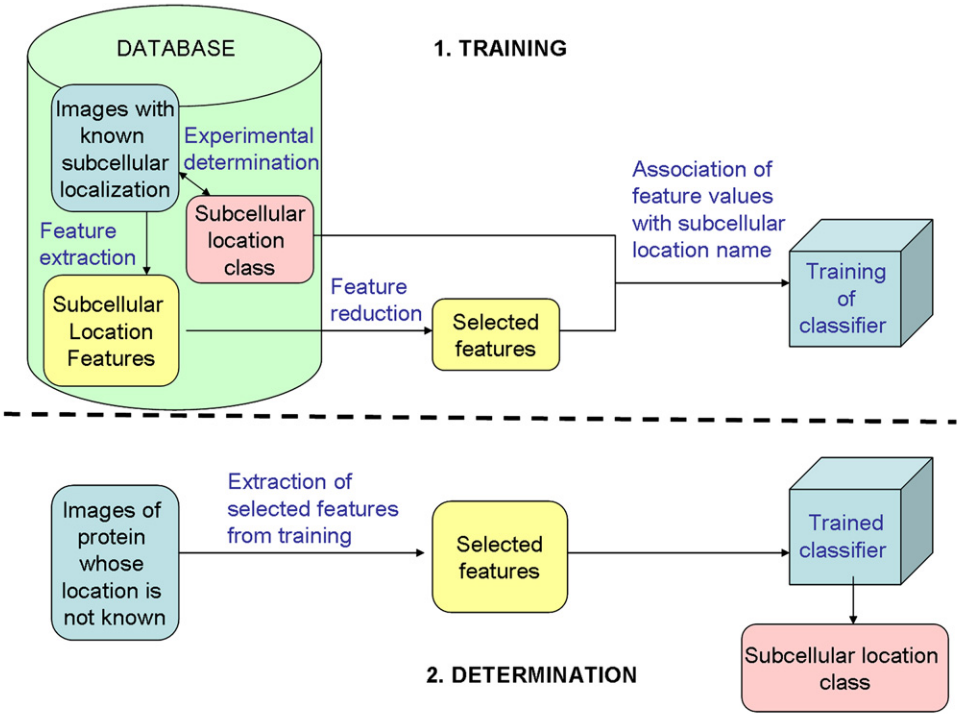
DATABASE
1. TRAINING
Images with known subcellular localization
Experimental determination
Subcellular location class
Feature extraction
Subcellular Location Features
Feature reduction
Selected features
Association of feature values with subcellular location name
Training of classifier
Images of protein whose location is not known
Extraction of selected features from training
Selected features
Trained classifier
Subcellular location class
2. DETERMINATION

**Figure 2. This Schematic Describes the Two Steps of the Supervised Classification for Determining the Subcellular Location Pattern of a Protein**

The first step requires a set of images which represents the different classes of subcellular location patterns to be recognized. The feature extraction provides a numerical description for each image. The classifier is trained to distinguish the subcellular location patterns given the values of a selected set of the most discriminative features. The second step determines the subcellular location class of a target protein from its fluorescent microscope images. The selected features are computed and used as inputs of the trained classifier. The classifier assigns one of the known classes to the protein in each image. The accuracy is improved when 3D stacks, 2D time sequences, or several images are used.

### *Statistical Tests: Comparison*

By defining a reference pattern from a protein under specific conditions, a screening application can compare it with the pattern of other proteins, known or unknown, to see if they share the same profile. This method is useful in identifying proteins likely to be involved in the same pathway, potential interacting proteins, or two components of the same protein complex. Another approach is to compare the reference of a protein distribution pattern with patterns found under modified conditions (e.g., a different stage of differentiation, pathology, or drug addition). In both examples, two sets of images have to be compared with a statistical test to determine if the difference between the two patterns is significantly different. That task can be automatically carried out by the SImEC (Statistical Image Experiment Comparator) software, which can determine whether two sets of images are likely to have come from the same distribution using various multivariate hypothesis tests (Roques and Murphy, 2002; Zhao et al., 2006). To refine the interpretation of the results, the software displays the features ranked from the most to the lowest degree of difference (calculated with the t test) between the two image sets.

An impressive high-throughput comparison study used a different approach to compare the effects of drugs on the distributions of a marker protein for various pathways (Perlman et al., 2004). A large number of features were calculated, and a series of univariate tests of the degree to which the histograms of those features differed between control and drug-treated samples were performed. The results formed a vector describing the responses for each drug, and these were clustered to identify categories of drugs that shared the same basic mechanism of action.

### *Supervised Learning: Classification*

Assuming that a protein is found in only one subcellular compartment, the determination of that compartment can be obtained by using the previous tool, SImEC, to compare the image set of the target protein with image sets of proteins with known subcellular locations. This is an inefficient process because it requires as many statistical tests as there are classes of known subcellular locations.

In the field of machine learning, this can be solved in one step with a supervised classifier. A supervised classifier is designed to assign a class to an unknown input, given a previous training set consisting of examples of each class. The input is the set of subcellular location features extracted from a fluorescent microscope image (see Figure 2). The performance of a classifier is measured by its accuracy in giving the correct class for known inputs that were not used in training. It can be easily estimated

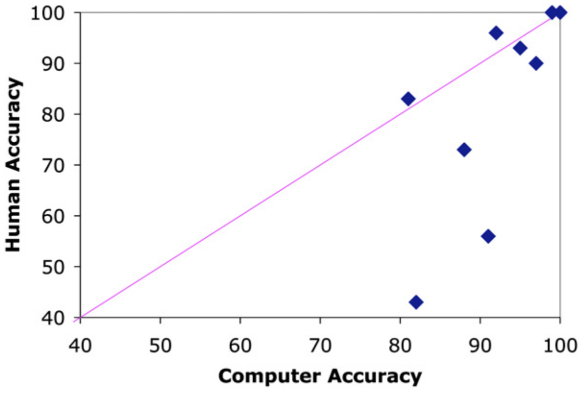

**Figure 3. Comparison of Classification Accuracies from an Automated System and from Visual Examination**
Accuracies for an automated classifier are presented versus the accuracies for the same images obtained by visual examination. Each symbol represents a different pattern class. In increasing order of human accuracy these are: gpp130, Giantin, LAMP2, TfR, ER, Tubulin, mitochondria, nucleolin, and DNA (both at 100% for human and 99% for computer accuracy), and actin (100% for both). From Murphy (2004).

by repeated splitting of the collection of known protein location images into one part for training and one for testing.

Many different types of classifiers exist, including linear classifiers, decision trees, k-nearest neighbor classifiers, one- and two-hidden-layer backpropagation neural networks, modular topology neural networks, and support vector machines (SVM). SVM have been observed to provide the best performance on two large image datasets (Huang and Murphy, 2004b). Most of these classifiers are readily available within image databases or statistical packages.

Comparison of the results on a dataset of 2D images of HeLa cells using an automated classifier (Huang and Murphy, 2004b) and visual examination (Murphy et al., 2003) is shown in Figure 3. Each symbol represents the accuracies for a different subcellular pattern. The overall accuracy for the automated classifier is 92%, as compared to 83% for visual examination. The computer's performance is very similar to the human performance for distinguishing the major patterns, but it is better for similar patterns, like those of endosomes and lysosomes or two Golgi proteins. Performance of automated classifiers on 3D images of the same patterns is even better, 98% (Huang and Murphy, 2004c).

### *Unsupervised Learning: Clustering*

The inherent restrictions of the supervised classification approach are the need for training set generation and the limitation that images can be assigned only to predefined classes. To tackle this problem, unsupervised learning methods, also called clustering methods, are able to define classes based only on the distance between the objects in the feature space. The distance between two points in the feature space can be used to estimate the degree of similarity between the two protein distributions they represent in the real world.

Clustering approaches have been applied to fluorescent microscope images of a large collection of cell lines generated by CD tagging (Jarvik et al., 2002) in order to create a subcellular localization tree which groups together similar protein location patterns (Chen and Murphy, 2005; Chen et al., 2003). The classes were autogenerated by computing the Mahalanobis distance between images and the mean vector for each class, while the connections of the hierarchical tree were determined by calculating the distance between each pair of classes. Since many images of each cell line (each expressing a different tagged protein) can be collected, a particularly robust form of clustering, consensus clustering (Thorley and Page, 2000), was used (Chen and Murphy, 2005). This ensures that the clusters are not adversely affected by one or a small number of atypical images. Clustering of proteins by their location patterns on a proteome-wide basis will enable the determination of the set of all possible (normal) location patterns and the identification of protein location families that share a given location.

## Extending Single-Cell Methods to Tissues

In addition to systematic studies of single cells, initial approaches to automated analysis of subcellular patterns in intact tissues have been described. Multicolor fluorescent staining of tissue microarrays has been used to distinguish tumor samples with different distributions of estrogen receptor or β-catenin (Camp et al., 2002). A powerful new robotic technology for sequential fluorescent staining of as many as a hundred proteins in fixed tissue samples has been shown to enable discrimination of disease-specific localization patterns (Schubert et al., 2006). In contrast to these studies, the Human Protein Atlas project (http://www.proteinatlas.org) has used immunocytochemical staining to obtain images of over 700 monospecific antibodies in 48 normal human tissues and 20 tumors (Uhlen et al., 2005). Each image was evaluated by a pathologist to identify the cell types and rough subcellular location of the targeted protein.

These studies have utilized fixed tissues and immunostaining, and the approaches promise to provide important information (especially for distinguishing various disease states). However, we can imagine that extensions of the live cell methods discussed above may be needed in order to understand the full dynamic behavior of all proteins in living tissues.

## Database versus File System Approaches

All of the methods described above can be implemented as stand-alone applications that use the operating system's file system to organize image files or as an analysis layer on top of an image database. Relational database models are convenient and widely used to capture biological information. Examples include the organism-specific genome databases and LIFEdb (*L*ocalization, *I*nteraction, *F*unctional assays and *E*xpression of Proteins) (http://www.LIFEdb.de). It is an example of a database that links genomic information, protein sequences, experimentally determined location, and predicted location with a sequence comparison tool (Mehrle et al., 2006). Another example is OrganelleDB (http://organelledb.lsi.umich.edu),

**Table 1. Examples of Questions about Subcellular Location Successfully Addressed by Computational Methods Incorporated in Publicly Available Fluorescence Microscope Image Databases**

| Questions | Method |
|---|---|
| Can I find images of a particular protein in this database? | Context-based image retrieval |
| How can images that look like a specific image be retrieved? | Content-based image retrieval |
| Do these two proteins have the same location pattern? | Statistical tests |
| Does the modification of the biological protocol change the location pattern of the target protein? | Statistical tests |
| What is the most representative image of this experiment/set of images? | Measuring distance in feature space from population mean |
| In what subcellular compartment is this protein? | Supervised classification |
| How can proteins that have the same location pattern be grouped together into families? | Clustering (unsupervised classification) and tree generation |

which gathers data from over a hundred organisms on the organelle, subcellular structure, or protein complex in which proteins are found.

Relational databases can also be used to associate images with their context. Two of the earliest image database systems for fluorescence microscope images were PSLID (Protein Subcellular Location Image Database) (Huang et al., 2002) and OME (Open Microscopy Environment) (Swedlow et al., 2003). OME is a general purpose, open source image database system that can be downloaded and installed in local imaging facilities. It includes excellent support for importing images from many microscope sources and contains tools for a wide range of microscope applications. On the other hand, PSLID is a specialized database dedicated to subcellular location images. The open source PSLID system can be downloaded and installed locally, but the PSLID website (http://pslid.cbi.cmu.edu) also provides public access to large collections of tagged protein images. It also provides access to tools to carry out all of the analysis described in this paper. These are summarized in Table 1.

Tools that do not use a database architecture often cannot provide the same level of sophistication of search and archiving that database systems can, but they have the major advantage that they can be significantly easier to install and use. An example is the CellProfiler system designed for analyzing changes in cell phenotypes during RNAi screens (Carpenter et al., 2006).

## Conclusions

The systematic analysis of protein location has received far less attention than other characteristics of proteins, such as structure and binding partners. The combination of powerful protein tagging methods, automated microscopes, and automated image interpretation methods provides the ability to comprehensively and automatically determine the location of tagged molecules within living cells.

The automation of protein localization by automated fluorescent microscopy image analysis is a powerful way to identify and understand the actors of pathways involved in the different stages of differentiation. Given the tools reviewed in this paper, it becomes possible to compare the localization of a target protein in immature cells with localization in cells engaged in different differentiation pathways. From these experiments, spatial and temporal models can be created. A particular challenge is deciding upon a strategy that specifies the numbers of proteins, cell types, developmental stages, and genetic backgrounds that should be acquired in order to build a comprehensive understanding of the variation in protein subcellular location during development.

The majority of available high-throughput microscope acquisition systems have been optimized for fixed cell applications; however, there is growing interest in live cell kinetic assays, and several systems have already successfully penetrated this application area. The methods presented in this article are completely adaptable to the available information in the resulting images (various spatial resolutions, 2D or 3D spatial information, and temporal resolution). Such analyses provide valuable information to feed standardized databases designed to tackle the challenges of systems biology.

## REFERENCES

Adiga, U., Malladi, R., Fernandez-Gonzalez, R., and de Solorzano, C.O. (2006). High-throughput analysis of multispectral images of breast cancer tissue. IEEE Trans. Image Process. *15*, 2259–2268.

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. Science *313*, 1642–1645.

Boland, M.V., and Murphy, R.F. (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics *17*, 1213–1223.

Boland, M.V., Markey, M.K., and Murphy, R.F. (1997). Classification of protein localization patterns obtained via fluorescence light microscopy. 19th Annu. Intl. Conf. IEEE Eng. Med. Biol. Soc., 594–597.

Boland, M.V., Markey, M.K., and Murphy, R.F. (1998). Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry *33*, 366–375.

Bouthemy, P., and Fablet, R. (1998). Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. 14th Intl. Conf. Pattern Recogn. *1*, 905–908.

Camp, R.L., Chung, G.G., and Rimm, D.L. (2002). Automated subcellular localization and quantification of protein expression in tissue microarrays. Nat. Med. *8*, 1323–1327.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

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.

Chen, S.-C., Zhao, T., Gordon, G.J., and Murphy, R.F. (2006). A novel graphical model approach to segmenting cell images. 2005 IEEE Symp. Comput. Intell. Bioinf. Comput. Biol., 1079.

Chou, K.C., and Shen, H.B. (2006). Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. *347*, 150–157.

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.

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. 2006 IEEE Intl. Symp. Biomed. Imaging, 566–569.

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.

Dufour, A., Shinin, V., Tajbakhsh, S., Guillen-Aghion, N., Olivo-Marin, J.-C., and Zimmer, C. (2005). Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. IEEE Trans. Image Process. *14*, 1396–1410.

Ehrlich, J.S., Hansen, M.D.H., and Nelson, W.J. (2002). Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion. Dev. Cell *3*, 259–270.

Giepmans, B.N., Adams, S.R., Ellisman, M.H., and Tsien, R.Y. (2006). The fluorescent toolbox for assessing protein location and function. Science *312*, 217–224.

Harris, M., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. *32*, D258–D261.

Hoja, M.R., Wahlestedt, C., and Hoog, C. (2000). A visual intracellular classification strategy for uncharacterized human proteins. Exp. Cell Res. *259*, 239–246.

Hu, Y., Carmona, J., and Murphy, R.F. (2006). Application of temporal texture features to automated analysis of protein subcellular locations in time series fluorescence microscope images. 2006 IEEE Intl. Symp. Biomed. Imaging, 1028–1031.

Huang, K., and Murphy, R.F. (2004a). Automated classification of subcellular patterns in multicell images without segmentation into single cells. 2004 IEEE Intl. Symp. Biomed. Imaging, 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., and Murphy, R.F. (2004c). From quantitative microscopy to automated image understanding. J. Biomed. Opt. *9*, 893–912.

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. 2002 IEEE Intl. Symp. Biomed. Imaging, 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.

Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Welssman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

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.

Jones, T.R., Carpenter, A.E., and Golland, P. (2005). Voronoi-based segmentation of cells on image manifolds. ICCV Workshop Comput. Vision Biomed. Image Appl., 535–543.

Koroleva, O.A., Tomlinson, M.L., Leader, D., Shaw, P., and Doonan, J.H. (2005). High-throughput protein localization in Arabidopsis using Agrobacterium-mediated transient expression of GFP-ORF fusions. Plant J. *41*, 162–174.

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.

Mehrle, A., Rosenfelder, H., Schupp, I., del Val, C., Arlt, D., Hahne, F., Bechtel, S., Simpson, J., Hofmann, O., Hide, W., et al. (2006). The LIFEdb database in 2006. Nucleic Acids Res. *34*, D415–D418.

Meijering, E., Smal, I., and Danuser, G. (2006). Tracking in molecular bioimaging. IEEE Signal Process. Mag. *23*, 46–53.

Murphy, R.F. (2004). Automated interpretation of subcellular location patterns. 2004 IEEE Intl. Symp. Biomed. Imaging, 53–56.

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. Intl. Conf. Intell. Syst. Mol. Biol. *8*, 251–259.

Murphy, R.F., Velliste, M., Yao, J., and Porreca, G. (2001). Searching online journals for fluorescence microscope images depicting protein subcellular locations. 2nd IEEE Intl. Symp. BioInf. Biomed. Eng., 119–128.

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.

Murphy, R.F., Kou, Z., Hua, J., Joffe, M., and Cohen, W.W. (2004). Extracting and structuring subcellular location information from on-line journal articles. IASTED Intl. Conf. Knowl. Sharing Collab. Eng., 109–114.

Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man. Cybernet. *9*, 62–66.

Pepperkok, R., and Ellenberg, J. (2006). High-throughput fluorescence microscopy for systems biology. Nat. Rev. Mol. Cell Biol. *7*, 690–696.

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.

Price, J.H., Goodacre, A., Hahn, K., Hodgson, L., Hunter, E.A., Krajewski, S., Murphy, R.F., Rabinovich, A., Reed, J.C., and Heynen, S. (2002). Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. J. Cell. Biochem. Suppl. *39*, 194–210.

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

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.

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.

Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., et al. (2006). Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nat. Methods *3*, 525–531.

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.

Stapley, B.J., Kelley, L.A., and Sternberg, M.J. (2002). Predicting the sub-cellular location of proteins from text using support vector machines. Pac. Symp. Biocomput. 374–385.

Starkuviene, V., Liebel, U., Simpson, J.C., Erfle, H., Poustka, A., Wiemann, S., and Pepperkok, R. (2004). High-content screening microscopy identifies novel proteins with a putative role in secretory membrane traffic. Genome Res. *14*, 1948–1956.

Suzuki, K., Kamada, Y., and Ohsumi, Y. (2002). Studies of cargo delivery to the vacuole mediated by autophagosomes in Saccharomyces cerevisiae. Dev. Cell *3*, 815–824.

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

Taylor, D.L., Woo, E.S., and Giuliano, K.A. (2001). Real-time molecular and cellular analysis: the new frontier of drug discovery. Curr. Opin. Biotechnol. *12*, 75–81.

Thorley, J.L., and Page, R.M. (2000). RadCon: phylogenetic tree comparison and consensus. Bioinformatics *16*, 486–487.

Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C.A.-K., Amini, B., Andersen, E., Andersson, A.-C., Angelidou, P., Asplund, A., Cerjan, D., et al. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. Amer. Soc. Biochem. Mol. Biol. *4*, 1920–1932.

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

Wu, J.-Q., Kuhn, J.R., Kovar, D.R., and Pollard, T.D. (2003). Spatial and temporal pathway for assembly and constriction of the contractile ring in fission yeast cytokinesis. Dev. Cell *5*, 723–734.

Yarrow, J., Totsukawa, G., Charras, G., and Mitchison, T. (2005). Screening for cell migration inhibitors via automated microscopy reveals a Rho-kinase inhibitor. Chem. Biol. *12*, 385–395.

Yu, C.S., Chen, Y.C., Lu, C.H., and Hwang, J.K. (2006). Prediction of protein subcellular localization. Proteins *64*, 643–651.

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.

Zhao, T., Soto, S., and Murphy, R.F. (2006). Improved comparison of protein subcellular location patterns. 2006 IEEE Intl. Symp. Biomed. Imaging, 562–565.

Zhou, X., and Wong, S.T.C. (2006). High content cellular imaging for drug development. IEEE Signal Process. Mag. *23*, 170–174.

Zimmer, C., Labruyere, E., Meas-Yedid, V., Guillen, N., and Olivo-Marin, J.-C. (2002). Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing. IEEE Trans. Med. Imaging *21*, 1212–1221.