Automated Unmixing of Complex Protein Subcellular Location Patterns

A Written Report for Machine Learning Data Analysis Project

Student: Tao Peng
Advisor: Robert F. Murphy

Abstract

Many proteins and macromolecules exhibit complex subcellular distributions, which may include localizing in more than one organelle and varying in location depending on the cell physiology. Estimating fraction of fluorescence in each pattern of subcellular location is essential to quantitatively understand protein dynamics such as synthesis, degradation and relocation. Previously the only available method to access such information relies on using colocalization with other fluorescent markers. However, this technique is limited by the probes' specificity, availability and spectral overlap, rendering automated analysis of protein localization difficult with this strategy. We therefore developed both supervised and unsupervised learning methods to quantify the fractions of a fluorescent marker in different organelles in automated fashion. The principle of these methods is representing the protein distribution into distinct object types combined with the assumption that mixed patterns were formed by additive combinations of object types. The supervised method then uses fundamental patterns of organelles to unmix the mixed patterns by linear regression or multinomial inference; while the unsupervised method, which assumes no pre-knowledge of fundamental patterns available, discloses the fundamental patterns from the mixed patterns and fractional composition of each image set with graphical model learning strategy. To test our approaches, cells were tagged by combinations of two organelle specific probes that had the same fluorescent properties and imaged to simulate multiple patterns of subcellular localization. The results indicate that we can unmix the complex subcellular distribution with reasonably high accuracy using both supervised and unsupervised methods.

Keywords: Microscope image analysis; Machine learning; Subcellular location

1 INTRODUCTION

Eukaryotic cells are organized into a number of distinct subcellular compartments and structures that play critical roles in cellular functions. Compartmentalization of proteins plays an important role in regulating pathway activities and modulating cross-talk between regulatory networks. Therefore, protein localization is a tightly regulated process whose failure may lead to severe pathologies [1]. Unfortunately detecting and quantifying the amount of macromolecules in smaller organelles are much harder tasks to perform using traditional techniques, and they are not readily automated. The problem is made more difficult by the fact that proteins (and other macromolecules) are often found in more than one subcellular structure.

The traditional approach to determining the amount of a macromolecule in specific compartments is measuring colocalization of that molecule with organelle-specific markers labeled with a different fluorophore. This approach is often used for one or a few proteins whose possible locations are known a priori, and recent work describing methods for automatically measuring the fraction of colocalization have been described [2, 3]. It is difficult to apply on a proteome-wide basis since it requires the availability of probes for all possible compartments and the collection of separate images of each protein in combination with each marker. An exciting automated approach that builds on the basic colocalization method is the MELK technology, which uses robotics to carry out successive rounds of staining for different macromolecules on a given cell sample or tissue [4]. However, this approach is restricted to fixed samples and cannot be used for analysis of dynamic pattern changes in living cells. Hand-tuned algorithms for distinguishing particular subcellular regions are also widely used as an alternative to colocalization in high content screening [5], but they typically are only able to distinguish major cellular regions and are not easily transferred to the analysis of other regions or cell types.

Beginning with the demonstration that automated recognition of subcellular patterns was feasible [6, 7], the Murphy group and others have created systems that are able to classify all major subcellular location patterns and to do so with a higher accuracy than visual analysis [8, 9, 10, 11]. Automated systems can also learn what subcellular patterns are present in large collections of images without prior knowledge of the possible patterns [10, 12]. However, such pattern clustering approaches have two major limitations. First, they treat each unique combination of major, fundamental patterns as a new pattern since they cannot readily identify the fundamental patterns of which it is composed. Second, they are not designed to handle cases where the fraction of mixing between two patterns can vary continuously, because such continua are either considered as one large pattern or arbitrarily divided into sub-patterns.

To develop tools to quantify the amount of fluorescence in each compartment for images containing a mixture of fundamental patterns, assuming that sets of images containing each fundamental pattern are available. Zhao *et al* have previously proposed an object-based approach [13] to this problem that consisted of two learning stages: learning what object types are present in the fundamental patterns, and learning how many objects are present for each type in each pattern. The fraction of fluorescence in each fundamental pattern for a mixed image was then estimated by determining the mixture coefficients that were most likely to have given rise to that image. This method was tested on synthetic images created from known amounts of many patterns, which permitted the accuracy of unmixing of a given image to be determined by comparison with the mixture coefficients used to synthesize that image. However, the effectiveness of this approach on real images with multiple patterns was not determined due to the lack of availability of real images for which mixture fractions were known. In this study, we used high throughput automated microscopy to create an image dataset for cells labeled with varying mixtures of fluorescent mitochondrial and lysosomal probes containing essentially the same fluorophore. This controlled experiment mimics typical cases such as a protein that distributes among different organelles

or that changes its location upon drug stimulation. Therefore, these data offer the perfect conditions to test our algorithm in the context of automated image acquisition. We were successfully able to unmix the two different patterns and quantify the relative amount of probes in the two organelles. We were also able to show that the method can identify objects and patterns that are distinct from those used for training. This work should therefore open the door to a more comprehensive approach to subcellular pattern analysis for automated microscopy. In part, the strategy described here should help quantify subtle translocation and localization defects that could not be readily measured previously.

These automated unmixing methods were observed to perform well on real data in recovering the underlying mixture coefficients. However, designed in supervised learning fashion, they still require the researcher to specify the fundamental patterns of which the mixed patterns are composed. For example, for the quantitative analysis of translocation experiments as a function of time or drug concentration, the extreme points could be easily identified as the patterns of interest. However, they are still inapplicable to proteome-wide studies where it would be a difficult (and perhaps impossible) task to identify all fundamental patterns that are present. Therefore, it is necessary to tackle the unsupervised pattern unmixing problem: Given a large collection of images, where none has been tagged as being a representative of a fundamental pattern, map all images into a set of mixture coefficients automatically derived from the data. The approaches we have designed to solve it are thought to be unsupervised because not only mixing fractions for each set of images are estimated but also fundamental pattern representations are discovered during the learning process. Using the same test dataset previously created to test supervised unmixing methods, we were able to show the unsupervised unmixing methods perform competitively to supervised methods in quantifying the relative amount of probes in the two organelles. The unsupervised unmixing as an advanced and generalized approach, should be more effective in monitoring protein statics and dynamics quantitatively using microscope imaging and image analysis in fully automated fashion.

2 RESULTS

In this paper, a pattern designates the subcellular distribution of a protein, or of a set of proteins whose distributions are statistically indistinguishable. We define a fundamental pattern as a pattern that cannot be represented as the sum of the patterns of other proteins, while a mixed pattern refers to a distribution consisting of two or more fundamental patterns. A pattern is characterized by a collection of fluorescent objects whose shape, size, and intensity vary within cells. For example, nuclei are typically large ellipsoidal objects while lysosomes are small and generally have spherical shapes. The method we describe here seeks to estimate the components in a given mixed image based on two assumptions. The first is that the set of discrete object types resulting from segmentation of images containing a mixed pattern is essentially the same as the union of the sets of object types found in images of each of its fundamental patterns (i.e., that any new object types that might be found only in mixed images do not contain a significant amount of fluorescence and can be safely ignored). The second is that the amount of fluorescence in each object type in a mixed image is approximately the sum over all fundamental patterns of the product of the fraction of total protein in that pattern and the number of objects of that type in images of that pattern (i.e., that any differences between the actual and expected sums are sufficiently small and uncorrelated with the mixture fractions that they do not systematically affect estimates of the fractions).

The approaches are illustrated in Figure 1 for supervised unmixing and in Figure 2 for unsupervised unmixing. Both approaches are tested for mixtures of two patterns, but they generalize to any number of patterns. The details for each step of this process are described in the section of Method. We assume that we are provided with a collection of images of cells containing varying combinations of fluorescent

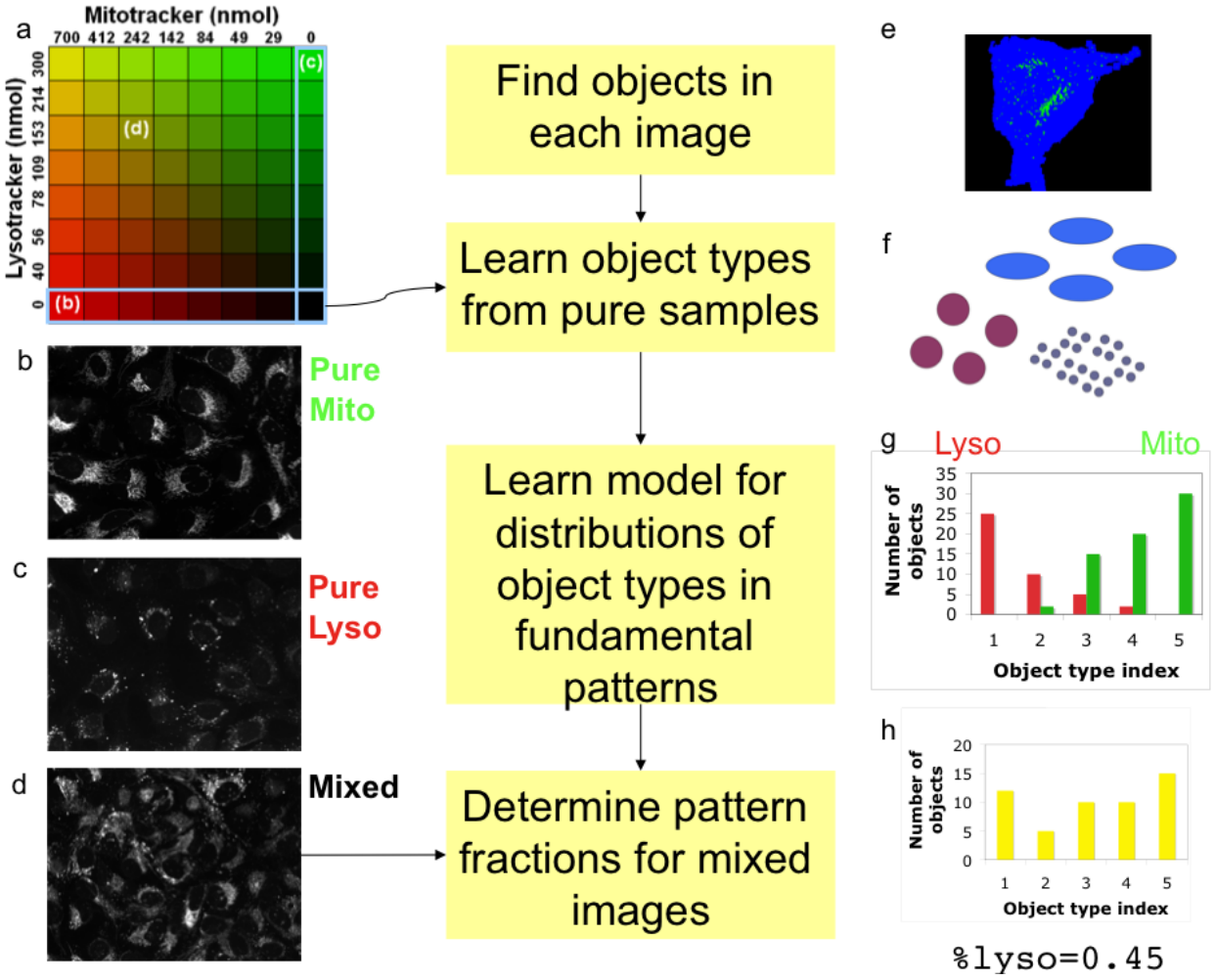

Figure 1: Supervised unmixing approach. (a) The starting point is a collection of images (typically from a multi-well plate) in which various concentrations of two probes are present (the concentrations of the Mitotracker and Lysotracker probes are shown by increasing intensity of green and red, respectively). Example images are shown for wells containing just Lysotracker (b), just Mitotracker (c), or a mixture of the two probes (d). The steps in the analysis process are shown: finding objects (e), learning object types (illustrated schematically as objects with different sizes and shapes), learning the object type distributions for the two fundamental patterns (g), and unmixing a mixed object type distribution (h).

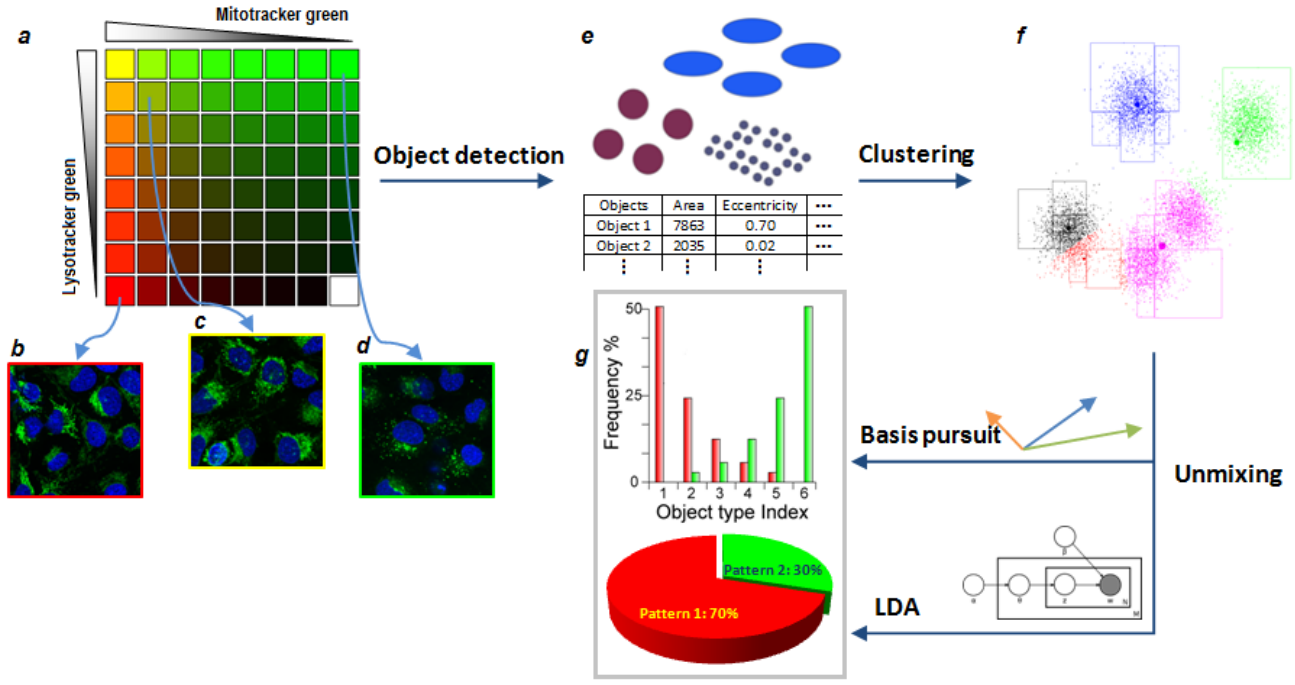

Figure 2: Unsupervised unmixing approach. (a) The algorithms use a collection of images as input in which various concentrations of two probes are present (the concentrations of the Mitotracker and Lysotracker probes are shown by increasing intensity of red and green, respectively). Example images are shown from wells containing only Mitotracker (b), only Lysotracker (c) and a mixture of the two probes (d). (e) Objects with different size and shapes are extracted and object features are calculated. (f) Objects are clustered into groups in feature space, shown with different colors. (g) Fundamental patterns are identified and the fractions they contribute to each image are estimated.

probes such as for the multi-well plate of pure and mixed samples depicted in Figure 1a. The two probes are assumed to be imaged using filters that do not distinguish between the probes. Wells containing pure probes are symbolically represented along the outer row and column (with example images shown in Figure 1b and c), and the other locations represent mixed conditions (a mixed image is shown in Figure 1d). The starting point for unmixing is finding all objects in each image by thresholding (Figure 1e). Each object is described using a set of numerical features that measure characteristics like its size and shape.

Under supervised framework, the two sets of fundamental patterns (pure probes) are then used to learn the types of objects that can be found (Figure 1f). Given this list of object types, the distribution of object types across each fundamental pattern is then learned as a count of the number of objects in each object type (Figure 1g). Given this, the distribution of fluorescence in each object type in a mixed image (Figure 1h) can be used to estimate the fraction of probe in each fundamental type. Comparing with it, under unsupervised settings, all sets of images are used to learn the types of objects and either fundamental patterns or mixed patterns are represented by distributions of object types(Figure 2a-f). The unsupervised unmixing methods then use "bag of words" model to extract the fundamental pattern representations and the fraction of probe in each fundamental type for each set of images with similar fractional compositions(Figure 2g).

To test the performance of the approaches applied to real images, a dataset of mixed patterns is needed in which the mixture fractions are known (at least approximately) so that estimates obtained by unmixing can be compared with expectation. We have therefore constructed such a dataset using high-throughput microscopy[1]. We chose two fluorescent probes (Lysotracker green and Mitotracker green) that stain distinct subcellular compartments (lysosomes and mitochondria) but that contain similar fluorophores so that they can be imaged together. The dataset contains images for cells incubated with each probe separately (at different concentrations) as well as images for cells incubated with mixtures of the probes. We assume that the amount of probe fluorescence in each compartment is proportional to the concentration of that probe added. This represents a good simulation of the images expected for a protein that can be found in varying amounts between two compartments.

2.1 Learning object types

2.1.1 Object extraction and feature calculation

As outlined above, the starting point is to identify each fluorescence-containing object in all images. We use an automatically-chosen global threshold for each image since this approach does not require segmentation of the image into individual cell regions. Each object is then described by a set of eleven features that characterize its size, shape and distance from the nearest nucleus (see Methods). If more than one image (field) is available for a given condition, the objects from all fields are combined.

2.1.2 Object type learning

Having identified the individual objects, we next determine how many types of objects are present. We define an object type as a group of objects with similar characteristics that form a cluster in the feature space. Rather than specifying the object types a priori, we used cluster analysis to learn clusters from the set of all of objects in all of the training images. While many different clustering methods might be used for this step, we have used k-means clustering due to the large number of objects in the training

[1]We acknowledge G. M. C. Bonamy *et al* for performing the experiment, including specimen preparation and imaging. Detail of the experiment is available in [14]

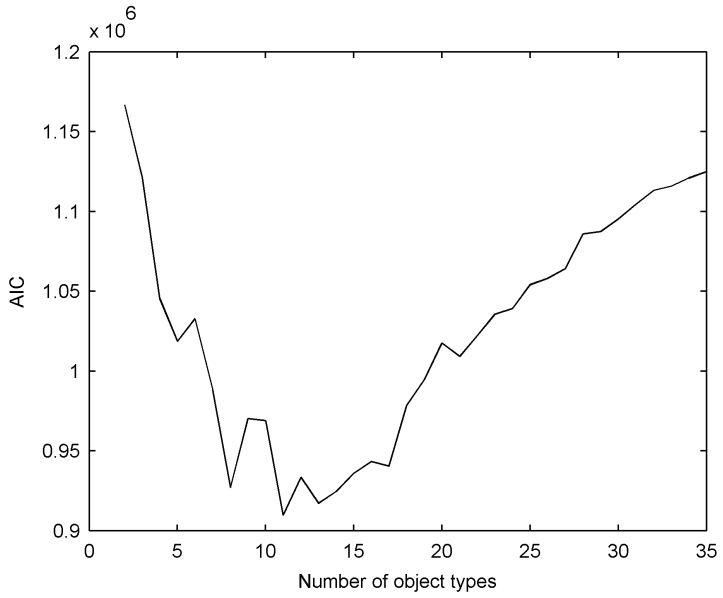

Figure 3: Learning the number of object types k-means clustering was carried out for varying numbers of clusters (k) for all objects in the combination of images receiving only Mitotracker or Lysotracker. The AIC value (which balances the tightness of the clustering against the number of clusters required to achieve it) was then calculated for each clustering. The optimal number of clusters (minimum AIC value) is found to be 11.

set. The optimal number of clusters k was determined by minimizing the Akaike Information Criterion (AIC), which specifies a tradeoff between complexity of the model (number of clusters) and goodness of the model (compactness of the clusters). As shown in Figure 3, the AIC value declines with increasing k until reaching a minimum at a k value of 11 (after which it rises).

2.2 Supervised unmixing

2.2.1 Learning the object composition of fundamental patterns

Once k is known, each fundamental pattern p can be represented as a vector of length k consisting of the frequency of each object type. As shown in Figure 4, the frequency of each object type is quite different between the lysosomal and mitochondrial patterns.

2.2.2 Estimating unmixing fractions

While the type of each object in the training images is known (since all objects in the training images were used for clustering), the type of objects in the testing images is not. Each protein object in a testing image was therefore assigned to the cluster whose center was closest to it in the feature space. The frequency and the total fluorescence of all objects belonging to the same type were calculated for each test image.

The average fraction of mitochondrial and lysosomal patterns in each well was then estimated by three different approaches using both the object types and object features (see Methods). Two of these

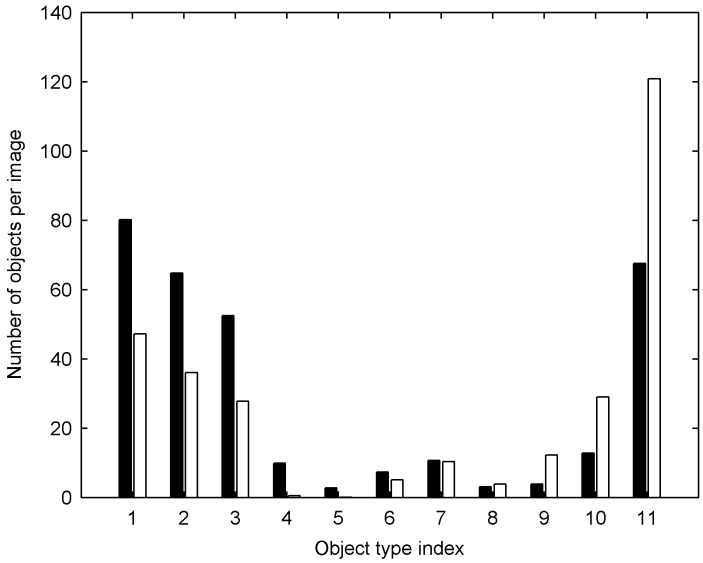

Figure 4: Distribution of object types within fundamental patterns. The average number of objects of each type is shown for the combination of all images stained with either Mitotracker (black) or Lysotracker (white). The object types are sorted according to the difference between the numbers of objects in the two patterns. Thus the lowest numbered object types are primarily found in Mitotracker stained cells, while the highest numbered object types are primarily found in Lysotracker stained cells.

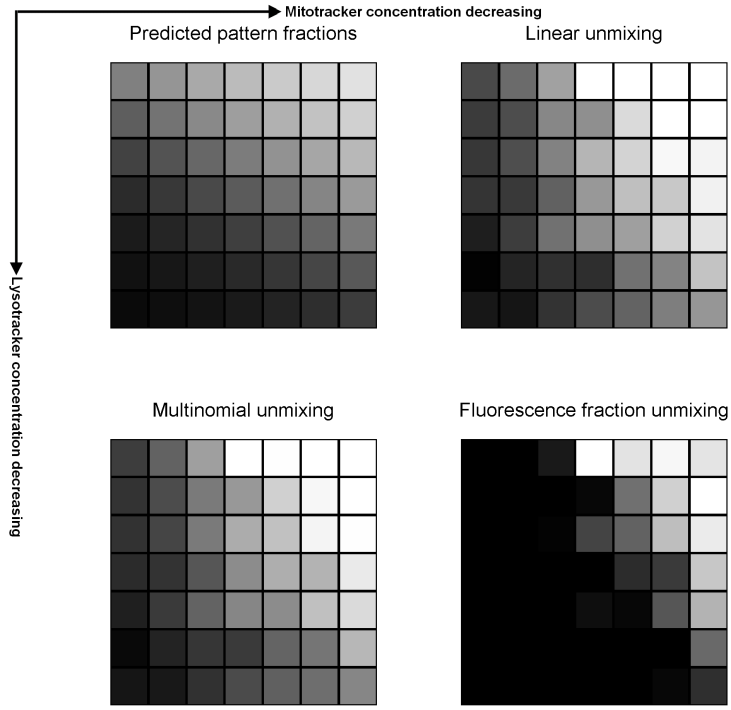

Figure 5: Expected and estimated pattern fraction for three unmixing methods. The balance between lysosomal and mitochondrial pattern (either expected or estimated) is represented as the fraction of the total pattern (black=100% mitochondrial, white=100% lysosomal). For the expected fraction, this is estimated as linearly proportional to the ratio of the relative concentration of the mitochondrial probe to the sum of the relative concentration of the lysosomal and mitochondrial probes (where relative concentration is defined as fraction of the maximum subsaturating concentration).

methods use the number of objects of each type to estimate mixing fractions. The third uses the amount of fluorescence in each type, which depends on the assumption that this amount is linearly dependent upon the concentration of each probe. The results for the three methods are compared with those expected from the relative probe concentrations in Figure5. The correlation coefficients between estimated and expected fractions are 0.71, 0.77 and 0.83.

2.2.3 Removing outliers

The unmixing method can be applied without any restrictions to any test images. However if the mixed patterns in test images contain additional fundamental patterns not present in the training sets, the estimates of mixing fractions will be incorrect. Our solution consists of two steps. The first is to exclude objects in test images that do not appear to belong to any of the object types found in the training images (outlier objects). The second is to flag as outlier patterns those test images that during unmixing have large fitting error (i.e., to reject images that cannot be successfully decomposed into the user-specified fundamental patterns). Since we cannot know a priori what kinds of new objects might be encountered,

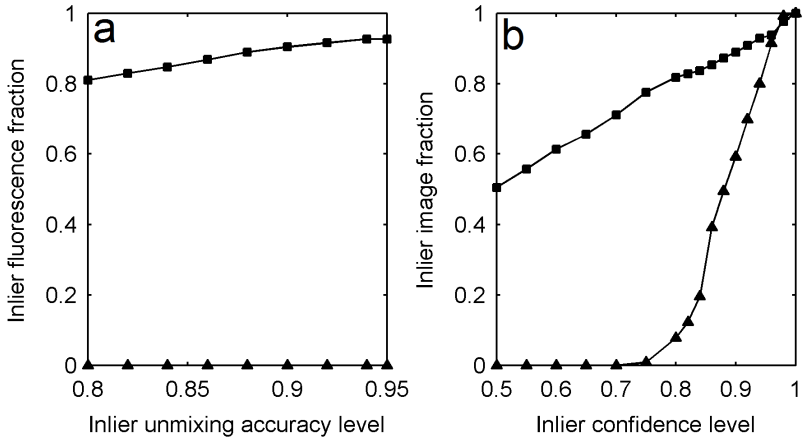

Figure 6: Effectiveness of outlier removal methods. (a) Nuclear images were used as outliers. Unmixing accuracies for both inliers (squares, mitochondrial and lysosomal objects) and outlier objects (triangles) with first-level outlier exclusion were approximated by cross validation under different chosen accuracy levels. Nuclear fluorescence was totally removed at all accuracy levels. (b) ER pattern images were used as outliers. Average outlier recognition testing-accuracies for both inliers (squares, mitochondrial and lysosomal images) and outlier images (triangles) with second-level outlier exclusion were approximated by cross validation under different chosen accuracy levels. The best separation is obtained using a 75-80% inlier confidence level.

both steps use hypothesis-based tests to find thresholds that retain high accuracy for unmixing the training patterns. To test this approach, we used images in which either the nucleus or the endoplasmic reticulum (ER) was marked. Our results show that this methodology can completely remove nuclear objects during the first level outlier detection (Figure 6a). For the more difficult case of ER staining, the second level detection recognizes most of the ER-containing images as outlier patterns but retains high accuracy of fundamental pattern unmixing (Figure 6b).

2.3 Unsupervised unmixing

2.3.1 Latent Dirichlet Allocation

Discovering the object composition of fundamental patterns and estimating the fractions of mixtures are achieved simultaneously. Latent Dirichlet allocation (LDA) which is a popular technique of topic modeling in text analysis was used to solve the unsupervised unmixing problem. Using the LDA approach (see Methods) with arbitrarily assigned number of fundamental patterns $B = 2$, which is the groundtruth, the overall correlation coefficient between estimated and actual pattern fractions was found to be 0.91. The results of LDA unmixing and basis pursuit[2] are compared in Figure 7 and Figure 8.

It is notable that both unsupervised methods led to higher correlation with the underlying coefficients than the supervised methods. A possible cause of this is the appearance of new object types in the mixture patterns. Under the unsupervised framework, with massive clustering, these objects might be assigned labels different from the ones of the fundamental patterns, while in the supervised version they are forced

[2]An approach developed by L. P. Coelho, see [15].

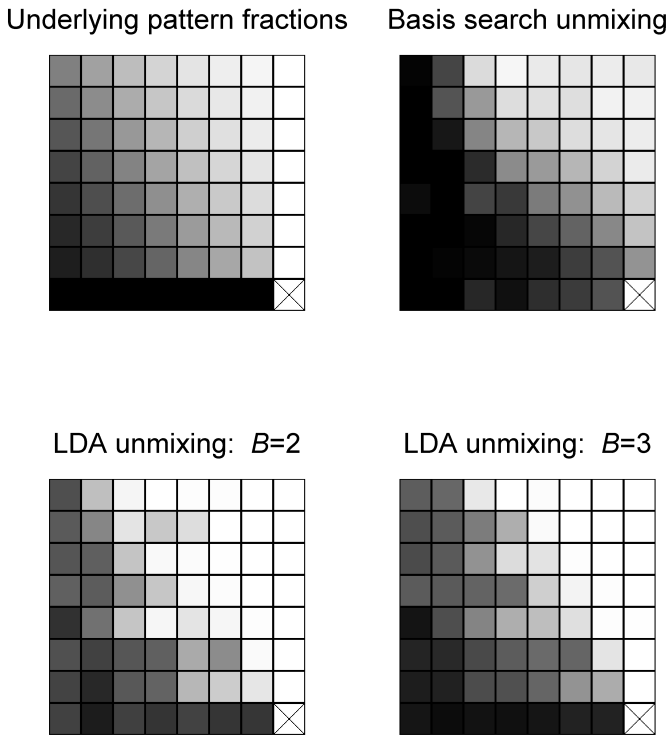

Figure 7: Results for unsupervised unmixing methods. The balance between two patterns is represented as the fraction of the total pattern (black, 100% pattern 1; white, 100% pattern 2). The discovered fundamental patterns were re-indexed to guarantee positive correlations between underlying fractions and unmixing results.

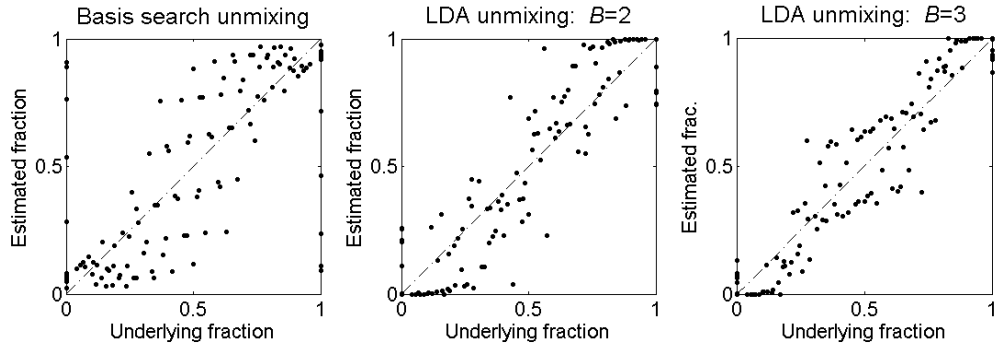

Figure 8: Estimated concentration as a function of the underlying relative probe concentration. Perfect result would be along the dashed diagonal.

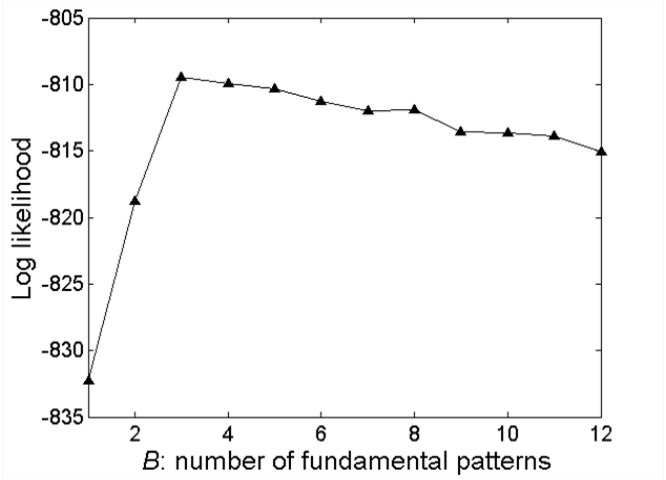

Figure 9: Cross-validated test log-likelihood as a function of the number of fundamental patterns.

	Mitochondrial	Lysosomal	Mixture
Pattern 0	0.0%	0.0%	1.5%
Pattern 1	8.8%	**99.9**%	58.8%
Pattern 2	**91.2**%	0.1%	39.7%

Table 1: Unmixed coefficients for fundamental patterns and mixed samples for the discovered patterns (using LDA method). For the two fundamental patterns and the mixed cases, we display the average coefficient for the 3 discovered fundamental patterns.

to be one of the object types present in the fundamental patterns. To prove this conjecture, we assumed that such new types of objects really exist and applied the outlier removal technique stated in last section to perform supervised unmixing again, in the hope of removing the influence of these objects. The correlations increased to 0.91 and 0.88 with linear and multinomial unmixing approaches, respectively, which are comparable to the unsupervised results.

2.3.2 Determining number of fundamental patterns

To estimate the number of fundamental patterns using the LDA approach, we measured the log likelihood of the dataset for different numbers of bases using cross validation. The results are shown in Figure 9. We can see that the best result is obtained for $B = 3$, while the underlying dataset only has two fundamental patterns.

Table 1 shows the discovered mixing fractions when the method was applied to images of fundamental patterns. Pattern 1 obviously corresponds to the lysosomal component, while pattern 2 corresponds to the mitochondrial component. Pattern 0 appears to be a "non-significant" pattern capturing the new object types arising in the mixture patterns. The overall correlation coefficient is 0.95 with pattern 0 removed.

3 METHODS

3.1 Image Analysis

3.1.1 Preprocessing

Images containing no nuclei, and out of focus images, were removed by thresholding on total Hoechst fluorescence. Shading and skew corrections were performed on remaining images to compensate for non-homogeneous illumination and differences in registration between channels. Background fluorescence was removed by subtracting the most common pixel value from each pixel.

3.1.2 Object detection

An automated threshold method [16] was used to distinguish probe-containing from non-probe-containing pixels. In the resulting binary images, each set of connected above-threshold pixels (an object) was identified. Objects containing less than 5 pixels were ignored. The same approach was applied on the DNA channel to identify DNA objects.

3.1.3 Object feature calculation

To describe the properties of each object, a set of numerical features previously defined as SOF1 (**S**ubcellular **O**bject **F**eatures 1), was calculated (Table 2). This set is composed of nine features based upon morphological properties of the object and two features describing the spatial relationship between objects and the cell nucleus. However, since in the experiments described here images were not segmented into single cells, feature SOF1.2 was replaced by the average distance between each object and the nearest nucleus. All features were normalized to zero mean and unit standard deviation calculated using the training data (Z-scores).

Table 2: List of Subcellular Object Features

Index	Feature description
SOF1.1	Size (in pixels) of the object.
SOF1.2	Distance of object center of fluorescence to DNA center of fluorescence.
SOF1.3	Fraction of object that overlaps with DNA.
SOF1.4	Eccentricity of object hull.
SOF1.5	Euler number of object.
SOF1.6	Shape factor of convex hull.
SOF1.7	Size of object skeleton.
SOF1.8	Fraction of overlap between object convex hull and object.
SOF1.9	Fraction of binary object that is skeleton.
SOF1.10	Fraction of fluorescence contained in skeleton.
SOF1.11	Fraction of binary object that constitutes branch points in the skeleton.

3.1.4 Object type learning

The features for all objects from the singly-stained samples were clustered using the NetLab k-means function. The quality of the clustering was assessed using the AIC as described previously [13]. For test images, objects were assigned to the cluster whose center was the smallest Euclidean distance from it.

3.1.5 Supervised unmixing: linear approach

Once the k object types are defined, each image can be represented as a vector $\mathbf{y} = (y_1, \ldots, y_k)$ of the frequency of each object type in that image. We define u as the total number of fundamental patterns. For $u = 2$, a mixture of pattern 1 (lysosomal) with n_1 objects of a specific object type and pattern 2 (mitochondrial) with n_2 objects of the same object type is assumed to generate a mixed pattern with $n_1 + n_2$ objects of this type. We assume that mixed pattern object frequencies are linear combinations of fundamental pattern object frequencies $\mathbf{f}_p$ as follows:

$$\begin{cases} \mathbf{y} = \sum_{p=1}^{u} \alpha_p \mathbf{f}_p \\ \sum_{p=1}^{u} \alpha_p = 1 \end{cases} \tag{1}$$

where p represents the proportion of fundamental pattern p in the composition of the mixture. Therefore, a mixed pattern can be represented by a vector of coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_u)$, which represents the fraction of fundamental patterns it is composed of. Unmixing the mixture pattern consists of solving the linear equation above. Since we have k equations for all object types and only two fundamental patterns ($u = 2$), a reasonable solution is to minimize the squared error $SE = \sum_{i=1}^{k} (\hat{y}_i - y_i)^2$. This can be achieved using the pseudo-inverse matrix method by solving $\hat{\boldsymbol{\alpha}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{y}$, where $\mathbf{F}$ is the matrix of object frequencies $\mathbf{f}_p$ for all patterns. To avoid negative $\boldsymbol{\alpha}_p$ coefficients since the contribution of a fundamental pattern cannot be negative, fundamental patterns with negative coefficients were set to zero and other coefficients are re-estimated using the same approach. All coefficients were finally normalized to sum to 1.

3.1.6 Supervised unmixing: multinomial unmixing

An alternative method for unmixing is based on the fact that the number of objects of each type varies between cells even within the same pattern (e.g., the number of small lysosomes can vary from cell to cell). We can reasonably assume that if we learn the distribution of the number of objects per cell of a given type for a given fundamental pattern that it will also apply to the distribution of that object type in a mixed sample. In a multinomial distribution, each object belongs in exactly one of the k possible object types with the probabilities $(\theta_1, \ldots, \theta_k)$ (so that $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$). Therefore each fundamental pattern is represented by a multinomial distribution $(\theta_1^p, \ldots, \theta_k^p)$ where θ_i^p is the probability that an object from pattern p belongs to the object type i. This can be estimated by the maximum likelihood estimator of a multinomial distribution:

$$\hat{\theta}_i^p = \frac{n_i^p}{\sum_{i=1}^{k} n_i^p} \tag{2}$$

where n_i^p corresponds to the number of objects of pattern p which are of type i. Mixed patterns are represented by a multinomial distribution composed of a linear combination of fundamental pattern distribution parameters,

$$\boldsymbol{\theta} = \left(\sum_{p=1}^{u} \alpha_p \theta_1^p, \ldots, \sum_{p=1}^{u} \alpha_p \theta_k^p \right) \tag{3}$$

To reach a distribution which best fits the data of mixed pattern conditions, we adjust the coefficients α_p to maximize the likelihood of the object frequency for all k types:

$$\hat{\boldsymbol{\alpha}} = \max_{\boldsymbol{\alpha}} \prod_{i=1}^{k} \left(\sum_{p=1}^{u} \alpha_p \theta_i^p \right) \tag{4}$$

An equivalent problem would be to maximize the log likelihood. We have previously deduced closed forms of the first-order derivative and the Hessian matrix, and proved concaveness of the log likelihood function (16). Therefore, Newton's method was adopted to solve the optimization problem.

3.1.7 Supervised unmixing with fluorescence fraction

In addition to using the number of objects of each object type to estimate pattern fractions, we can also use the amount of fluorescence in each object type. To do this, we first find the average fraction of fluorescence f_i^p within each object type i for each fundamental pattern p. We also determine a constant L_p for each probe that relates the concentration of that probe to the expected amount of total fluorescence in all objects of the pattern p labeled by that probe. We combine these to calculate $\lambda_i^p (= f_i^p \cdot L_k)$ as the amount of fluorescence within each object type that is expected per unit of probe added (assuming that fluorescence is roughly linear over the range of probe concentrations added). The fluorescence expected in each object type F_k is then given by

$$\begin{pmatrix} F_1 \\ \vdots \\ F_k \end{pmatrix} = \sum_{p=1}^{u} C_p \begin{pmatrix} \lambda_1^p \\ \vdots \\ \lambda_k^p \end{pmatrix} \tag{5}$$

where Cp is the concentration of probe that labels pattern p. We used the pseudo-inverse approach to estimate it.

3.1.8 Outlier detection

To address the possibility that a particular pattern being unmixed is not a mixture of the fundamental patterns used during training (that is, that it might contain other patterns as well), we developed a two-level outlier detection method. The unmixing results were expressed as $(\alpha_0, \alpha_1, \ldots, \alpha_u)^T$ where α_0 is the fluorescent fraction of any unrecognized pattern (outliers) and u is the total number of fundamental patterns. Both levels use statistical hypothesis tests to determine outliers. First, we used a chi-square test to remove outlier objects that were not similar to any of the object types learned from the fundamental pattern images. The chi-square statistic is defined as $\chi^2 = \sum_{j=1}^{m} x_j^2$, where x_j is the j^{th} feature of an object and m is the total number of features. Statistics of test object χ_T^2 was tested under every chi-square distribution learned from each object type in the training images to see if it was from that distribution. An object was considered an outlier it was rejected by all tests at a specified confidence level α. Since a proper value for α is hard to determine a *priori*, we chose it by a linear search using unmixing of the fundamental patterns. The fundamental pattern images were split into training and test sets and the accuracy was reported as the fraction of objects that were associated with the correct pattern. For various α, we used cross validation to get averaged accuracies. Accuracy improves with decreasing α cut-offs, in other words, with a stricter criterion, more objects are excluded as outliers. We chose an arbitrary acceptable accuracy level and its associated level α to remove outlier objects in testing images.

Similarly, we performed a hypothesis test to exclude mixed patterns that had large fitting errors when decomposed into fractional combination of fundamental pattern fluorescence fractions. This fitting error statistics was defined as:

$$E = \left\| \mathbf{F} - \sum_{p=1}^{u} \alpha_p \mathbf{\Lambda}^p \right\|, \quad \mathbf{F} = (F_1, \ldots, F_k)^T \text{ and } \mathbf{\Lambda}^p = (\lambda_1^p, \ldots, \lambda_k^p) \tag{6}$$

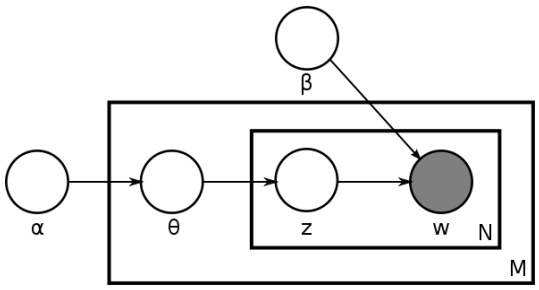

Figure 10: Latent Dirichlet allocation for unmixing. α represents the prior on the topics, θ is the topic mixture parameter (one for each of M images), z represents the particular object topic which is combined with β, the topic distributions to generate an object of type w.

Statistics of a test pattern fluorescence fraction E_T was compared with the empirical distribution of the fitting error. Pattern was rejected to be a mixture of fundamental patterns if E_T was beyond a certain threshold T_2 ($E_T > T_2$). A large T_2 value tolerates more possible real mixture patterns but also risks accepting more unknown patterns. We defined the accuracy of this level as the fraction of training images not rejected. We next learned the empirical distribution of the fitting error on the training set. We then chose T_2 corresponding to an appropriate accuracy level.

3.1.9 Unsupervised unmixing: LDA

Topic modeling in text using latent Dirichlet allocation LDA is a popular technique to solve an analogous class of problems [17]. In this framework, documents are seen as simple "bags of words" and topics are distributions over words. Observed bags of words can be generated by choosing mixture coefficients for topics followed by a generation of words according to: pick a topic from which to generate, then pick a word from that topic.

In our setting, we view object classes as visual words over which to run LDA. This is similar to work by other researchers in computer vision which use keypoints to define visual words [18, 19, 20].

The process of generating objects in images to represent mixtures of multiple fundamental patterns follows the Bayesian network in Figure 10. The generative process is as follows: for each of M images, a mixture $\boldsymbol{\theta}_i$ is first sampled (conditioned on the hyper-parameter α). $\boldsymbol{\theta}_i$ is a vector of fractions of the fundamental pattern distributions $\boldsymbol{b}$. N_i objects are sampled for each image in two steps: select a basis pattern according to $\boldsymbol{\theta}_i$ and then an object is sampled from the corresponding object type distribution.

To invert this generative process, we used the variational-EM algorithm of [17] to estimate the model parameters of fundamental patterns $\boldsymbol{\beta}$ and mixture fractions $\boldsymbol{\theta}$. It should be noted that this is an approximation approach liable to getting trapped in local maxima and returning non-optimal results. Therefore, we ran the algorithm multiple times with different random initializations and chose the one with the highest log-likelihood.

We choose the number of fundamental patterns B to maximize the log-likelihood on a held-out dataset (using cross-validation to obtain more accurate estimate).

4 CONCLUSION

The Murphy group has previously described approaches to cluster proteins based on their subcellular distribution [10, 12]. A logical extension of this work is to create tools to estimate the distribution of fluorescently labeled macromolecules between distinct compartments, and we have previously demonstrated such approaches provide good results for synthetic images [13]. Here we show that this improved approach and the extended unsupervised approach work well on real images obtained from mixed patterns and is suitable for high throughput microscopy, technology that would arguably benefit the most from such a strategy.

Our test mimicked the case of a tagged protein whose distribution varies between two organelles. Because we controlled the amount of both dyes applied to a given cell sample, it was easy to verify whether or not our predictions about the proportion of mitochondrial or lysosomal labeling were accurate. The successful results described here validate the effectiveness of the two-stage, object-based unmixing method on real image data.

The tool we have described requires only a set of images for each of the pure patterns and a set of images for mixed patterns acquired under the same conditions in supervised unmixing settings and only sets of images for mixed pattern with various mixing fractions. Moreover, incorporation of outlier tests and accuracy estimates makes the approach robust to unanticipated phenotypes.

The success of the experiments described here should provide the capacity to better describe what may be complex effects of drugs or disease on protein location. The tool offers a new way to determine a precise and objective subcellular distribution of gene products for various physiological contexts and genetic backgrounds. This approach can also aid large scale projects such as proteome-wide localization studies since it has been tested for images acquired using automated microscopy.

References

[1] J. R. Davis, M. Kakar, and C. S. Lim. Controlling protein compartmentalization to overcome disease. *Pharmaceutical Research*, 24(1):17–27, 2007.

[2] S. V. Costes, D. Daelemans, E. H. Cho, Z. Dobbin, G. Pavlakis, and S. Lockett. Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophysical Journal*, 86(6):3993 – 4003, 2004.

[3] J. W. D. Comeau, S. Costantino, and P. W. Wiseman. A guide to accurate fluorescence microscopy colocalization measurements. *Biophysical journal*, 91(12):4611 – 4622, 2006.

[4] W. Schubert, B. Bonnekoh, A. J. Pommer, L. Philipsen, R. Böckelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A. W. M. Dress. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature Biotechnology*, 24(10):1270–1278, October 2006.

[5] R. T. Dunlay, W. J. Czekalski, and M. A. Collins. Overview of informatics for high content screening. *Methods in Molecular Biology*, 356:269–280, 2006.

[6] M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33(3):366–375, 1998.

[7] R. F. Murphy, M. V. Boland, and M. Velliste. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, pages 251–259, 2000.

[8] M. V. Boland and R. F. Murphy. A neural network classifier capable of recongnizing all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17:1213–1223, 2001.

[9] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lörch, J. Ellenberg, R. Pepperkok, and R. Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res*, 14(6):1130–1136, June 2004.

[10] E. Glory and R. F. Murphy. Automated subcellular location determination and high throughput microscopy. *Developmental Cell*, 12:7–16, 2007.

[11] N. Hamilton, R. Pantelic, K. Hanson, and R. Teasdale. Fast automated cell phenotype image classification. *BMC Bioinformatics*, 8(1):110, 2007.

[12] X. Chen, M. Velliste, and R. F. Murphy. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytometry Part A*, 69A(7):631–640, 2006.

[13] T. Zhao, M. Velliste, M. V. Boland, and R. F. Murphy. Object type recognition for automated analysis of protein subcellular location. *IEEE Trans. Image Proc*, 14:1351–1359, 2005.

[14] T. Peng, G. M.C. Bonamy, E. Glory, D. R. Rines, S. K. Chanda, and R. F. Murphy. Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 107:2944–2949, 2010.

[15] L. P. Coelho, T. Peng, and R. F. Murphy. Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics*, 26:i7–i12, 2010.

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

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[18] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[19] Long Zhu, Yuanhao Chen, and Alan Yuille. Unsupervised learning of Probabilistic Grammar-Markov Models for object categories. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):114–28, January 2009.

[20] J. Sivic, A. Zisserman, and J. Philbin. Geometric lda: A generative model for particular object discovery. In *Proceedings of the British Machine Vision Conference*, 2008.