research articles Journal of proteome research

A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images

Justin Newberg[†,§] **and Robert F. Murphy**[*,†,‡,§,||]

Center for Bioimage Informatics, and Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15217

Received November 16, 2007

The systematic study of subcellular location patterns is required to fully characterize the human proteome, as subcellular location provides critical context necessary for understanding a protein's function. The analysis of tens of thousands of expressed proteins for the many cell types and cellular conditions under which they may be found creates a need for automated subcellular pattern analysis. We therefore describe the application of automated methods, previously developed and validated by our laboratory on fluorescence micrographs of cultured cell lines, to analyze subcellular patterns in tissue images from the Human Protein Atlas. The Atlas currently contains images of over 3000 protein patterns in various human tissues obtained using immunohistochemistry. We chose a 16 protein subset from the Atlas that reflects the major classes of subcellular location. We then separated DNA and protein staining in the images, extracted various features from each image, and trained a support vector machine classifier to recognize the protein patterns. Our results show that our system can distinguish the patterns with 83% accuracy in 45 different tissues, and when only the most confident classifications are considered, this rises to 97%. These results are encouraging given that the tissues contain many different cell types organized in different manners, and that the Atlas images are of moderate resolution. The approach described is an important starting point for automatically assigning subcellular locations on a proteome-wide basis for collections of tissue images such as the Atlas.

Keywords: location proteomics • immunohistochemistry • spectral unmixing • subcellular location • pattern recognition • machine learning • tissue microarrays

Introduction

Genomics and proteomics have provided important stepping stones in modeling cells at the systems level. However, more comprehensive and descriptive data are needed to drive future models, and many current approaches need to be amended to process these data for incorporation into systems models. Historically, one essential tool for biological and medical studies has been microscopy, which is appealing because it can capture spatial information under various conditions. Unfortunately, most studies utilizing microscopy methods focus on a specific task, and often significant amounts of information are lost in such studies. Since proteins localize to various organelles, location is important in characterizing a cell's proteome. The field of location proteomics is therefore concerned with the critical need to capture informative and defining characteristics of subcellular patterns on a proteome-wide basis.[1–3]

Over the past decade, methods for the systematic study of protein locations in cultured cells have been developed that combine fluorescence microscopy with pattern recognition and machine learning techniques.[4,5] These methods involve extracting subcellular location features (SLFs) from images.[6] SLFs are numerical characteristics that quantitatively describe subcellular distributions, and include morphological features, Zernike moment features, Haralick texture features, and wavelet features. Once extracted from the images, SLFs can be used for many purposes, including the training of classifiers to distinguish between protein patterns. When a classifier is able to perform this task well, it indicates that the features are informative.

Extensive demonstration of the feasibility of automated subcellular pattern classification has been carried out on a publicly available data set of fluorescence microscope images for proteins that localize to various different organelles in HeLa cells.[6] Over the past seven years, classification accuracies for this standard data set have been improved through the inclusion of new features and the use of different classification schemes.[7–9] An average accuracy of over 95% is now possible, which compares favorably to the 83% accuracy from a trained human observer.[4]

Images of cell cultures or tissues typically contain many cells in close proximity. Thus, segmenting such images into single cell regions can be an important step in subcellular pattern recognition, although such segmentation is well-recognized as

[*] To whom correspondence should be addressed. Tel: (412) 268-3480. E-mail: murphy@cmu.edu.
[†] Center for Bioimage Informatics, Carnegie Mellon University.
[‡] Department of Biological Sciences, Carnegie Mellon University.
[§] Department of Biomedical Engineering, Carnegie Mellon University.
[||] Department of Machine Learning, Carnegie Mellon University.

10.1021/pr7007626 CCC: $40.75

Published on Web 04/25/2008

a challenging task.[10,11] Alternatively, pattern recognition can be extended across multiple cell fields. The first work to test this latter approach was performed on a synthesized data set, in which 2−6 single cell regions displaying the same protein pattern were combined to form multicell image fields. Fields were synthesized from images in the aforementioned 10-class data set. From the synthetic images, field level features were extracted. These features were designed to be invariant to cell orientation and number. A classification accuracy of 95% was obtained with this method, demonstrating that the features effectively capture the multicell protein patterns.[12]

With the effectiveness of subcellular pattern recognition demonstrated, an important next step is to apply these methods to tissue images. However, creating a comprehensive set of protein images in tissue is challenging. This has in part been addressed by tissue microarray technology (TMA), in which multiple histological sections are used to make many slides, and each slide is immunohistochemically stained for a different protein.[13]

TMA has been used to build the Human Protein Atlas (HPA), an online database containing images of over 3000 proteins across various healthy and cancerous human tissues.[14,15] Analysis of HPA images to date has mainly consisted of determination of the level of staining for each protein in different cell types. Because of its size, the Atlas is an appealing and useful data source for location proteomics studies. Also, since the images are histological sections, there are many potential clinical benefits to automated analysis of the HPA.

In this work, we adapt methods for subcellular pattern recognition that have been extensively applied to fluorescence micrographs of cultured cells to images from the Atlas. HPA images are brightfield micrographs with two mixed stains per image, whereas fluorescence micrographs such as those used previously contain distinct fluorescence channels that reflect the signal from the protein of interest and from a parallel DNA-binding probe. We demonstrate that the HPA images can be processed so that they are suitable for automated analysis, and that excellent recognition of basic subcellular patterns can be achieved across a wide range of tissues. This is significant because demonstration of the efficacy of the method can subsequently enable an automated and thorough investigation of the subcellular patterns in the HPA.

Methods

Data Set. Images from the Human Protein Atlas (http://proteinatlas.org) were used in this study. The images were obtained using immunohistochemical staining of tissue microarrays.[14] Each slide containing a tissue microarray was incubated with a monospecific antibody (an antibody demonstrated to react with only a single protein), followed by washing and incubation with secondary antibody conjugated to horseradish-peroxidase. After washing, each slide was stained with hematoxylin, which nonspecifically stains DNA in cell nuclei a purplish color, and diaminobenzidine (DAB), which is converted to a brown precipitate in regions containing horseradish peroxidase. Brightfield images depicting the circular tissue sections 0.6−1.5 mm in diameter were obtained for 45 normal human tissue types using an RGB camera. The Atlas contains anywhere from 0−6 (typically 3) images for each protein in each tissue type. The images are roughly 3000 × 3000 pixels in size, with each pixel approximately representing a 0.5 × 0.5 μm region in the sample plane. Images are stored as compressed JPEGs.

A set of 16 proteins that localize specifically to one of eight major organelles was chosen from the HPA. Atlas proteins were selected based on locations specified by their UniProt-associated Gene Ontologies (GO) and on how well these GO terms matched with comments in the Atlas regarding the specificity of organelle staining. The proteins, with their Atlas antibody identification numbers and their organelle class, are the following: Sterol-4-alpha-carboxylate 3-dehydrogenase (248, endoplasmic reticulum/ER), Dyskerin (447, nucleolus), Golgin-84 (992, Golgi apparatus), Mitochondrial matrix protein P1 (1523, mitochondrion), Golgi phosphoprotein 4 (1677, Golgi), Cathepsin F precursor (2141, lysosome), Zinc finger protein 146 (3358, nucleolus), GlcNAc-1-phosphotransferase subunit gamma (4055, lysosome), Actin-related protein 2/3 complex subunit 1A (4334, cytoskeleton), AFG3-like protein 2 (4479, mitochondrion), DNA replication licensing factor MCM4 (4497, nucleus), the 78 kDa glucose-regulated protein precursor (5221, ER), Endosome-associated protein p162 (5861, endosome), Bcl-2-associated transcription factor 1 (6669, nucleus), Cytokeratin-9 (7261, cytoskeleton), and Cathepsin E precursor (8021, endosome).

Linear Spectral Unmixing. The RGB images contain a mixture of purplish hematoxylin staining and brownish DAB staining. Since our goal was to analyze the subcellular pattern of each specific protein, some form of color unmixing had to be performed to separate the DAB staining from the hematoxylin staining before we could apply automated recognition methods. The source images contain m-by-n-by-c pixels, where the number of colors c is 3 for RGB images. These can be reshaped into a $(m \times n)$-by-c matrix $\bar{V}$ and then used to form a matrix V that contains only the unique rows of $\bar{V}$. Given a color-bases matrix W with dimensions c-by-r, where r is 2 since there are two source stains to be separated (DNA and protein), we seek to find a matrix $\bar{H}$ of the same dimensions as V that represents the solution of:

$$V = W \times H \tag{1}$$

As long as the unique rows are indexed, H can be easily mapped back to $\bar{H}$, the unmixed image.

Simple linear unmixing works under the assumption that W is known for each image and the signals are linearly separable, and it has been applied to various color separation problems.[16,17] Because of its simplicity, and the fact the same types of stains are used for all of the Atlas images, we used linear unmixing as a starting point after first determining W for all of the images. First, the brightfield images were inverted to make the background black (so low concentrations of protein correspond to low pixel values). The color-basis vectors of the stains were determined by converting a set of HPA images into the HSV color space and then creating a histogram of the resulting image hues. Two peaks in the histogram were identified by the bins with most counts above and below a hue threshold of 0.3. The locations of the peaks were used to define hues of DAB and hematoxylin. Since each hue corresponds to various different sets of saturation and brightness values, we calculated the corresponding saturation and brightness values and then mapped these HSV coordinates back into the RGB space to define the stain color coefficients, k, which signify the relative amounts of red, green, and blue inherent in each stain. The stain color-vectors were used to make the color-bases matrix, W:

$$W = \begin{bmatrix} k_1^{Hem} & k_1^{DAB} \\ k_2^{Hem} & k_2^{DAB} \\ k_3^{Hem} & k_3^{DAB} \end{bmatrix} = \begin{bmatrix} 57.1 & 43.7 \\ 56.5 & 55.6 \\ 42.5 & 64.2 \end{bmatrix} \qquad (2)$$

After the stain color-bases were defined, we linearly unmixed the data using the transpose of the Moore-Penrose pseudoinverse (pinv) of W:

$$H = V \times \mathrm{pinv}(W)' \qquad (3)$$

Finally, each channel of H was scaled so the minimum intensity value was 0 and the maximum value was 255, and H was used to generate the linearly unmixed protein and DNA channels.

Blind Spectral Unmixing by Non-Negative Matrix Factorization. Because of experimental variance, the spectra of immunocytochemical dyes are often not consistent across every image, rendering simple linear unmixing potentially unsuitable. Therefore, we also evaluated non-negative matrix factorization (NMF) as an approach to blindly unmix the images. NMF works under the constraint that each stain contributes non-negatively to the overall image intensity. This method has been shown to be effective in unmixing brightfield images with two stains.[18] Other blind unmixing methods that correct for measurement noise exist.[19] However, in this work, these are not considered, as the source images contain compression artifacts that complicate the modeling of such noise.

The image preprocessing steps for the blind unmixing were similar to linear unmixing. However, instead of calculating a common W for all images, we calculated a unique W for each image sample. We then randomly initialized H, and applied NMF to solve for the non-negative matrix factors W and H by iteratively minimizing the L2 distance between V and $W \times H$.[20] The postprocessing of H into a new unmixed image was similar to the linear unmixing, in which each channel was scaled and then remapped into the image data channels.

Note that performing blind unmixing on V and not $\bar{V}$ is beneficial for two reasons. First, less data is unmixed, meaning the overall computation time decreases. More importantly, by reducing redundant pixels, the unmixing process is less biased by the relative levels of staining. For example, if there is much more purple than brown in the image, unmixing on V^* might separate one hue of purple from another, rather than separate purple from brown. By using V, there need be no concern over the relative amounts of staining (so long as both brown and purple are present in the image).

Data Set Partitioning. Some of the Atlas images contain a third stain. To remove these images from our data set, we performed linear unmixing with a third color-basis on all of the images. From the resulting images, we measured the average intensity of the signal in the third channel and removed images with a value greater than some threshold. Next, for each location class, we removed tissues that had less than two images. Finally, we determined the number of remaining images for each protein by tissue. We then placed alternating occurrences of these protein images into the training and sets. If two of a particular protein in a particular tissue were put in the training set, and only one was put in the testing set, the next time an imbalance was encountered, the testing set received the additional image. This partitioning approach guarantees that either (1) the training and testing sets will be the same size in the event of an even number of image samples, or (2) the training set will contain only one more sample than the testing set in the event that there are an odd number of image samples.

Subcellular Location Features (SLFs). Field level features were extracted from the whole tissue images. Because the images contain many different cells, the SLFs must be robust to cell rotations and translations, as well as the number of cells in the tissue. Haralick texture features were used for this purpose.[12] Since wavelets have been shown to be effective at subcellular pattern recognition,[8] multiresolution (MR) texture features were also used. Finally, since there is nuclear information for each protein image, DNA features that relate protein and nuclear object overlap and distance were extracted.

First, the most common pixel value was subtracted from each image to remove background. To calculate the DNA overlap

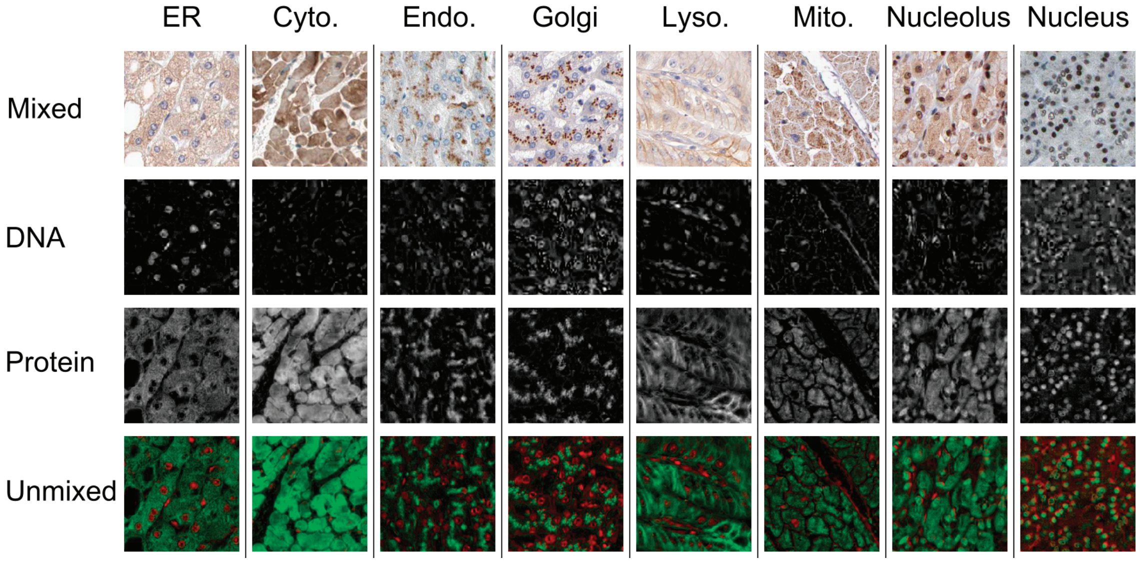

Figure 1. Example image regions from the HPA and results from linear spectral unmixing. These tissues are immunohistochemically stained for various different proteins which localize to the organelles denoted by the column titles. The top row shows unprocessed, mixed Atlas images, while the remaining rows show results from linear unmixing. The second row shows the DNA staining channel, while the third row shows the protein channel. The bottom row is a composite of the two separate channels, with DNA shown in red and protein in green. Images in rows 2−4 have been background-subtracted and contrast-stretched prior to cropping the segments.

Table 1. Features Selected by Stepwise Discriminant Analysis Using the Training Set[a]

rank	feature number	feature description	decomp. level	coeff. type	direction
1	2	Nuclear colocalization (area)	-	-	-
2	613	Correlation	8	H	0, 90
3	10	Sum average	-	-	0, 90
4	177	Angular second moment	2	D	-
5	571	Correlation	7	D	-
6	665	Correlation	8	D	-
7	17	Information measure of correlation 2	-	-	0, 90
8	596	Energy	7	V	-
9	837	Information measure of correlation 2	10	D	45, 135
10	27	Difference variance	-	-	0, 90
11	14	Difference variance	-	-	0, 90
12	7	Correlation	-	-	-
13	827	Correlation	10	D	0, 90
14	595	Energy	7	H	45, 135
15	600	Correlation	8	H	45, 135
16	720	Correlation	9	V	-
17	28	Difference entropy	-	-	0, 90
18	4	Avg. dist. between protein and DNA objects	-	-	0, 90
19	694	Correlation	9	H	-
20	775	Correlation	10	H	45, 135
21	814	Correlation	10	D	0, 90
22	716	Information measure of correlation 1	9	V	45, 135
23	609	Information measure of correlation 1	8	H	45, 135
24	486	Information measure of correlation 1	6	V	45, 135
25	717	Information measure of correlation 2	9	V	-
26	757	Energy	9	H	45, 135
27	839	Energy	10	V	0, 90
28	788	Correlation	10	V	-
29	597	Energy	7	D	0, 90
30	677	Energy	8	V	-
31	676	Energy	8	H	0, 90
32	758	Energy	9	V	45, 135
33	3	Nuclear colocalization (intensity)	-	-	-
34	707	Correlation	9	V	45, 135
35	8	Sum of squares	-	-	45, 135
36	659	Difference variance	8	D	45, 135
37	838	Energy	10	H	0, 90
38	681	Correlation	9	H	45, 135
39	756	Information measure of correlation 2	9	D	45, 135
40	762	Correlation	10	H	0, 90
41	543	Angular second moment	7	V	-
42	759	Energy	9	D	45, 135
43	670	Sum entropy	8	D	45, 135
44	79	Difference variance	1	V	-
45	790	Inverse difference moment	10	V	-
46	546	Sum of squares	7	V	45, 135
47	15	Difference entropy	-	-	-
48	633	Difference variance	8	V	0, 90
49	26	Entropy	-	-	0, 90
50	532	Correlation	7	H	-
51	733	Correlation	9	D	-
52	840	Energy	10	D	-
53	646	Difference variance	8	V	0, 90
54	553	Difference entropy	7	V	-
55	657	Sum entropy	8	D	45, 135
56	652	Correlation	8	D	0, 90
57	723	Sum average	9	V	0, 90

[a] This was done on a feature set originally consisting of the four DNA overlap features and 836 multiresolution texture features calculated using the Daubechies 8 filter.

features, we thresholded the protein and nuclear image channels using Otsu's method. On the resulting binary images we calculated the following four features:

1. The ratio of the area occupied by protein to that occupied by DNA.
2. The fraction of above threshold protein area that colocalizes with above threshold DNA.
3. The fraction of the protein fluorescence that colocalizes with DNA (SLF2.22).[6]
4. The average distance (in pixels) between above threshold protein pixels and the nearest above threshold nuclear pixel.

Next, we scaled the background-subtracted protein images to 32 gray-levels and decomposed them down to 10 levels by the discrete Wavelet Transform using the Daubechies 8 filter. Twenty-six Haralick texture features were calculated—13 from averaged horizontal and vertical co-occurrences and 13 from averaged diagonal co-occurrences[8,21]—and the total energy was calculated on each of the three sets of detail coefficients at each level of decomposition (giving 27 × 3 × 10 = 810 features). Twenty-six Haralick features were also calculated on the original image. Combining all of the texture features with the four DNA overlap features gave a total of 840 features per image.

For the images in the training set, Stepwise Discriminant Analysis (SDA) was used to select the most discriminating features across the eight location classes.[9] In a comparison of

Table 2. Confusion Matrix for Classification of Linearly Unmixed Images Using a Classifier Based on a Single Wavelet Basis Function[a]

	output of classifier							
	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus
ER (131)	**68.7**	11.5	6.1	3.8	5.3	2.3	2.3	0
Cyto (125)	16	**52.8**	7.2	0	9.6	10.4	2.4	1.6
Endo (111)	9	14.4	**61.3**	0.9	2.7	3.6	0.9	7.2
Golgi (126)	8.7	0.8	0.8	**79.4**	1.6	3.2	4	1.6
Lyso (127)	7.1	13.4	6.3	5.5	**63**	2.4	1.6	0.8
Mito (125)	4.8	8.8	0.8	3.2	2.4	**76**	2.4	1.6
Nucleolus (120)	0.8	0	0	9.2	5.8	2.5	**81.7**	0
Nucleus (117)	0.9	2.6	8.5	0	0	2.6	0	**85.5**

[a] Agreement between the true class and classifier output is shown in bold. The number of samples per testing class is shown beneath the class names. The overall accuracy for the single classifier is 71%, as 697 out of 982 testing images are classified correctly. Because of rounding error, rows may not sum to 100%.

Table 3. Confusion Matrix for Classification of Linearly Unmixed Images with Voting Across 10 Classifiers Using Different Wavelet Bases Function[a]

	output of classifier							
	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus
ER (131)	**83.2**	7.6	3.1	1.5	2.3	0.8	1.5	0
Cyto (125)	14.4	**64**	3.2	0	10.4	7.2	0	0.8
Endo (111)	8.1	9.9	**75.7**	0	2.7	0	0	3.6
Golgi (126)	1.6	0	0	**87.3**	1.6	0	9.5	0
Lyso (127)	3.9	9.4	1.6	7.9	**75.6**	0	0.8	0.8
Mito (125)	3.2	4	0	3.2	0.8	**85.6**	1.6	1.6
Nucleolus (120)	0.8	0	0	5.8	4.2	1.7	**87.5**	0
Nucleus (117)	0	0.9	8.5	1.7	0	0.9	0	**88**

[a] Agreement between the true class and classifier output is shown in bold. The number of samples per testing class is shown beneath the class names. The overall accuracy is 81%, as 794 out of 982 testing images are classified correctly. Because of rounding error, rows may not sum to 100%.

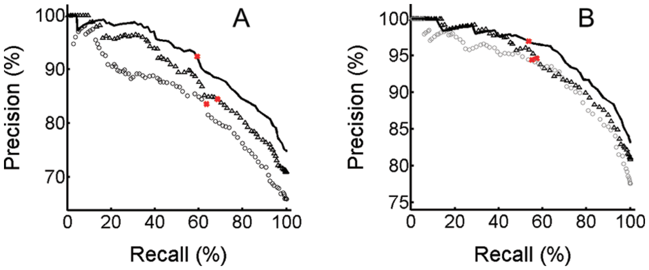

Figure 2. Tradeoff between recall and precision for different automated classifiers and unmixing methods. Classifiers based on a single wavelet basis function (A) perform less well than their voting counterparts (B). Shown are classifiers trained on linearly unmixed data (Δ), blindly unmixed data (○), and both linearly and blindly unmixed data (—). Systems trained with both types of unmixed data consistently perform better in the high recall regime. Recall is defined as the number of images correctly classified with a probability above a threshold divided by the total number of images correctly classified. Precision is the ratio of correctly classified images to the images with labels with probabilities above some threshold, and it is equivalent to classifier accuracy. When only samples with label probabilities/likelihoods greater than 0.5 are considered in the linear, blind, and combined unmixing sets, the simple classifiers have 85%, 84%, and 92% precision, respectively (red x), while the voting classifiers give precisions of 95%, 94%, and 97% (red x).

eight methods for feature reduction, SDA was previously observed to perform best for subcellular pattern classification.[9]

Simple Classification. A linear support vector machine (SVM) classifier was trained on the training set after the SVM slack penalty was determined by parameter variation with 10-fold cross validation (using the LIBSVM toolbox from http://www.csie.ntu.edu.tw/~cjlin/libsvm). Output of classification for the testing set was used to determine overall classification accuracy. For each testing sample, the classifier outputs a probability that a sample belongs to each class. These posterior probabilities are determined using a modified version of Platt scaling.[22] The sample was labeled according to which class had the highest probability.

Classification with Voting. To create additional classifiers, additional sets of features were extracted using different Daubechies filters (with lengths from 2−20 in increments of 2). For features of each filter length, a classifier was trained as detailed above, giving 10 separate SVMs. For each image sample, the probability outputs of the classifiers were summed and divided by the number of classifiers, and the class corresponding to the resulting highest likelihood was assigned as the sample label. Unlike a prior multiresolution classification scheme, which incorporates classifiers with MR voting,[8] this method uses separate classifiers trained on multiresolution features with a simple voting scheme.

Software. This work was done in Matlab 7.1 in Linux with the LIBSVM library, version 2.84. The code and data in this paper are available at http://murphylab.web.cmu.edu/software/.

Results

From the Atlas, we selected a set of 16 proteins which localize to major specific organelle classes (see Methods). These proteins were chosen so we could create classifiers that distinguish between eight location types: the ER, cytoskeleton, endosome, Golgi apparatus, lysosome, mitochondrion, nucleolus, and nucleus (two proteins were included for each location type). After the partitioning of data into training and testing sets, there are 122.8 ± 6.4 samples per class in the training set and 122.8 ± 6.4 samples per class in the testing set. Moreover, each of the 45 tissues is represented by 21.8 ± 6.6 and 21.8 ± 6.3 samples in the training and testing sets, respectively. Finally, each of the 16 proteins is represented by 61.4 ± 5.5 and 61.4 ± 6.4 images, again in the training and testing sets. Example images are shown in Figure 1.

Simple Classification Test. To determine whether automated analysis of the HPA images is feasible, we linearly unmixed the images as described in Methods. We extracted various features from the resulting protein and nuclear channels from each

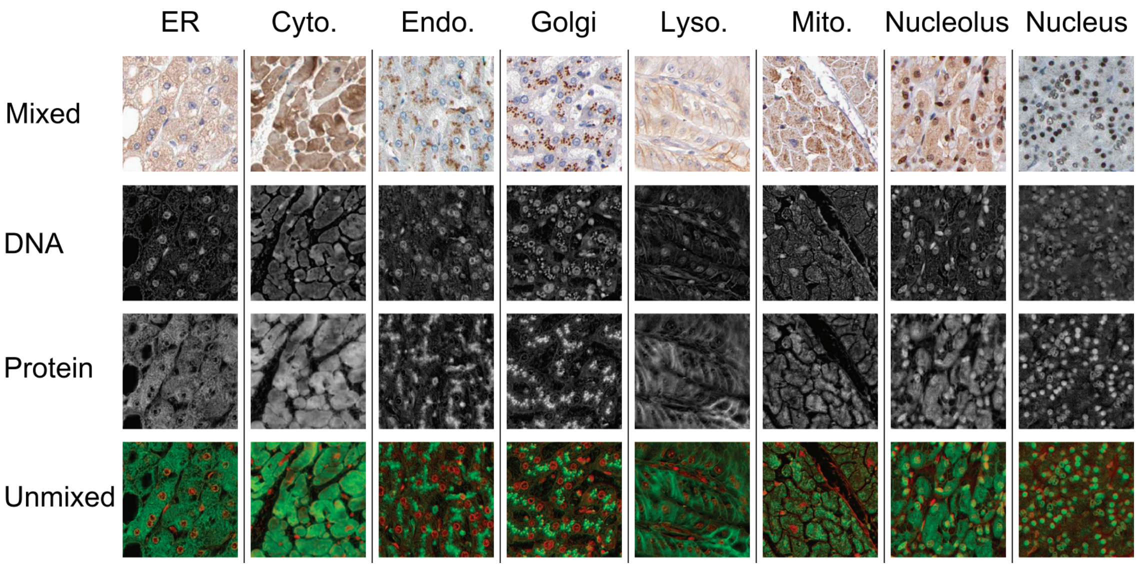

Figure 3. Example unmixing results from blind spectral unmixing. The mixed images (top row) are the same as in Figure 1. The second and third rows show the separated DNA and protein channels, respectively, and the fourth row shows the composite of the two channels, with DNA shown in red and protein in green. Note that the bottom row shows more signal overlap between the two channels than in the images from linear unmixing (Figure 1).

Table 4. Confusion Matrix for Classification of Images with Using Both Linearly and Blindly Unmixed Data When Only High Likelihood Labels Are Considered[a]

	output of classifier							
	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus
ER (53)	**100**	0	0	0	0	0	0	0
Cyto (21)	4.8	**76.2**	0	0	14.3	4.8	0	0
Endo (2)	0	0	**100**	0	0	0	0	0
Golgi (88)	1.1	0	0	**98.9**	0	0	0	0
Lyso (52)	0	1.9	0	0	**96.2**	0	1.9	0
Mito (64)	0	0	0	0	0	**98.4**	1.6	0
Nucleolus (94)	0	0	0	2.1	2.1	1.1	**94.7**	0
Nucleus (78)	0	0	0	0	0	0	0	**100**

[a] Outputs from the two previously-described voting classifiers were summed, and labels with the highest likelihood were assigned to the samples. Then, only images with labels with greater than 0.5 likelihood were considered. Agreement between the true class and classifier output is shown in bold. The number of samples per testing class is shown beneath the class names. The overall accuracy is 97%, as 438 out of the 452 testing images (that have high confidence) are classified correctly. Because of rounding error, rows may not sum to 100%.

unmixed image sample, and then selected informative features using SDA (Table 1). Many correlation and energy features were selected, and features were chosen across various different resolutions.

Next, we trained and tested a classifier using these selected features. The resulting classification accuracy is 71% (Table 2), with an estimated 95% confidence interval of 2.8%. Class accuracies range from 53% (for cytoskeletal patterns) to 86% (for nuclear patterns). Cytoplasmic patterns are highly confused with each other, while nuclear patterns are better distinguished. Surprisingly, the nuclear pattern is still frequently mislabeled as an endosomal pattern. When only samples whose labels have a probability above some threshold are considered, the precision increases (Figure 2A). (Precision is defined as the number of correct label assignments considered divided by the total number of considered assignments, and it is equivalent to classifier accuracy.) Recall drops as fewer images are considered. (Recall is defined as the correct number of considered labels divided by the maximum number of correct labels.) Considering samples that have classification probability greater than 0.5 boosts the classification accuracy to 85%.

Voting Across Classifiers. Voting schemes have been shown to be effective at increasing classification accuracy in subcellular pattern recognition.[8,23] To see how voting across classifiers trained on different features affects classification, we used features calculated from nine different wavelet bases function to train nine additional classifiers. Each of our 10 classifiers reports for a test sample the probabilities that the sample belongs to any of the 10 classes. Summing these probabilities across the 10 classifiers, we chose the label with the highest likelihood as the assignment for a sample. Using this scheme, we obtained an overall classification accuracy of 81% (Table 3). Class accuracies range from 64% (cytoskeleton) to 88% (nucleus). While there is still significant confusion between cytoplasmic patterns, all of the locations are better distinguished under this voting scheme. Considering samples that have classification likelihood greater than 0.5 boosts the classification accuracy to 95% (Figure 2B).

Evaluation of Unmixing Methods. Having determined that this classification approach can recognize eight organelle patterns in the HPA images, we used it to compare linear and blind unmixing. A visual comparison of the blind (Figure 3) and linearly unmixed images (Figure 1) shows that both give similar results. The blind method seems better at detecting low signal levels and, as a result, produces images with more channel overlap. When we used the blindly separated images in the classification framework, we found that a simple classifier gives an accuracy of 66%, and the voting classifier gives 78% (data not shown). The overall accuracy is lower than the linearly unmixed data, and only the Golgi pattern is better distinguished

Table 5. Accuracy of Location Class Recognition in Each Tissue[a]

	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus	Accuracy (all data)	Accuracy (data with likelihood >0.5)
Adrenal gland (16/6)	100/0	100/100	100/0	100/100	50/0	100/100	100/100	100/100	93.8	100
Appendix (21/14)	100/100	50/0	100/0	100/100	100/100	100/100	100/100	100/100	95.2	100
Bone marrow (17/7)	50/100	0/0	50/0	66.7/100	100/100	100/0	100/100	100/100	70.6	100
Breast (13/4)	100/0	66.7/0	100/0	100/100	0/0	100/100	100/0	100/100	76.9	100
Bronchus (13/6)	100/0	100/0	100/0	100/100	100/100	100/100	0/0	100/100	92.3	83.3
Cerebellum (23/14)	100/100	100/100	66.7/100	100/100	100/100	66.7/100	100/100	100/100	91.3	100
Cerebral cortex (23/14)	66.7/100	100/100	66.7/0	100/100	66.7/100	66.7/100	66.7/66.7	100/100	78.3	92.9
Cervix uterine (17/6)	50/0	33.3/0	100/0	100/100	66.7/0	100/100	100/100	0/0	76.5	100
Colon (20/8)	100/100	100/0	100/0	100/100	100/0	100/100	100/100	50/100	95	100
Duodenum (23/14)	100/100	66.7/0	66.7/0	100/100	100/100	100/100	100/100	66.7/100	87	100
Endometrium (42/21)	100/100	50/0	80/100	100/100	100/0	100/100	80/80	100/100	88.1	90.5
Epididymis (16/7)	100/0	100/0	100/0	100/100	50/100	100/100	100/100	100/0	93.8	100
Esophagus (17/8)	100/100	66.7/100	100/0	100/100	0/0	100/100	50/100	100/100	76.5	100
Fallopian tube (23/13)	100/100	66.7/0	100/0	100/100	100/100	100/100	100/100	100/100	95.7	100
Gall bladder (19/5)	100/100	100/0	66.7/0	100/0	100/100	100/100	50/100	50/0	84.2	100
Heart muscle (24/11)	100/100	66.7/0	66.7/0	100/100	0/0	100/100	66.7/100	100/100	75	90.9
Hippocampus (23/14)	33.3/100	66.7/50	100/0	100/100	100/100	66.7/100	66.7/50	100/100	78.3	85.7
Kidney (24/10)	100/100	33.3/0	100/0	100/100	100/100	66.7/100	100/100	66.7/100	83.3	100
Lateral ventricle (23/12)	66.7/100	33.3/50	100/0	100/100	100/100	33.3/100	100/100	66.7/100	73.9	91.7
Liver (24/7)	66.7/100	100/100	66.7/0	100/100	100/100	33.3/100	66.7/100	33.3/0	70.8	100
Lung (24/8)	100/0	66.7/0	33.3/0	100/100	100/100	100/100	100/100	66.7/100	83.3	100
Lymph node (20/12)	100/100	100/0	100/0	100/100	100/100	50/100	100/100	100/100	95	100
Nasopharynx (14/5)	50/0	50/0	50/0	100/100	0/0	100/100	100/100	100/100	71.4	100
Oral mucosa (20/9)	100/100	100/0	100/0	100/0	66.7/100	100/100	100/100	100/100	95	100
Ovary (23/12)	66.7/0	66.7/0	100/0	100/100	100/100	66.7/100	66.7/50	100/100	82.6	91.7
Pancreas (22/11)	66.7/100	66.7/0	50/0	100/100	66.7/100	100/0	100/100	100/100	81.8	100
Parathyroid gland (17/6)	66.7/100	100/100	100/0	100/0	50/100	50/100	66.7/100	50/0	70.6	100
Placenta (23/11)	66.7/100	50/100	100/0	66.7/0	100/100	66.7/50	100/100	100/100	82.6	90.9
Prostate (24/13)	66.7/100	100/0	0/0	100/100	100/100	100/100	100/100	66.7/100	79.2	100
Rectum (21/7)	100/100	66.7/0	100/0	100/100	100/0	100/0	100/100	100/100	95.2	100
Salivary gland (23/10)	100/100	66.7/0	100/0	100/100	33.3/0	100/100	100/100	100/100	87	90
Seminal vescicle (19/8)	100/0	66.7/0	66.7/0	100/100	100/100	100/0	100/100	100/0	89.5	87.5
Skeletal muscle (20/8)	66.7/100	50/0	33.3/0	50/50	0/0	0/0	50/100	66.7/100	40	75
Skin (18/5)	33.3/100	66.7/0	100/0	100/100	33.3/0	100/100	100/100	100/100	72.2	100
Small intestine (21/8)	50/0	0/0	66.7/0	100/100	100/100	100/100	66.7/100	100/100	71.4	100
Smooth muscle (20/9)	66.7/100	33.3/100	50/0	100/100	66.7/100	100/100	100/100	100/0	75	100
Soft tissue (41/11)	66.7/100	60/0	100/0	75/100	20/0	100/100	80/100	80/100	73.2	100
Spleen (24/13)	100/100	100/100	100/0	66.7/100	66.7/100	100/100	100/100	100/100	91.7	100
Stomach (43/19)	100/100	83.3/0	100/0	100/100	100/100	100/100	80/100	50/100	88.4	100
Testis (24/11)	100/0	100/0	100/0	100/100	100/100	66.7/100	100/100	100/100	95.8	100
Thyroid gland (23/10)	100/100	33.3/0	100/0	100/100	100/100	66.7/100	100/100	33.3/0	78.3	100
Tonsil (23/13)	100/100	100/100	0/0	33.3/100	100/100	100/100	100/100	100/100	82.6	100
Urinary bladder (15/10)	100/100	100/100	100/0	100/100	100/0	100/100	100/100	100/100	100	100
Vagina (20/10)	66.7/100	66.7/0	100/0	100/100	100/0	50/100	100/100	100/100	85	100
Vulva anal skin (19/12)	66.7/0	100/100	100/0	100/100	100/100	100/100	100/100	100/100	94.7	100

[a] The number of testing images per tissue class is shown next to the tissue names. Values are shown as pairs (separated by '/') of accuracies when all images were considered and accuracies when only labels with a likelihood greater than 0.5 were considered. Accuracies are given in percentages.

by both the simple and voting classifiers using blindly unmixed data (data not shown). Classification accuracy when only samples with greater than 0.5 likelihood were considered was also lower than with linearly unmixed data (Figure 2). However, by summing the likelihoods of the two voting classifiers (one for each unmixing method), and using the highest likelihood sample labels, we were able to further boost the classification accuracy over all testing images to 83%, with six of the eight classes (not cytoskeleton and lysosome) recognized with greater than 80% accuracy (data not shown). Considering samples that have classification likelihood greater than 0.5 boosts the classification accuracy to 97% (Table 4), with an estimated 95% confidence interval of 1.6%.

Classification by Tissue. We next used the voting classifiers summed across unmixing methods to determine in which tissues pattern recognition is least successful (Table 5). When the classifier is applied to all testing images, classification in all tissues except for skeletal muscle is greater than 70%. Moreover, when considering only labels above 0.5 likelihood, the accuracy increases in nearly all tissues, and the minimum classification accuracy becomes 75% in skeletal muscle.

Conclusions

The automated classification systems described in this work yield accuracies as high as 97%, showing that analysis of the basic subcellular patterns in the HPA is feasible using these learning approaches. The results show that the image level multiresolution texture and nuclear overlap features are informative in characterizing the subcellular location patterns in the HPA images, and different wavelet decompositions can be used in a voting scheme to improve classification accuracy. This indicates that the different decompositions capture different information about the patterns. However, while the features do a significantly better job of describing nuclear patterns, they are not as informative with respect to cytoplasmic patterns, especially the cytoskeletal class, which are confused with other cytoplasmic location classes. Part of the inability of the system to distinguish between the cytoplasmic classes could be due to the moderate resolution and JPEG compression of the images, which do not allow for some of the finer details of these patterns to be distinguished. Regardless, this work suggests that features better characterizing the cytoplasm may help to improve analysis. The benefit of considering samples with high confidence labels, which can boost classification accuracy to above 95%, along with the fact that even when all images are considered the classification systems can distinguish six of the eight classes with higher than 80% accuracy, indicates that analysis of protein patterns across different tissues and cell types is feasible using the features described here.

This work also outlines a framework that allows for different unmixing methods to be compared to determine which is best for automated analysis. Visual inspection of blindly and linearly unmixed images hints that blind unmixing is better at separating stains, as it seems to better detect low levels of staining than linear unmixing. Classification results, however, show that simple linear unmixing is slightly more effective at providing informative features for classification. This indicates that the assumptions behind linear unmixing (linearly separable signals, stain colors are consistent enough across images) are suitable for the automated analysis of HPA images. However, combining classifications from linearly and blindly unmixed data does provide a slight boost in accuracy, meaning the blind unmixing is uncovering some additional relevant information missed by linear unmixing.

Moreover, the classification results in the Atlas tissues indicate that subcellular pattern recognition can be done effectively in nearly all tissues. Only in skeletal muscle does the classifier perform poorly, suggesting that this tissue should not be considered for subsequent automated analysis.

Most importantly, this work provides a framework that can be used to analyze all of the proteins in the Human Protein Atlas. The developed classifiers can be applied across the Atlas to determine the subcellular locations of all Atlas proteins. We anticipate that any time the classifier encounters a mixture pattern (e.g., a protein in both the nucleus and cytoskeleton) or a new pattern (such as an extracellular distribution), it would assign a label with a low confidence score. Such assignments could be screened out. To create more universal classifiers, however, we will potentially need to choose additional organelle classes for training (such as an extracellular location class to include extracellular matrix proteins). Moreover, we will likely need to train the system on more than just two proteins per location class. Despite the fact that we chose the 16 proteins based on their specificity for localizing to the eight major organelles, these proteins' locations are likely not fully characteristic of the patterns of all proteins that localize to similar compartments. Thus, finding out the minimum number of classes needed for training a more universal classifier is an ongoing and essential task for analyzing the full HPA. Of course, unsupervised learning methods may be needed to provide a more accurate understanding of subcellular patterns on a proteome-wide basis.

Acknowledgment. We thank Drs. Brian Athey, David States, Gil Omenn, and James Cavalcoli for many helpful discussions and support for this project. We also thank Amina Chebira, Dr. Shann-Ching Chen, Dr. Estelle Glory and all of our colleagues in the Center for Bioimage Informatics for helpful discussions and advice. This work was done using images made publicly available by the HPR project, directed by Dr. Mathias Uhlen. This research was supported in part by NIH grant 1 U54 DA0215 to the National Center for Integrated Biomedical Informatics and by NSF ITR grant EF-0331657.

References

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

(2) Garcia Osuna, E.; Hua, J.; Bateman, N.; Zhao, T.; Berget, P.; Murphy, R. Large-scale automated analysis of location patterns in randomly tagged 3T3 cells. *Ann. Biomed. Eng.* **2007**, *35*, 1081–1087.

(3) Murphy, R. F. Systematic Description of Subcellular Location for Integration with Proteomics Databases and Systems Biology Modeling, In *Proceedings of the 2007 IEEE International Symposium on Biomedical Imaging*; 2007, pp 1052−1055.

(4) Glory, E.; Murphy, R. F. Automated subcellular location determination and high throughput microscopy. *Dev. Cell* **2007**, *12*, 7–16.

(5) Chen, X.; Velliste, M.; Murphy, R. F. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytometry* **2006**, *69A*, 631–640.

(6) Boland, M. V.; Murphy, R. F. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* **2001**, *17* (12), 1213–1223.

(7) Chen, S.-C.; Gordon, G.; Murphy, R. F. A Novel Approximate Inference Approach to Automated Classification of Protein Subcellular Location Patterns in Multi-Cell Images, In *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging*; 2006, pp 558−561.

(8) Chebira, A.; Barbotin, Y.; Jackson, C.; Merryman, T.; Srinivasa, G.; Murphy, R. F.; Kovacevic, J. A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinf.* **2007**, *8*, 210.

(9) Huang, K.; Velliste, M.; Murphy, R. F. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc. SPIE—Int. Soc. Opt. Eng.* **2003**, *4962*, 307–318.

(10) Bengtsson, E.; Wahlby, C.; Lindbald, J. Robust cell image segmentation methods. *Pattern Recognit. Image Anal.* **2004**, *14* (2), 157–167.

(11) Coulot, L.; Kirschner, H.; Chebira, A.; Moura, J. M. F.; Kovacevic, J.; Osuna, E. G.; Murphy, R. F. Topology preserving STACS segmentation of protein subcellular location images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging*; 2006, pp 566−569.

(12) Huang, K.; Murphy, R. F. Automated classification of subcellular patterns in multicell images without segmentation into single cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging*; 2004, pp 1139−1142.

(13) Kononen, J.; Bubendorf, L.; Kallioniemi, A.; Barlund, M.; Schraml, P.; Leighton, P.; Torhorst, J.; Mihatsch, M. J.; Sauter, G.; Kallioniemi, O.-P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **1998**, *4*, 844–847.

(14) Uhlén, M.; Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **2005**, *4* (4), 384–393.

(15) Uhlén, M.; Björling, E.; Agaton, C.; Al-Khalili Szigyarto, C.; Amini, B.; Andersen, E.; Andersson, A.-C.; Angelidou, P.; Asplund, A.; Asplund, C.; Berglund, L.; Bergström, K.; Brumer, H.; Cerjan, D.; Ekström, M.; Elobeid, A.; Eriksson, C.; Fagerberg, L.; Falk, R.; Fall, J.; Forsberg, M.; Björklund, M. G.; Gumbel, K.; Halimi, A.; Hallin, I.; Hamsten, C.; Hansson, M.; Hedhammar, M.; Hercules, G.; Kampf, C.; Larsson, K.; Lindskog, M.; Lodewyckx, W.; Lund, J.; Lundeberg, J.; Magnusson, K.; Malm, E.; Nilsson, P.; Odling, J.; Oksvold, P.; Olsson, I.; Öster, E.; Ottosson, J.; Paavilainen, L.; Persson, A.; Rimini, R.; Rockberg, J.; Runeson, M.; Sivertsson, Å.;

Sköllermo, A.; Steen, J.; Stenvall, M.; Sterky, F.; Strömberg, S.; Sundberg, M.; Tegel, H.; Tourle, S.; Wahlund, E.; Waldén, A.; Wan, J.; Wernérus, H.; Westberg, J.; Wester, K.; Wrethagen, U.; Xu, L.-L.; Hober, S.; Pontén, F. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* **2005**, *4*, 1920–1932.

(16) Berger, C. E.; Koejer, J. A.; Glas, W.; Madhuizen, H. T. Color separation in forensic image processing. *J. Forensic Sci.* **2006**, *51* (1), 100–102.

(17) Ruifrok, A. C.; Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **2001**, *23*, 291–299.

(18) Rabinovich, A.; Agarwal, S.; Laris, C. A.; Price, J. H.; Belongie, S. Unsupervised color decomposition of histologically stained tissue samples. In *Advances in Neural Information Processing Systems*; 2003, pp 667−674.

(19) Hochreiter, S.; Clevert, C. A.; Obermayer, K. A new summarization method for Affymetrix probe level data. *Bioinformatics* **2006**, *22* (8), 943–949.

(20) Seung, H.; Lee, D. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.

(21) Haralick, R. M. Statistical and structural approaches to texture. *Proc. IEEE* **1979**, *67* (5), 786–804.

(22) Wu, T.-F.; Lin, C.-J.; Weng, R. C. Probability estimates for multi-class classification by pairwise coupling. *J. Machine Learn. Res.* **2004**, *5*, 975–1005.

(23) Huang, K.; Murphy, R. F. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinf.* **2004**, *5*, 78.

PR7007626