Journal of Biomedicine and Biotechnology • 2005:2 (2005) 87–95 • DOI: 10.1155/JBB.2005.87

RESEARCH ARTICLE

# Objective Clustering of Proteins Based on Subcellular Location Patterns

Xiang Chen[1,3] and Robert F. Murphy[1,2,3]

[1]*Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA*
[2]*Department of Biomedical Engineering, 2100 Doherty Hall,*
*Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA*
[3]*Center for Automated Learning and Discovery, School of Computer Science,*
*Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891, USA*

Received 8 September 2004; accepted 4 November 2004

The goal of proteomics is the complete characterization of all proteins. Efforts to characterize subcellular location have been limited to assigning proteins to general categories of organelles. We have previously designed numerical features to describe location patterns in microscope images and developed automated classifiers that distinguish major subcellular patterns with high accuracy (including patterns not distinguishable by visual examination). The results suggest the feasibility of automatically determining which proteins share a single location pattern in a given cell type. We describe an automated method that selects the best feature set to describe images for a given collection of proteins and constructs an effective partitioning of the proteins by location. An example for a limited protein set is presented. As additional data become available, this approach can produce for the first time an objective systematics for protein location and provide an important starting point for discovering sequence motifs that determine localization.

## INTRODUCTION

The biotechnology revolution, especially the development of high-throughput technologies, has led to a rapid explosion of biological raw data that could not been imagined a few decades ago. For the first time in history, biologists can perform metaanalysis on available experimental data (largely unorganized) in order to generate hypotheses for the mechanisms by which cells, tissues, and organisms carry out their specialized functions. Until recently, systematic efforts to describe protein location have been limited to the assignment by database curators of a relatively small set of terms to each protein. While the recent development of restricted vocabularies for this purpose (most prominently the Gene Ontology Consortium cellular component ontology) has been an important step, such vocabularies do not have the ability to uniquely identify the many (probably on the order of a hundred) distinct, complex subcellular patterns displayed by proteins. To complement these approaches, we have applied pattern recognition and machine learning methods to this general problem, and coined the term "location proteomics" to describe the branch of proteomics that systematically and objectively studies the location patterns of individual proteins and their relationships [1].

Cells vary greatly in their size, shape, intensity, position and orientationation in fluorescent images, and consequently raw pixel intensity values are not very useful in location pattern recognition in general. The core of our group's previous work has been the development of sets of numerical features (termed subcellular location features, or SLFs) to represent the patterns of proteins seen in fluorescence microscope images without being overly sensitive to changes in intensity, rotation, and position of a cell [2, 3]. These numerical descriptions of subcellular location have been validated by developing automated classifiers that can correctly assign previously unseen images to the major classes of subcellular structures or organelles [2, 3, 4, 5].

With the development of automated high-resolution microscopy technology [6, 7], the capability now exists for capturing high-resolution 3D fluorescence microscope images of protein subcellular distributions. Coupled with technologies that create cell lines expressing randomly tagged proteins [8, 9], it is possible to collect large numbers of images for diverse proteins in a given cell type

Correspondence and reprint requests to Robert F. Murphy, Department of Biomedical Engineering, 2100 Doherty Hall, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA, Email: murphy@cmu.edu

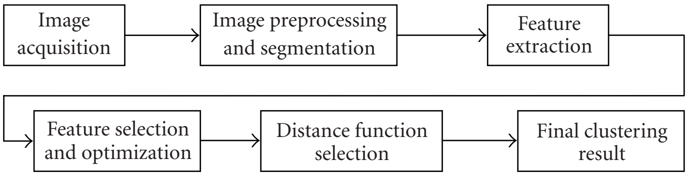
Image acquisition
Image preprocessing and segmentation
Feature extraction
Feature selection and optimization
Distance function selection
Final clustering result

Figure 1. Flow chart for clustering protein subcellular location patterns.

within a reasonable time scale. The work described in this paper tries to approach the ultimate goal of determining which proteins imaged in such a project share the same location pattern. The problem could be stated alternatively as follows. *Given a set of proteins, each with multiple image representations, find a partitioning of the protein set such that images from members in the same partition show a single location pattern.*

From the computational view, the task of finding the optimal grouping of a set of proteins based on their subcellular location patterns can be described as finding the maximum number of partitions so that the SLF features of proteins within the same partition are statistically indistinguishable while the features for any two proteins from different partitions are distinguishable. This is the classic clustering problem. It is formally identical to those appearing in many other fields, such as identifying gene clusters from mRNA expression levels. In our case, however, image features are measured on widely varying scales with different units while mRNA expression levels are at least all expressed in the same units. This inhomogeneity of units complicates the process of feature selection and distance definition.

Building on our initial work demonstrating the use of the SLF to build subcellular location trees [1], we describe here clustering approaches for constructing objective partitionings of proteins by location. We started by making what we consider to be a reasonable assumption: that the majority of images for a specific protein in interphase cells should show the same location pattern. Under this assumption, we propose a method (shown as a flow chart in Figure 1) for automatically determining the number of partitions for the dataset and performing the partitioning accordingly.

## METHODS

### *Image acquisition*

A set of NIH 3T3 cells clones each expressing a different GFP-tagged protein were obtained by CD-tagging [10]. The acquisition of high resolution 3D images of these clones by spinning disk confocal microscopy has been described [1]. Briefly, the pixel spacing in both directions in the image plane was 0.11 $\mu$m and the vertical spacing between adjacent planes (slices) was 0.5 $\mu$m. The gray level of each pixel is between 0 and 4095 (12 bits per pixel). The resulting 1280 × 1024 × 31 3D images each contained from 1 to 3 cells. Ninety differently tagged protein clones (with 8 to 33 cells per clone) were included in the current study. Example images are shown in Figure 2.

### *Image preprocessing and segmentation*

The procedures employed in image preprocessing have been described in detail previously [3, 11]. In brief, background in each image was removed and single-cell images were obtained through image segmentation (either automated or manual). Single-cell images were then thresholded using an automated method.

### *Feature extraction and optimization*

The SLFs used in the current study can be divided into three categories:

(i) morphological features [5], based on finding the fluorescent objects in an image. A fluorescent object is a set of connected pixels with above threshold intensities,

(ii) edge features [1], which capture the amount of fluorescence distributed along edges,

(iii) haralick texture features [1, 12], based on the gray level co-occurrence matrix of an image, which capture the correlation between adjacent pixel intensities.

This combination of features (previously defined as 3D-SLF11) were extracted on the preprocessed single-cell images according to procedures described previously [1, 5, 12] with one modification. Previous studies suggested that pixel resolution and gray levels could potentially influence the discriminating power of the Haralick texture features [12, 13]. Therefore, the Haralick texture features were calculated at various degrees of downsampling (to 0.5, 1.0, 1.5, 2.0, and 2.5 $\mu$m pixel size) and various numbers of gray levels (16, 64, and 256). The optimal values for the current dataset were determined as 0.5 $\mu$m pixel size and 64 gray levels using the method described in [12].

### *Feature selection*

For an arbitrary collection of proteins, it is not a trivial task to identify the optimal feature set to use for clustering them. One idea is to generate a number of different feature

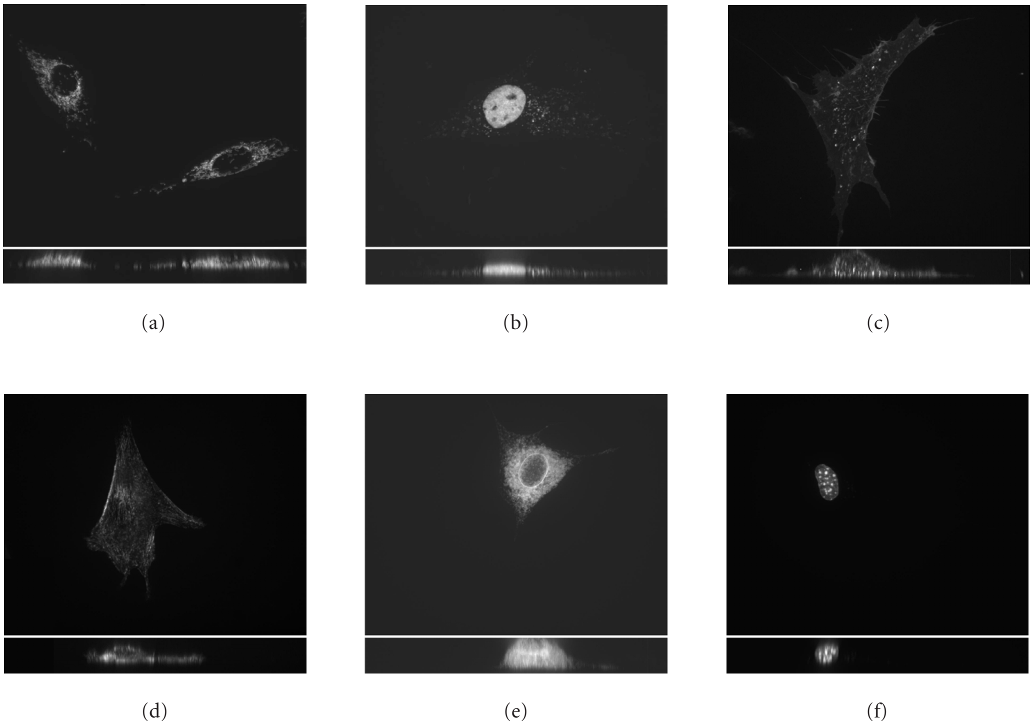

FIGURE 2. Selected images from the 3D 3T3 image dataset. Tagged protein names are shown with a hyphen followed by a clone number if the same protein was tagged in more than one clone in the dataset. Representative images are shown for (a) Atp5a1-1, (b) Ewsh, (c) Glut1, (d) Tubb2-1, (e) Canx, and (f) Hmga1-1. The top portion of each panel shows a projection on the $x$-$y$ plane and the bottom shows a projection on the $x$-$z$ plane.

sets and to use them to train a classifier that tries to distinguish every protein in the collection. We then consider the best feature set available to be the one with the highest overall classification accuracy. Of course, the overall accuracy is not an accurate estimate of the classifier's true discriminating power since some proteins in the collection may share a single location pattern. These proteins would be indistinguishable for a classifier and the classification result among these proteins could be largely random (lowering the overall accuracy). However, that accuracy is still a good metric for choosing a feature set since it will increase as informative features are added and decrease as they are removed.

Stepwise discriminant analysis (SDA) was used for selection of those "informative" features that support the discrimination between proteins with different patterns. The stepdisc function of SAS (SAS Institute, Cary, NC) was used with default parameter values (stepwise selection method, all variables included in the model calculation, start from no variable in the model, use 0.15 as the significance level for adding or retaining variables). The input to SDA was the full feature matrix for all cells for all clones and the output was a ranked list of features that were considered to contribute to distinguishing the clones.

To select the optimal feature subset for clustering, increasing numbers of the ranked features were used to train classifiers to try to distinguish each protein clone as described previously [12] with one exception: instead of the neural network classifier, we used a support vector machine (SVM) classifier with max-win strategy [14].

### Distance function

As a starting point for this work, we used (1) a Euclidean distance function, which calculates the distance between each pair as the square root of the sum of squares of the feature differences over the whole feature set (each feature is normalized to zero mean and unit variance across the entire image collection) and (2) a Mahalanobis distance function, which further takes the correlations between features into account.

In the current 3D 3T3 dataset, most morphological, edge, and texture mean features were either in a single-mode, bell, shaped distribution (Figure 3a) or in a single-mode, exponential-shaped distribution (Figure 3b). The distributions for some texture range features were more complex with double modes (Figure 3c).

To avoid excessive weighting of features whose absolute values happen to be larger than the other features (compare Figures 3b and 3c) in Euclidean distance calculation, we first normalized all features to z-scores (subtracting the mean for each feature and dividing

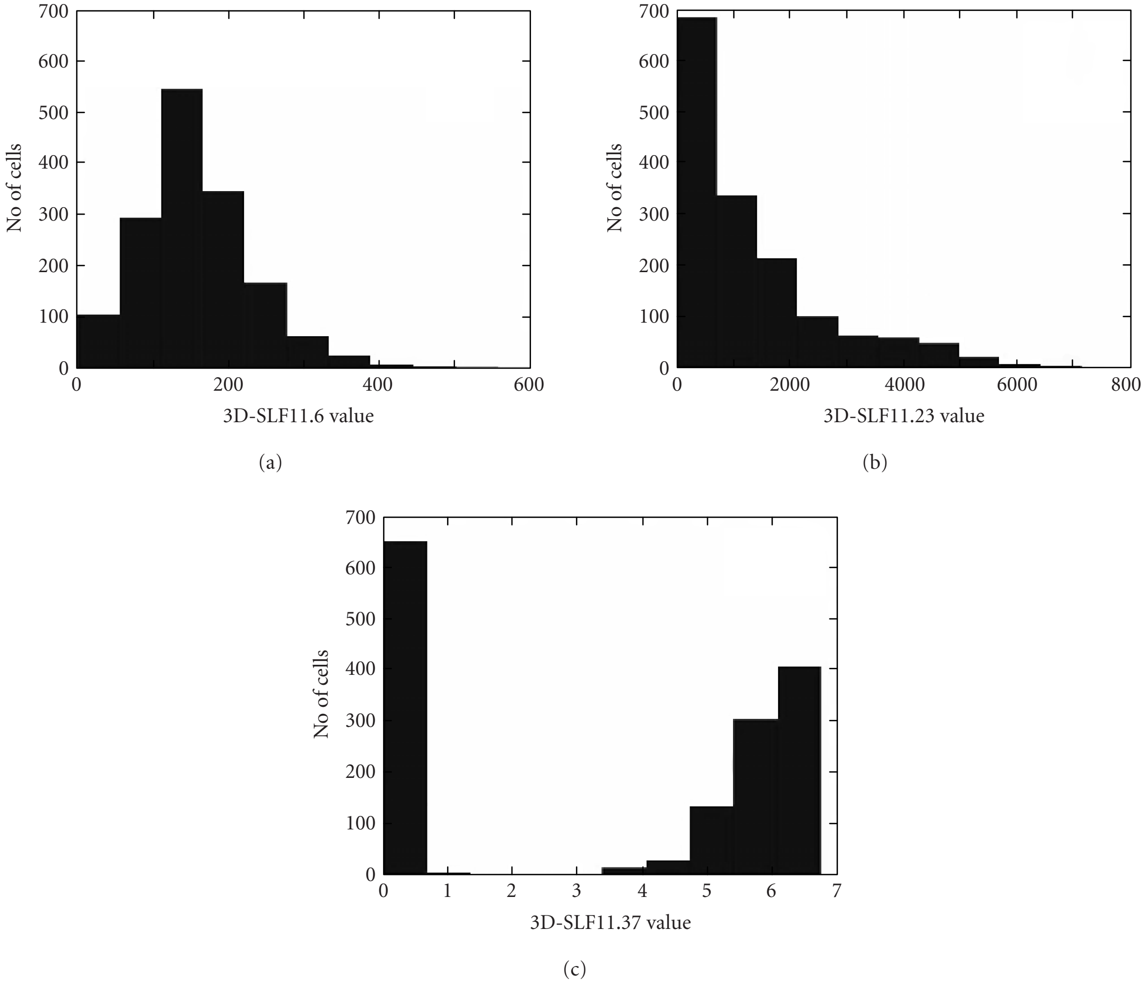

FIGURE 3. Histograms of selected features before z-score normalization. Examples of features with (a) a roughly Gaussian distribution (3D-SLF11.6, average object to center of fluorescence distance), (b) a roughly Poisson distribution (3D-SLF11.23, texture feature average of co-occurrence matrix sum variance), and (c) a biomodal distribution (3D-SLF11.37, texture feature range of co-occurrence matrix sum entropy).

by its standard deviation, both calculated across all clones).

Due to differences in the number of single-cell images for each clone, we randomly selected five images for each clone to construct a global covariance matrix. This process was repeated 100 times and the mean value was taken as the final covariance matrix used in the Mahalanobis distance calculation.

### *Clustering/partitioning algorithms*

Each single-cell image from all clones was first converted to a feature vector and $k$-means clustering was performed on the entire image set using varying $k$ (from 2 to the total number of clones). Akaike information content (AIC) was then used as a criterion to select the optimal value of $k$ [15].

As a parallel approach, hierarchical clustering was performed on *mean* feature vectors for each clone. Since the image collection contains multiple images for each individual protein, we can construct many estimates of the mean feature vectors by randomly selecting half of the images for each protein. For each randomly chosen set, the simplest tree building algorithm, unweighted pair-group method with arithmetic mean (UPGMA) algorithm, was used to construct a distance tree (dendrogram). These were used to form a consensus tree [16] which contains those structures with general agreement in the set of trees for all random trials. Different partitionings of the protein set could be obtained by cutting the consensus tree at different heights (lower height yields more clusters). Optimal partitioning was selected using the AIC criteria.

A third approach (Algorithm 1) started with the confusion matrix created by the classifier described in the "feature selection" section. It is expected that some clones share a single location pattern. Consequently, images from different clones with the same pattern will be expected to be assigned to one of those clones largely at random. To cluster these together, the confusion matrix was searched for off-diagonal elements that were above a threshold and the clones corresponding to these elements were merged. We select the threshold to yield a similar number of clusters to the optimal $k$ obtained in the $k$-means/AIC algorithm.

The last approach we took was visual inspection where one or more descriptive term (e.g., uniform, cytoplasmic, nucleolar) was assigned for each clone after visually examining all images sequentially displayed on a monitor. Clones with the same combination of descriptive terms were grouped into the same cluster.

### *Evaluation of distance functions*

The choice of distance function is critical to any clustering task. We do not intend to propose an optimal distance function here since it is possible that the best distance function should be determined individually for different datasets, either theoretically or experimentally. Instead we proposed to evaluate the effectiveness of different distance functions by measuring the agreement of the partitioning using the same distance function with different algorithms. The intuition behind this method is that a good distance function should be able to yield consistent partitioning of the dataset with different clustering algorithms.

In order to measure the degree of agreement among different clustering results, we used Cohen's $\kappa$ statistic [17, 18] to compare two partitionings A and B:

$$\kappa(\mathrm{A}, \mathrm{B}) = \frac{\text{Observed agreement}-\text{expected agreement}}{1-\text{expected agreement}} \tag{1}$$

Observed agreement is defined as the portion of protein pairs where the two clustering results agree (the pair belongs either to a single cluster or to two distinct clusters for both results). Expected agreement can be defined as the agreement between two random partitionings with the same distribution frequencies as A and B, respectively. Calculating the expected agreement is difficult but it can be estimated by simulation. The $\kappa$ statistic represents the portion of agreement in the two clustering results beyond chance, with a maximum value of 1 for perfect agreement. By running multiple simulations (randomly, independently assigning the set of clones to different partitions in A and B based on their marginal distribution probabilities and then calculating the observed $\kappa$ statistic), it is also feasible to estimate the variance of the $\kappa$ statistics under the null hypothesis that partitionings A and B are independently and randomly distributed.

```
Procedure clustering_on_confusionmatrix (ConfusionMatrix, threshold)
Initialize cluster[i] = i for each i
While(max(off diagonal values in ConfusionMatrix) >= threshold) do
    normalize the ConfusionMatrix so that the
  sum of each row is 100
    select i < j such that ConfusionMatrix(i, j) is
  the largest above threshold off diagonal value of
  ConfusionMatrix
    set cluster[i] = cluster[i] ∪ cluster[j]
  clear cluster[j]
    set ConfusionMatrix[i,:] = ConfusionMatrix[i,;] + ConfusionMatrix[j,:]
    set ConfusionMatrix[:,i] = ConfusionMatrix[:,i] + ConfusionMatrix[:,j]
    clear ConfusionMatrix[j,:]
    clear ConfusionMatrix[:,j]
 End While
 return cluster
End Procedure
```

Algorithm 1. Procedure: clustering on confusionmatrix (ConfusionMatrix, threshold).

## RESULTS

The 3D 3T3 dataset, consisting of 90 randomly tagged protein clones, was obtained using CD-tagging techniques [8]. We first constructed the optimal feature subset to use in clustering these proteins by their location patterns, as described in the "Methods" section. Since morphological, edge, and texture features have all been shown to be useful for classifying both 2D [3] and 3D images [12], we began our search for discriminating features using a set of 42 features drawn from all three types. Using the method described before, a subset of 34 features (which we defined as 3D-SLF18, see Table 1) gave the best overall classification accuracy on a subset of 46 clones from the 3D 3T3 dataset (data not shown). This feature subset, consisting of 9 morphological features, 1 edge feature, and 24 texture features, was used for subsequent clustering procedures in this study.

We next consider approaches to clustering these proteins. As an initial approach, we propose clustering all individual images and determining an optimal number of clusters (the large number of individual images makes this estimate feasible). To do this, individual images were first converted to feature vectors and $k$-means clustering was then performed on the whole image set using various $k$ values (from 2 to the total number of proteins included in the collection). Under the reasonable assumption that a majority of the cells in a clone share a single location pattern, the range of $k$ should cover the optimal number of clusters/partitionings in the image set. Each value of $k$ gave a specific clustering of the images with different cluster compactness (measured by the variances within the clusters). Akaike information content (AIC) was then used as a criterion to select the optimal value of $k$. AIC measures the fitness of the current model given the data, adjusted by the number of parameters included in the model (to avoid overfitting). This

Table 1. Optimal feature set for distinguishing the 3D 3T3 images (3D-SLF18). The features are listed in decreasing order of discriminating power as evaluated by SDA.

| Feature name | Feature description |
|---|---|
| 3D-SLF11.16 | The fraction of fluorescence in above threshold pixels that are along an edge |
| 3D-SLF11.19 | Average of correlation |
| 3D-SLF11.23 | Average of sum variance |
| 3D-SLF11.31 | Range of contrast |
| 3D-SLF11.5 | Ratio of maximum object volume to minimum object volume |
| 3D-SLF11.28 | Average of info measure of correlation 1 |
| 3D-SLF11.3 | Average object volume (average number of above threshold pixels per object) |
| 3D-SLF11.21 | Average of inverse difference moment |
| 3D-SLF11.24 | Average of sum entropy |
| 3D-SLF11.33 | Range of sum of squares of variance |
| 3D-SLF11.22 | Average of sum average |
| 3D-SLF11.29 | Average of info measure of correlation 2 |
| 3D-SLF11.25 | Average of entropy |
| 3D-SLF11.34 | Range of inverse difference moment |
| 3D-SLF11.2 | Euler number of the cell |
| 3D-SLF11.41 | Range of info measure of correlation 1 |
| 3D-SLF11.27 | Average of difference entropy |
| 3D-SLF11.26 | Average of difference variance |
| 3D-SLF11.37 | Range of sum entropy |
| 3D-SLF11.40 | Range of difference entropy |
| 3D-SLF11.35 | Range of sum average |
| 3D-SLF11.36 | Range of sum variance |
| 3D-SLF11.20 | Average of sum of squares of variance |
| 3D-SLF11.32 | Range of correlation |
| 3D-SLF11.4 | Standard deviation (SD) of object volumes |
| 3D-SLF11.38 | Range of entropy |
| 3D-SLF11.10 | SD of absolute value of the horizontal component of object to protein center of fluorescence (COF) distances |
| 3D-SLF11.9 | Average absolute value of the horizontal component of object to COF distance |
| 3D-SLF11.18 | Average of contrast |
| 3D-SLF11.13 | SD of signed vertical component of object to protein center of fluorescence (COF) distances |
| 3D-SLF11.6 | Average object to COF distance |
| 3D-SLF11.17 | Average of angular second moment |
| 3D-SLF11.42 | Range of info measure of correlation 2 |
| 3D-SLF11.12 | Average signed vertical component of object to protein center of fluorescence (COF) distances |

gives a maximum likelihood estimate of the number of clusters given the data. Once the partitioning of the images was determined, all of the images belonging to the same protein were considered and the protein was allocated to the cluster that contained the maximum number of images from this protein as long as it accounted for at least 1/3 of the total images. Only those images belonging to this cluster were retained. When a given protein's images were found in several clusters so that none of the clusters had at least 1/3, that protein's location pattern was considered undetermined and it was dropped from further consideration. This reflects our initial assumption (or condition) that a protein has a unique pattern. The result of this stage is a clustering for only those protein images for which an assignment can be made with confidence.

As a parallel approach, we can perform hierarchical clustering on the average feature values for each protein (after eliminating the proteins considered too variable in the previous stage). We used the mean feature vector of each protein to construct a dendrogram. Since the image collection contains multiple images for each individual protein, we can construct many trees each of which is for a randomly selected half of the images for each protein. These are used to form a consensus tree [16], which contains the common structures with general agreement in

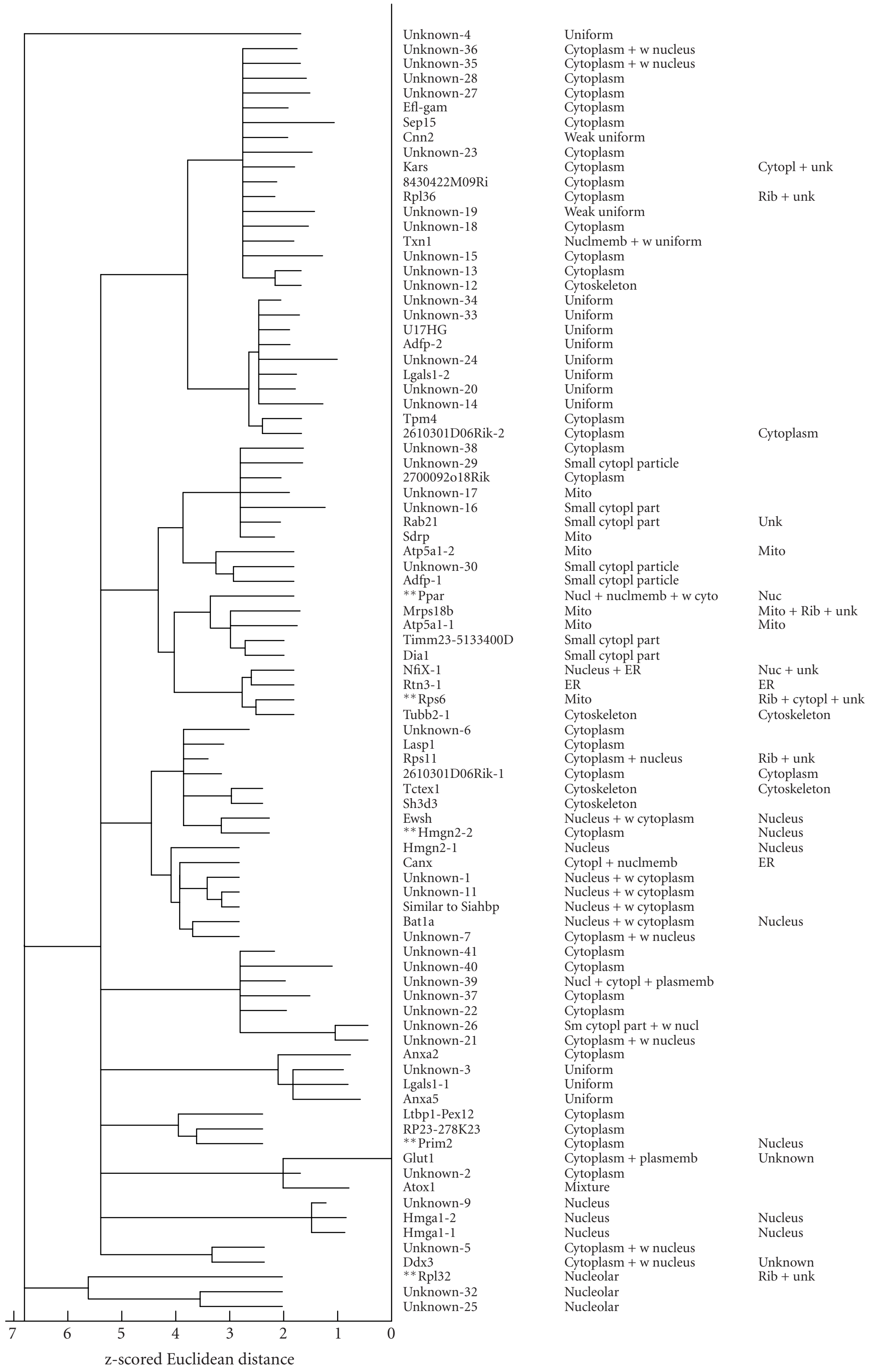

FIGURE 4. A consensus subcellular location tree generated on the 3D 3T3 image dataset using SDA-selected 3D-SLF11 features. The columns show the protein names (if known), human observations of subcellular location, and subcellular location inferred from gene ontology (GO) annotations. The sum of the lengths of horizontal edges connecting two proteins represents the distance between them in the feature space. Proteins for which the location described by human observation differs significantly from that inferred from GO annotations are marked (∗∗).

Table 2. Comparison of clustering methods and distance functions. The agreement between the sets of clusters resulting from the four clustering methods described in the text was measured using the $\kappa$ test. The standard deviations of the statistic under the null hypothesis were estimated to range between 0.014 and 0.023 from multiple simulations.

| Clustering approaches compared | z-scored Euclidean distance | Mahalanobis distance |
|---|---|---|
| | $\kappa$ | $\kappa$ |
| $k$-means/AIC versus consensus | 1 | 0.5397 |
| $k$-means/AIC versus ConfMat | 0.4171 | 0.3634 |
| Consensus versus ConfMat | 0.4171 | 0.1977 |
| $k$-means/AIC versus visual | 0.2055 | 0.1854 |
| Consensus versus visual | 0.2055 | 0.1156 |

the set of original trees. The clusters found in this tree can be compared to those obtained from clustering individual images.

We first compared the performance of the two different distance measures. It is reasonable to assume that a better distance function should produce greater agreement among clustering results using different algorithms. Since only the $k$-means/AIC and consensus hierarchical clustering algorithms utilized the distance function, we compared the agreement between the two clustering results using the $\kappa$ statistic. In addition, we also compared these results against the results obtained using the other two algorithms (visual assignment and clustering using confusion matrix). Table 2 summarizes the results. Clearly the z-scored Euclidean distance function produced larger agreement than the Mahalanobis distance function. Therefore, we used the z-scored Euclidean distance function for the rest of the study. Another major point from Table 2 is that the agreements between visual clustering and the other approaches were clearly lower than the agreement between any pair of the machine clustering algorithms. The consistency seen among the automated methods confirms their value for generating location pattern annotations in proteomics projects.

When Euclidean distance was used as the distance function with the $k$-means/AIC algorithm for individual images, the optimal number of clusters found was 30. However, 13 of the 30 clusters contained only outliers from protein clones and therefore we obtained 17 clusters from this set of proteins. Out of all 90 clones, 3 were removed by the consistency requirement described above. The corresponding consensus tree obtained in parallel using average features is shown in Figure 4. The consensus tree was drawn in an additive style in which the sum of length of edges connecting pairs of proteins represents the distance between them.

Examination of the consensus tree (and the clusters obtained from $k$-means/AIC algorithms, not shown) reveals that proteins expected to have similar location patterns were mostly grouped properly. For example, the only three nucleolar proteins (Rpl32, Unknown-25, and Unknown-32) are grouped together. It should also be noted that there are two major nuclear protein clusters, one with Hmga-1, Hmga-2, and Unknown-9 and the other with Unknown-1, Unknown-11, similarly to Siahbp1 and Bat1a. The first cluster contained proteins with an exclusively nuclear distribution while the second cluster contained nuclear proteins with minor cytoplasmic distributions. The separation of these proteins into two clusters indicates that they are statistically distinguishable, in agreement with our previous results [1].

The consensus tree in Figure 4 has been incorporated into a web interface (http://murphylab.web.cmu.edu/services/PSLID/) that allows the underlying images for any branch to be displayed interactively.

## CONCLUSIONS AND DISCUSSION

We have previously shown that the major protein subcellular location patterns can be described numerically by SLFs. Automated classifiers trained on these features can determine protein location patterns from previously unseen fluorescence images.

The observation that the SLFs used for this automated classification were clearly effective in distinguishing subcellular patterns suggested that a properly chosen partitioning of proteins using SLFs would group proteins based on their location patterns. We describe automated methods to create such a partitioning objectively. Our initial trial on a modest set of randomly tagged proteins using a set of morphological, edge, and texture features largely validates this method.

It should be pointed out that by increasing the dimensionality of protein images (e.g., by adding time as a fourth dimension and the presence of various drugs as a fifth dimension), proteins currently in the same cluster would be potentially distinguishable. This will of course require development of new features that reflect the characteristics of the higher dimensions.

In closing, we suggest that the development of an automated, systematic, and objective clustering approach for protein location patterns is critical to finding potential targeting motifs in protein sequences, just as automated clustering of gene expression data has been a prerequisite for automated detection of regulatory elements [19, 20, 21].

## ACKNOWLEDGMENTS

We thank Jonathan Jarvik, Peter Berget, and all of our colleagues in the CD-tagging project for helpful discussions and providing images of tagged cell lines. This work was supported in part by NIH grant R01 GM068845, NSF grant EF-0331657, and a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund. X. Chen was supported by a graduate fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation.

## REFERENCES

[1] Chen X, Velliste M, Weinstein S, Jarvik JW, Murphy RF. Location proteomics - building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc SPIE.* 2003;4962:298–306.

[2] Murphy RF, Boland MV, Velliste M. Towards a systematics for protein subcelluar location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:251–259.

[3] Boland MV, Murphy RF. 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.

[4] Huang K, Murphy RF. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics.* 2004;5(1):78.

[5] Velliste M, Murphy RF. Automated determination of protein subcellular locations from 3d fluorescence microscope images. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*. New York, NY: IEEE; 2002:867–870.

[6] Nakano A. Spinning-disk confocal microscopy–a cutting-edge tool for imaging of membrane traffic. *Cell Struct Funct.* 2002;27(5):349–355.

[7] Price JH, Goodacre A, Hahn K, et al. Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. *J Cell Biochem Suppl.* 2002;39:194–210.

[8] Jarvik JW, Adler SA, Telmer CA, Subramaniam V, Lopez AJ. CD-tagging: a new approach to gene and protein discovery and analysis. *Biotechniques.* 1996;20(5):896–904.

[9] Rolls MM, Stein PA, Taylor SS, Ha E, McKeon F, Rapoport TA. A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J Cell Biol.* 1999;146(1):29–44.

[10] Jarvik JW, Fisher GW, Shi C, et al. In vivo functional proteomics: mammalian genome annotation using CD-tagging. *Biotechniques.* 2002;33(4):852–854, 856, 858–860 passim.

[11] Hu Y, Murphy RF. Automated interpretation of subcellular patterns from immunofluorescence microscopy. *J Immunol Methods.* 2004;290(1-2):93–105.

[12] Chen X, Murphy RF. Robust classification of subcellular location patterns in high resolution 3d fluorescence microscope images. *Proceeedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New York, NY: IEEE; 2004:1632–1635.

[13] Murphy RF, Velliste M, Porreca G. Robust classification of subcellular location patterns in fluorescence microscope images. *Proceeedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP 2002)*. New York, NY: IEEE;2002:67–76.

[14] Huang K, Murphy RF. Automated classification of subcellular patterns in multicell images without segmentation into single cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging (ISBI 2004)*. New York, NY: IEEE; 2004:1139–1142.

[15] Ichimura N. Robust clustering based on a maximum-likelihood method for estimating a suitable number of clusters. *Syst Comp Jpn.* 1997;28(1):10–23.

[16] Thorley JL, Page RM. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics.* 2000;16(5):486–487.

[17] Cook R. Kappa. In: P. Armitage and T. Colton, eds. *The Encyclopedia of Biostatistics.* New York, NY: Wiley; 1998:2160–2166.

[18] Cook R. Kappa and its dependence on marginal rates. In: P. Armitage and T. Colton, eds. *The Encyclopedia of Biostatistics*. New York, NY: Wiley; 1998:2166–2168.

[19] Jelinsky SA, Estep P, Church GM, Samson LD. Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes. *Mol Cell Biol.* 2000;20(21):8157–8167.

[20] Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx. *Curr Biol.* 2000;10(6):301–310.

[21] Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.* 2003;34(2):166–176.