Interpretation of Protein Subcellular Location Patterns in 3D Images Across Cell Types and Resolutions

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

[1] Center for Bioimage Informatics and Departments of Biological Sciences and Machine Learning, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213
xiang.chen@yale.edu, murphy@cmu.edu

Abstract. Detailed knowledge of the subcellular location of all proteins and how they change under various conditions is essential for systems biology efforts to recreate the behavior of cells and organisms. Systematic study of subcellular patterns requires automated methods to determine the location pattern for each protein and how it relates to others. Our group has designed sets of numerical features that characterize the location patterns in high-resolution fluorescence microscope images, has shown that these can be used to distinguish patterns better than visual examination, and has used them to automatically group proteins by their patterns. In the current study, we sought to extend our approaches to images obtained from different cell types, microscopy techniques and resolutions. The results indicate that 1) transformation of subcellular location features can be performed so that similar patterns from different cell types are grouped by automated clustering; and 2) there are several basic location patterns whose recognition is insensitive to image resolution over a wide range. The results suggest strategies to be used for collecting and analyzing images from different cell types and with different resolutions.

Keywords: Location Proteomics, Pattern Recognition.

1 Introduction

The central goal of proteomics is to characterize all proteins in a given cell type, which includes but is not restricted to the characterization of protein sequence, structure, expression, localization, function and regulation. Subcellular location is a critical property of a protein, and knowledge of the location pattern is essential to the complete understanding of its function. However, unlike many other properties for which systematic and automated approaches have been developed [1, 2], the systematic study of protein location patterns has just started [3]. Most of the approaches that exist today are based on assignment by database curators of terms from a restricted vocabulary, such as that generated by the Gene Ontology Consortium (http://geneontology.org). These approaches generally cannot distinguish location patterns beyond the level of major subcellular organelle categories. The development of random tagging techniques [4, 5] enables us to produce on a reasonable time scale a large number of clones from a given cell type that each express a different fluorescently-tagged

S. Hochreiter and R. Wagner (Eds.): BIRD 2007, LNBI 4414, pp. 328–342, 2007.

protein [6, 7]. Advances in microscope technologies have further enabled us to record high resolution, 3D fluorescence images of living cells. These techniques combined together can provide a collection of high resolution images for a large number of proteins in a given cell type. Machine learning techniques have been applied to provide tools for automated and systematic analysis of,such images, initially by our group [8-10] and more recently by others [11, 12]. These methods are suitable for automated analysis of subcellular location on a proteome-wide basis.

The core of our approach is the design of Subcellular Location Features (SLF), numerical descriptors that can quantitatively describe the distribution of proteins inside a cell. Cells vary greatly in their size, shape, position and rotation in the field. The total intensity of a cell is also affected by the labeling and imaging techniques. We designed SLFs so that they are able to represent location patterns while at the same time not being too sensitive to changes in cell intensity, position and orientation. Using the SLFs, we have built automated classifiers that are able to distinguish the major subcellular organelles and structures with high accuracy. An important conclusion of this work is that some classifiers built on SLFs are able to distinguish location patterns which are essentially indistinguishable to human visual inspection [13].

Besides automated classifiers, we have also described approaches for objectively partitioning an arbitrary protein collection by subcellular location pattern. Initial trials with a limited set of randomly tagged proteins in the NIH 3T3 cell line demonstrated the feasibility of this approach [14, 15].

Our previous trials have validated the effectiveness of SLFs in describing protein subcellular location patterns in different data sources [9, 10, 15] Although each of these trials was based on a single dataset, they suggest that SLFs describe general characteristics of location patterns, regardless of the data source. However, we have observed that classifiers trained on one dataset cannot simply be applied to another dataset taken for a different cell type or under different microscopy conditions. We therefore sought features and/or transformations of features that would be robust against variations in cell type or labeling and imaging protocols. This would be an important step towards our ultimate goal of systematically studying protein location patterns across all cell types. It would enable researchers to utilize classification or clustering procedures without first building their own reference set. Furthermore, as more and more published literature becomes available on line, we can potentially acquire and analyze many images without performing a single experiment. Such an effort (the SLIF, or Subcellular Location Image Finder, project) is ongoing in our group [16]. Given a suitable feature set and/or transformation, images retrieved from a range of sources could be analyzed to give a more comprehensive understanding of location.

A separate but related issue is the relationship between the resolution of an image and the resolution of the subcellular location assignment that can be made from it. Our initial experiments on 3D images were performed on a set of high resolution confocal fluorescence microscope images (0.05 – 0.1 μm horizontal pixel resolution) [10, 14]. However, even with recent advances in automated imaging technologies, the data acquisition step for high resolution fluorescence images is still the bottleneck for location studies at the proteome level. There is frequently a tradeoff between speed and resolution during image collection yet image resolution is one of the most important factors affecting the discriminating power of any interpretation method. For example, although we have shown that using our feature based classification approaches, two Golgi proteins (giantin and gpp130) could be distinguished with near-perfect accuracy in the high resolution 3D HeLa dataset [17], it is unlikely that these proteins would be distinguishable using images where the Golgi complex is barely separated from other subcellular organelles. Therefore, it is essential to establish guidelines for the maximum discrimination between subcellular patterns that can be expected for images at different resolutions. Strategies for image collection appropriate to a given problem can be properly chosen once such a guideline is established.

We have previously evaluated the effects of several image manipulations on 2D SLFs, including compression, resizing, and intensity rescaling [18]. The results suggested that while most individual features were sensitive to resizing, neural network classifiers trained on resized images were robust to a minor degree of resizing. It is therefore of interest to more fully study the performance of automated interpretation tools using images at different resolutions, and to consider 3D images.

This paper describes initial approaches to the problem of analyzing protein location patterns across different cell types and conditions. We combined 3D image data from two sources: a HeLa cell dataset of 11 different location patterns obtained by immunofluorescence labeling and a 3T3 cell dataset for 90 clones expressing different GFP-tagged proteins obtained by random tagging techniques. We searched for a proper transformation between features calculated for the 3D HeLa dataset and their counterparts for the 3D 3T3 dataset. We also evaluated the effects of pixel resolution on protein location pattern interpretation using the 3D HeLa dataset.

2 Methods

2.1 Datasets

The acquisition of the 3D HeLa dataset has been described previously [10]. To create it, antibodies were chosen for 8 different proteins found in 7 subcellular organelles and structures: ER (endoplasmic reticulum protein p63), Golgi (giantin and gpp130), lysosomes (LAMP2), endosomes (transferrin receptor), mitochondria (a mitochondrial inner membrane antigen), nucleoli (nucleolin), and microtubules (tubulin). A secondary antibody conjugated with Alexa 488 dye was then applied. A ninth protein (F-actin, present in microfilaments) was labeled using phalloidin directly-conjugated to Alexa 488. In parallel, total DNA and total cellular protein were labeled using propidium iodide (PI) and Cy5 reactive dye, respectively. The two Golgi proteins were included in the image dataset to test the ability of any classification system to separate protein location patterns which we have shown are essentially indistinguishable to visual inspection [13]. Images were taken with a three-channel confocal laser scanning microscope with lateral pixel resolution of 0.049 μm and axial pixel resolution of 0.203 μm. The 3D HeLa dataset contains 11 different subcellular location patterns with 50 – 52 cells per pattern (558 cells in total). Each cell is represented as a 3-color stack of from 14 to 24 2D slices each consisting of 1024 × 1024 pixels with 8-bit intensities. Example images are shown in Figure 1.

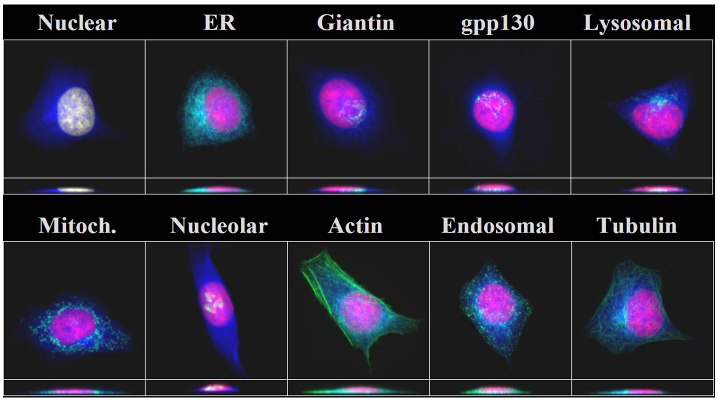

Fig. 1. Typical images from the 3D HeLa image dataset. Red, blue and green colors represent DNA, total protein and target protein staining, respectively. Projections on the X-Y (top) and the X-Z (bottom) planes are shown. Reprinted by permission of Carnegie Mellon University.

The acquisition of the 3D 3T3 dataset has also been described previously [14]. NIH 3T3 cell clones expressing GFP-tagged proteins were generated using CD-tagging techniques [4, 6]. A population of cells was infected with a retrovirus that creates a new GFP exon if inserted into intronic sequence. After the infection, cells expressing GFP-tagged proteins were isolated and subcloned. The identity of the tagged protein was determined by RT-PCR and sequencing. GFP images were taken with a single-channel spinning disk confocal microscope with lateral pixel resolution of 0.11 µm and axial pixel resolution of 0.5 µm. The 3D 3T3 dataset contains 90 different GFP expressing clones with 9 – 33 cells per clone (1554 cells in total). Each cell is represented as a 1-color 1280 × 1024 × 31 stack with 12-bit pixel intensities. Example images are shown in Figure 2.

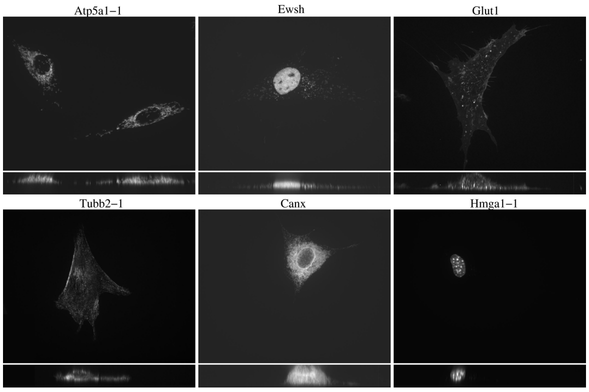

Fig. 2. Selected images from the 3D 3T3 image dataset. Tagged protein names are shown with a hyphen followed by a clone number if the same protein was tagged in more than one clone in the dataset. Projections on the X-Y (top) and the X-Z (bottom) planes are shown. Reprinted from reference 15.

2.2 Image Processing

Before feature extraction, the raw images were first corrected for background fluorescence, segmented into single cell regions, and thresholded. The segmentation for the 3D HeLa dataset was performed using parallel DNA and total protein images and a seeded watershed algorithm as described previously [10]. The segmentation of the 3D 3T3 dataset was achieved by manually defining a cropping mask for each cell since DNA and total protein images were not available.

For experiments on clustering proteins from different sources, the 3D HeLa dataset was down-sampled to the resolution of the 3D 3T3 dataset before feature extraction. When studying effects of pixel resolution, each preprocessed image from the 3D HeLa dataset (after background correction and segmentation) was resized from 2 to 7-fold (corresponding to a reduction in pixel size of from 50% to 14%).

2.3 Feature Extraction

For clustering of images from different data sources, feature set SLF11 [14] was calculated for all images. It consists of 14 morphological features from set SFL9 [10], 2 edge features [14] and 26 Haralick texture features [14, 15]. The 26 texture features in SLF11 consist of mean and range features for the 13 directions of pixel adjacency; the range features were not used in this study. Texture features for both the HeLa and 3T3 datasets were calculated after downsampling to 0.5 μm pixel size and 64 gray levels, which were determined to be the optimal settings for the 3T3 dataset [15].

For assessment of effects of pixel resolutions, feature set SLF19 [19] was calculated for the 3D HeLa dataset for the original images as well as for images downsampled from 2 to 7 fold (downsampling ratios refer to the lateral dimensions; downsampling in the in the axial dimension was only done as needed to match the lateral dimensions). It consists of the 42 SLF11 features plus an additional 14 morphological features calculated with reference to the parallel DNA image. For the original images and the 2-fold down-sampled images, texture features were calculated after additional downsampling to 0.4 μm pixel resolution and 256 gray levels, the optimal settings for the HeLa dataset [17]. For the other down-sampled images, texture features were calculated without further downsampling.

2.4 Feature Transformation Between Cell Lines

To find a set of features and a transformation function that could make HeLa cell features comparable to 3T3 cell features, four sets of subcellular patterns that were expected to be similar between the two cell types were identified. For each set, the mean feature values were calculated for both cell lines (since all values of feature SLF11.28, the average information measure of correlation 1, were negative, we converted them to absolute values). For each feature, we then fit the following transformation:

$$\log(f_{3T3}) = \alpha + \beta \times \log(f_{HeLa}) \tag{1}$$

where α and β are fitted regression coefficients. We chose a simple linear regression model since we have only four data points for each feature. The log form was used since the dynamic range of each feature can be quite large. Once the intercepts and coefficients for each individual feature were determined, we transformed the features calculated for all cells using the formula. Four of the features were eliminated: SLF9.2 (Euler number of objects) had both positive and negative values and could not be log transformed, and SLF9.8, SLF9.17, SLF9.20 were dropped because they were not observed to be discriminatory before transformation.

2.5 Feature Selection

We have previously observed after characterization of a number of methods for feature selection that Stepwise Discriminant Analysis (SDA) performed the best in the domain of subcellular location features [20]. We formed a merged HeLa-3T3 dataset but in which the four classes common to both datasets were joined (as discussed above). This gave a total of 82 classes (90 3T3 classes plus 11 HeLa classes minus 19 3T3 classes that belong to the four common classes). We applied SDA to the 25 features for the 82 classes, resulting in a list of ranked features with descending order of discriminating power. We selected the optimal feature set using classifiers with increasing numbers of the SDA ranked features [15].

A similar procedure was followed for the downsampled HeLa images.

2.6 Implementation of Classification and Clustering Algorithms

A support vector machine was trained and tested with an rbf kernel and max-win strategy and the overall classification accuracy was estimated by 10-fold cross validation.

Two clustering approaches were used as described previously [15]. The first approach started by performing *k*-means clustering using standardized (z-scored) Euclidean distance on all single cell observations with *k* increasing from 2 to either 20 for the resampled 3D HeLa dataset or 101 for the combined dataset (90 3T3 clones and 11 HeLa classes). Akaike Information Content (AIC) was used to select the optimal *k*. For each class, the cluster containing the largest membership was found. If that cluster contained at least 33% of the cells for that class, those cells were retained and the class was assigned to that cluster. Otherwise the class was dropped from further consideration. The second approach started from the cells retained in the first approach. It consisted of building a set of dendrograms using randomly selected halves of the retained dataset, and then constructing a majority consensus tree from the set of dendrograms. AIC was used to select the optimal cutting of the consensus tree into disjoint clusters.

2.7 Creation of Cluster Relationship Maps

Cluster relationship maps were created to visualize the effect of pixel size on the number of distinguishable location classes. The input for create these maps is a list of original classes and a matrix showing which classes are clustered together as a function of degree of downsampling. This matrix was created for the 3D HeLa dataset using *k*-means/AIC clustering as described above. A graph is then created with a node for each cluster found at each level of downsampling. Edges are created between nodes at different levels that contain the same class. In theory this graph could be quite complex, but in practice it was observed for the 3D HeLa data to be composed of distinct cliques. A cluster relationship map is created from the graph in two steps. The class indices are reindexed so that classes in the same clique have continuous indices. Coordinates for plotting are then assigned to each node: the downsampling level as the ordinate and the average index of its members as the abscissa. To facilitate comparison between levels, the overall accuracy of a classifier across all classes at a given downsampling ratio is noted adjacent to the left axis and the accuracy of a classifier trained on just the number of nodes at each level is noted on the right side. Visual clues are also provided to faciliate interpretation of the map. The width of the edge is drawn proportional to the number of classes shared between two clusters. If classes in a lower level (smaller downsampling ratio, higher resolution) node split into multiple nodes at the next higher level (larger downsampling ratio, lower resolution), the numerical IDs of the classes represented by each edge are written on the edge. If all classes at a lower level node still belong to a single node at the next higher level, the numeric IDs are omitted. Using this labeling scheme, membership of each node as well as individual edge could be inferred from the plot without ambiguity.

3 Results

3.1 Feature Conversion and Selection for the Combined HeLa-3T3 Dataset

The 3D HeLa dataset, consisting of 11 distinct location patterns and the 3D 3T3 dataset, consisting of 90 randomly tagged protein clones were derived independently on two different instruments. Table 1 lists the major differences between the two datasets. The two datasets were derived from cell lines that have different cell size, were labeled with different techniques and were imaged with different microscopes and protocols. Consequently, we expect some differences in features calculated for the same location pattern in the two datasets.

Our initial approach was to find an optimal subset of features that are not sensitive to these variations by feature selection rather than transformation (data not shown). Starting from feature set SLF11 (after reduction to 29 features as described in the Methods) and following procedures described previously [15], we obtained a mixed result. Although the nucleolar and nucleus location patterns from 3T3 and HeLa cell lines partitioned into the same clusters, most of the HeLa location patterns formed a distinct cluster that was isolated from the rest of the 3T3 clusters. This result suggested that some (if not most) of the features are dependent on the cell type or methodology used for acquisition.

Table 1. Differences between the datasets used in this study

Name	Labeling Method	Microscopy Method	Objective	Resolution (µm)
3D HeLa	Immunofluorescence	Laser Scanning Confocal	100×	0.049 × 0.049 × 0.203
3D 3T3	CD-tagging	Spinning Disk Confocal	60×	0.11 × 0.11 × 0.5

We therefore sought to find a suitable transformation that could be used to convert features for HeLa cells into their counterparts for 3T3 cells. Given the limited number of classes in common between the two, we sought to fit a linear regression model to the log of the feature values for corresponding pattern classes.

In order to achieve this goal, we started by identifying several sets of proteins that share the same location pattern in the two different cell lines. After visual inspection of the images, we identified four calibration sets, namely cytoskeletal, mitochondrial, nucleolar and uniform (whole cell) patterns. Although both the 3D HeLa dataset and the 3D 3T3 dataset have a nuclear pattern, we did not consider them to be corresponding. This is because the nuclear pattern in the HeLa cell set was generated by DNA staining (Figure 1) while the nuclear pattern in the 3T3 cell dataset was from a specific nuclear protein, Hmga1-1 (Figure 2), and the former pattern shows more uniform staining across the nucleus.

We calculated the mean feature vector for each location pattern of both datasets. For each feature, we used the feature values for the four calibration sets to estimate the intercepts and coefficients for a log linear regression model.

In order to get the optimal subset of features for the combined dataset, we followed the feature selection procedure for clustering described previously [15]. The principle behind this approach is to use increasingly numbers of the features ranked by SDA to train a classifier for all classes and choose the subset that gives the best classification accuracy (recognizing that some classes may not be distinguishable). When we used the optimal feature subset to train a neural network classifier to recognize the 4 combined patterns and the remaining single patterns, the overall classification accuracy was 62%. The classification accuracy for the combined cytoskeletal, mitochondrial, nucleolar and uniform cellular location patterns were 65%, 83% 98% and 97% respectively, suggesting that the classifier could learn decision boundary for both cell types reasonably well (particularly for nucleolar and uniform cellular patterns). The inspection of classification errors within the cytoskeletal class also revealed that the highest error occurred for the HeLa F-actin class, which can be expected since its pattern is similar but not identical to that of tubulin.

3.2 Clustering of the Combined Dataset

A robust set of features would group the protein images based on their location patterns. We tested this hypothesis by using two different approaches to clustering.

The first approach is the *k*-means/AIC algorithm. Using the selected optimal subset of transformed features, we found the minimum AIC was achieved at $k = 31$. However, 14 of these 31 clusters contained only minority images (images from a particular class that were not in the plurality cluster for that class) and were eliminated. This left 17 protein clusters that contained at least 33% of the images for at least one class. The number of classes per cluster ranged from 1 to 12 and the 11 HeLa classes were distributed into 8 clusters.

A parallel approach was to perform consensus clustering as we have described previously [15]. Since we had multiple images for each class in our datasets, we constructed a set of dendrograms in which each was built using a random half of the images for each class. A single majority consensus tree, which contains only the frequently-observed structures, was constructed from the set of dendrograms (Figure 3).

A set of disjoint clusters could be selected from the consensus tree by finding that set of cuts that minimized AIC (these are shown as short vertical lines in Figure 3). Eleven clusters were obtained with 3 to 23 clones per cluster.

Comparison of the clusters obtained in the two approaches revealed that the partitioning obtained by consensus tree analysis is contained in the partitioning by k-means/AIC. In other words, while members of any single cluster from k-means/AIC belong to a single cluster from consensus clustering, and a cluster from consensus clustering contains one or more clusters from k-means/AIC.

Examination of the consensus tree (Figure 3) revealed that those proteins expected to have similar location patterns were grouped together. For example, we consider the DNA pattern in the HeLa dataset to be similar (but, as discussed above, not identical) to the nuclear proteins in 3T3 (two clones of Hmga1 and Unknown-9), and they formed a cluster in the consensus tree even though we did not use this pair to estimate the feature transformation. Since the 3T3 images did not have parallel DNA images, we could not use any features that use this as a reference point (an important aspect of our HeLa cell work). Without being able to calculate position to the nucleus, the Golgi patterns and nucleolar patterns from the two cell types were all considered similar in the tree (both are small, closely placed objects). Tubulin and actin are both cytoskeletal proteins and in the consensus tree, their patterns from HeLa cells form a cluster with a tubulin protein (Tubb2) from the 3T3 dataset.

3.3 Classification of Images at Different Resolutions

As discussed in the introduction, our second goal was to determine the relationship between pattern resolution and image resolution. For this purpose, we downsampled the 3D HeLa images to varying degrees and compared the performance of classifiers trained to distinguish all 10 classes using an optimal feature set selected for each resolution.

The overall classification accuracies across all 10 classes using images resized to different degrees are summarized in Table 2. As the texture features contained information of optimal resolution, we first considered classification on morphological and edge features but not texture features. A clear descending trend was detected with increasing downsampling. While the classifier using original resolution images achieved an overall accuracy of 94.5%, this number quickly dropped below 90% for a minor downsampling (2-fold) and finally dropped below 80% for downsampling of 6 and 7 fold. These results suggested that the discriminant power of our 3D morphological and edge features is modestly sensitive to the pixel resolution of the images.

Inspection of the confusion matrices from these classifiers revealed that the two Golgi proteins (giantin and gpp130) could only be distinguished with confidence at the original resolution and even a two-fold down-sample operation largely eliminated the separation (data not shown) between them. At a 7-fold down-sample ratio, the classification between these two classes was essentially random (data not shown). This result confirmed that these two location patterns are extremely similar and could only be distinguished at very high resolutions.

Previously we have shown that using a combination of morphological, edge and texture features, we could achieve 98% overall classification accuracy in the same HeLa dataset [17]. Therefore we next included DNA and texture features in the

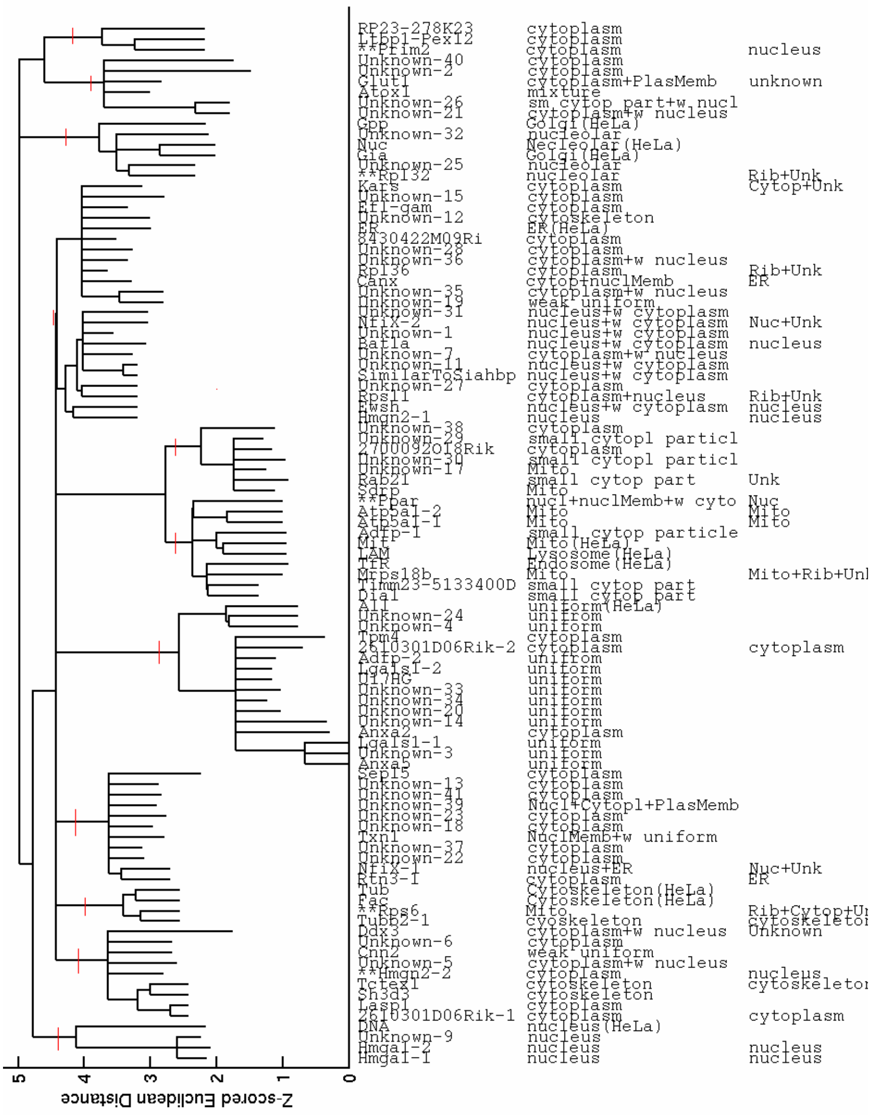

Fig. 3. A consensus Subcellular Location Tree for the combined image dataset using SDA selected features. The columns show the protein names (if known), human observations of subcellular location, and subcellular location inferred from Gene Ontology (GO) annotations. The sum of the lengths of horizontal 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 (**). The vertical lines show the 11 clusters selected by AIC from this tree.

scheme. Firstly, we observed that the inclusion of DNA related features did not further improve the classifier's performance at the original resolution. The same overall classification accuracy (98%) was achieved. Secondly, inclusion of texture features improved the classification performance, confirming the value of texture features in 3D image analysis. Finally, the same descending trend was detected with increasing downsampling ratios (from 98% at original resolution to 82% at 6-fold downsampled images), indicating that the discriminant power of the texture features is also modestly sensitive to the pixel resolution of the images.

3.4 Effects of Pixel Resolution on Cluster Stabilities

As an alternative way to examine what patterns are distinguishable as a function of pixel resolution, we next evaluated the effects of pixel resolution on solutions of clustering approaches.

We partitioned the 10 HeLa classes using both k-means/AIC and consensus tree algorithms at the original resolution as well as 6 downsampled resolutions. Results using only morphological and edge features revealed perfect agreements between the clusters obtained from k-means/AIC algorithm and the clusters obtained from the consensus tree algorithm at each of the 7 resolutions analyzed (data not shown). To graphically show the relationship between patterns and resolution and the stability of clusters, we constructed cluster relationship maps. The goal of these maps is to demonstrate which high-resolution patterns become confused as the resolution is decreased.

Figure 4 shows such a map constructed using only morphological and edge features. As might be expected, the DNA class formed a distinct cluster at all resolutions (as represented by the thin straight line on the left). The two Golgi proteins, giantin and gpp130, formed a cluster separated from all other classes. They were merged at all levels. The LAM class (lysosomal pattern), the Mit class (mitochondrial pattern) and the TfR class (endosomal pattern) formed a single cluster at the original image level, split into two clusters at downsample level 2, and merged again at level 6.

Table 2. Comparison of classification accuracy using images at different resolutions. For images at a specific resolution, the overall classification accuracy using a SVM with 10-fold cross validation and optimal feature subset (with and without texture features) selected by SDA was reported.

Downsampling Factor	**Pixel Size (µm)**	**Classification Accuracy (%)**	
		without Texture Features	**with Texture Features**
1 (None)	0.05 × 0.05 × 0.2	94.5	97.8
2	0.10 × 0.10 × 0.4	88.6	95.8
3	0.15 × 0.15 × 0.6	86.4	92.6
4	0.20 × 0.20 × 0.8	84.0	85.6
5	0.25 × 0.25 × 1.0	81.6	84.9
6	0.30 × 0.30 × 1.2	77.4	82.0
7	0.35 × 0.35 × 1.4	77.9	84.1

Figure 4 clearly revealed that by morphological and edge features we can distinguish 5 basic clusters within the 3D HeLa dataset down to a lateral pixel resolution of 0.35 μm: a cluster of DNA (uniform nuclear pattern), a cluster of giantin and gpp130 proteins (Golgi complex pattern), a cluster of nucleolin (nucleolar pattern), a cluster of ER, F-Actin and Tubulin proteins (cytoplasmic network pattern) and a cluster of Mit, LAM and TfR proteins (cytoplasmic punctate pattern). Although some basic clusters could be separated into sub-clusters at higher resolutions (for example, the separation of Mit and LAM from TfR at downsampling ratios of 2 to 5), this did not happen in a continuous manner.

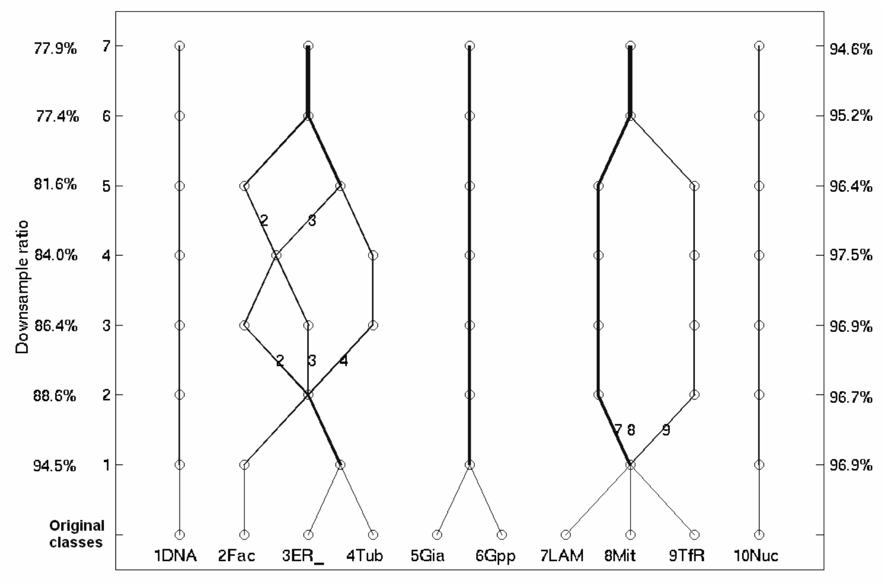

Fig. 4. A cluster relationship map for the 3D HeLa dataset using 3D morphological features (3D SLF 9) and 3D edge features (3D SLF11.15 and 3D SLF11.16). The overall classification accuracies for 10 classes at specific resolutions are shown on the left and the classification accuracies for 5 basic location pattern clusters are shown on the right.

The separation of the 5 basic clusters was further validated by training a classifier to distinguish these 5 clusters. The overall classification accuracy from 10-fold cross validation was labeled on the right side of Figure 4. These 5 clusters could be accurately distinguished with roughly the same accuracy at any resolution.

We repeated this analysis after adding texture features. Results indicated extensive agreement between k-means/AIC algorithm and consensus tree analysis algorithm with some slight differences (mainly relationships among LAM, Mit and TfR patterns). The relationship map using morphological, edge and texture features is shown in Figure 5. It indicated that inclusion of texture features greatly improved the resolution of the clustering solution. Generally at least 8 clusters could be distinguished from the dataset at different pixel resolutions. The stability of these clusters was also

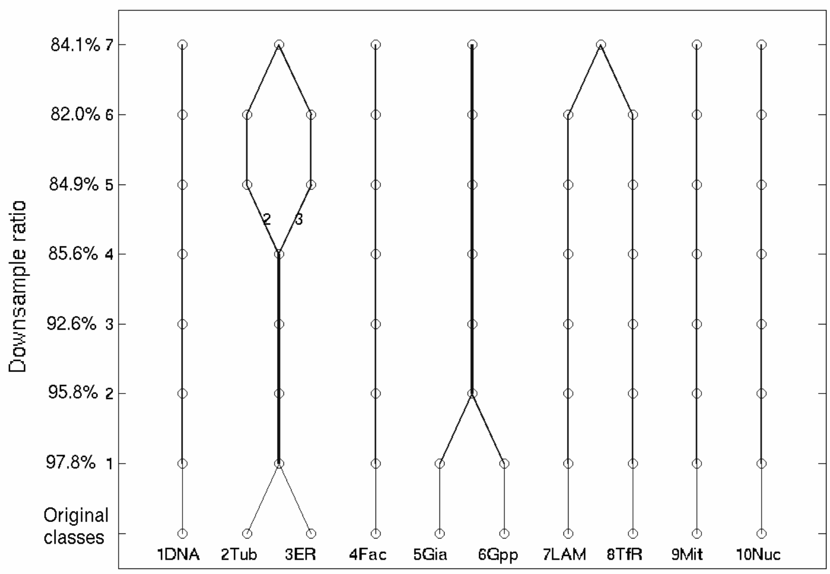

Fig. 5. A cluster relationship map for the 3D HeLa dataset using 3D morphological, edge and texture features. The overall classification accuracies for 10 classes at specific resolutions are shown on the left.

better. For example, the giantin and gpp130 patterns were placed in distinct clusters at the original resolution but were merged into a single cluster at lower resolutions. The LAM and TfR patterns were not merged until the lowest resolution tested.

4 Discussion and Conclusions

We have previously shown that SLFs can be used effectively to numerically describe protein subcellular location patterns using high resolution images in a single cell line. Automated classifiers trained on these features can recognize the location pattern in previously unseen images of HeLa cells with high accuracy. A clustering/partitioning scheme based on SLFs was shown to achieve objective clustering on a dataset of 3D images from 3T3 cells.

The effectiveness of SLFs in different datasets suggested that the SLFs are describing general aspects of the location patterns, which are not dataset dependent. In the current study, we extended our approaches to a multi-source dataset where the input data were from two cell types and two types of microscopes. We proposed a simple method to search for the proper transformation between features from different data sources. Using the transformed features, the partitioning of a combined 3T3 and HeLa dataset was largely based on intrinsic location patterns rather than on the data source. We also studied the effects of image resolution on automated interpretation. Overall classification accuracies for individual classes in the 3D HeLa dataset decreased with decreasing resolution. However, several basic clusters exist and these clusters are insensitive to image resolution, at least down to 0.35 μm pixel size. When texture features were combined with morphological and edge features, the improvement of the resolution as well as the stability of clustering solutions further validated the good discriminant power of texture features.

The results suggest the feasibility of constructing a combined subcellular location tree across many cell types with different resolutions. They also have implications for the collection of future image sets. These are that images of a reasonably large common set of proteins should be acquired for each cell type to be combined, images should be collected at relatively high resolution and that if at all possible parallel images of the DNA distribution should be acquired. The promise of the approach described here is that the similarities and differences of protein distributions in different cell types and different resolutions can be systematically organized and represented.

Acknowledgments. This work was supported in part by NIH grant R01 GM068845 and NSF grant EF-0331657. 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. Cutler, P.: Protein arrays: The current state-of-the-art. Proteomics **3** (2003) 3-18
2. Sali, A., Glaeser, R., Earnest, T., Baumeister, W.: From words to literature in structural proteomics. Nature **422** (2003) 216-225
3. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Welssman, J.S., O'Shea, E.K.: Global analysis of protein localization in budding yeast. Nature **425** (2003) 686-691
4. Jarvik, J.W., Adler, S.A., Telmer, C.A., Subramaniam, V., Lopez, A.J.: CD-Tagging: A new approach to gene and protein discovery and analysis. BioTechniques **20** (1996) 896-904
5. Rolls, M.M., Stein, P.A., Taylor, S.S., Ha, E., McKeon, F., Rapoport, T.A.: A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. J. Cell Biol. **146** (1999) 29-44
6. Jarvik, J.W., Fisher, G.W., Shi, C., Hennen, L., Hauser, C., Adler, S., Berget, P.B.: In vivo functional proteomics: Mammalian genome annotation using CD-tagging. BioTechniques **33** (2002) 852-867
7. Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Alaluf, I., Swerdlin, N., Perzov, N., Danon, T., Liron, Y., Raveh, T., Carpenter, A.E., Lahav, G., Alon, U.: Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nat Methods **3** (2006) 525-531
8. Boland, M.V., Markey, M.K., Murphy, R.F.: Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy. 19th Annu. Intl. Conf. IEEE Eng. Med. Biol. Soc. (1997) 594-597
9. 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 **17** (2001) 1213-1223

10. Velliste, M., Murphy, R.F.: Automated determination of protein subcellular locations from 3D fluorescence microscope images. 2002 IEEE Intl. Symp. Biomed. Imaging (2002) 867-870
11. Danckaert, A., Gonzalez-Couto, E., Bollondi, L., Thompson, N., Hayes, B.: Automated recognition of intracellular organelles in confocal microscope images. Traffic **3** (2002) 66-73
12. Conrad, C., Erfle, H., Warnat, P., Daigle, N., Lorch, T., Ellenberg, J., Pepperkok, R., Eils, R.: Automatic identification of subcellular phenotypes on human cell arrays. Genome Research **14** (2004) 1130-1136
13. Murphy, R.F., Velliste, M., Porreca, G.: Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J VLSI Sig Proc **35** (2003) 311-321
14. 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 **4962** (2003) 298-306
15. Chen, X., Murphy, R.F.: Objective clustering of proteins based on subcellular location patterns. J Biomed Biotechnol **2005** (2005) 87-95
16. Murphy, R.F., Kou, Z., Hua, J., Joffe, M., Cohen, W.W.: Extracting and Structuring Subcellular Location Information from On-Line Journal Articles. IASTED Intl. Conf. Knowl. Sharing Collab. Eng. (2004) 109-114
17. Chen, X., Murphy, R.F.: Robust classification of subcellular location patterns in high resolution 3D fluorescence microscopy Images. 26th Annu. Intl. Conf. IEEE Eng. Med. Biol. Soc. (2004) 1632-1635
18. Murphy, R.F., Velliste, M., Yao, J., Porreca, G.: Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Locations. 2nd IEEE Intl. Symp. BioInf. Biomed. Eng. (2001) 119-128
19. Nair, P., Schaub, B.E., Huang, K., Chen, X., Murphy, R.F., Griffith, J.M., Geuze, H.J., Rohrer, J.: Characterization of the TGN Exit Signal of the human Mannose 6-Phosphate Uncovering Enzyme. J. Cell Sci. **118** (2005) 2949-2956
20. Huang, K., Velliste, M., Murphy, R.F.: Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. Proc SPIE **4962** (2003) 307-318