# Location proteomics - Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins

Xiang Chen, Meel Velliste, Shmuel Weinstein, Jonathan W. Jarvik, and Robert F. Murphy*
Departments of Biological Sciences and Biomedical Engineering, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA, USA 15213

## ABSTRACT

The overall object of proteomics is to characterize all of the proteins expressed in a given cell type. With the rapid development of random gene tagging technology and high resolution fluorescence microscopy, it has become possible to generate libraries of digital images depicting the location patterns of most proteins in any given cell type. While the subcellular location of a protein is important to its function, no established methods exist for the systematic description, comparison or organization of protein location patterns. We have previously described classification methods that accurately recognize all major subcellular location patterns in both 2D and 3D images, as well as methods for rigorous statistical comparison of such patterns. We describe here the application of the numerical features from the previous work to images obtained by random tagging of proteins. Spinning disk confocal microscopy was used to collect images depicting the location patterns of 46 NIH 3T3 cell clones expressing proteins randomly tagged with a fluorescent protein. A set of 42 numerical features describing both image texture and object morphology were calculated and used to build subcellular location trees that group the tagged proteins by similarity of location pattern.
**Keywords:** proteomics, subcellular location pattern, CD-tagging, spinning disk confocal microscopy, green fluorescent protein, subcellular location features, subcellular location trees

## 1.INTRODUCTION

### 1.1 Location Proteomics

Proteomics is the term used to describe large scale documentation and characterization of many or all proteins expressed in a given cell type. It utilizes "mass-screening„ approaches to study global protein expression, structure, location, function, interaction, and other protein characteristics.

Each eukaryotic cell has many subcellular organelles with unique biochemical environments and specific functions. Some specific subcellular structures that are not strictly considered to be organelles, such as the cytoskeleton, also have distinct functions. The properties that a protein displays can be a function of its subcellular location, and thus knowledge of the subcellular location pattern for a protein can provide a critical clue in determining or predicting its function. However, little, if any, annotation is provided for subcellular location patterns in most genomics/proteomics projects where high throughput techniques are used. Currently, terms used for protein locations in literature and databases are descriptive and not suited to determining, for example, whether two proteins within the same organelle have different patterns within that organelle. High resolution comparison of location patterns between proteins is impossible without standardized methods to quantitatively describe these patterns. Therefore, automated classification and clustering methods are necessary to create the field of location proteomics, the branch of proteomics that describes the location pattern of individual proteins and their relationships.

### 1.2 Random tagging of proteins

A critical requirement for location proteomics is a means by which location patterns can be determined for most (or all) of the proteins expressed in a give cell type. While this can be approached by random tagging of cDNA libraries [1-4],

* murphy@cmu.edu; FAX 1 412 268 6571; murphylab.web.cmu.edu



Manipulation and Analysis of Biomolecules, Cells, and Tissues,
Dan V. Nicolau, Jörg Enderlein, Robert C. Leif, Daniel L. Farkas, Editors,
Proceedings of SPIE Vol. 4962 (2003) © 2003 SPIE · 1605-7422/03/$15.00

there are significant advantages of using random tagging of genomic DNA. One advantage is that the expression of a genomic tag is regulated by normal transcriptional controls. Gene trapping methods have therefore been used to analyze subcellular location[5, 6], but this approach relies upon accurate localization of N-terminal fusions. CD-tagging [7-9] is a particularly powerful approach to generating randomly tagged proteins that creates internal fusions with a tag of interest. Genomic CD-tagging involves the random insertion into the genome of a CD-cassette via an engineered retroviral vector. The CD-cassette is flanked by acceptor and donor splicing sequences so that it creates a new exon if it is inserted into an intron. We have used a CD-cassette containing the coding sequence of a green fluorescent protein (GFP) to generate a number of clones of NIH 3T3 cells expressing GFP-tagged fusion proteins [9]. By recovering and sequencing mRNA and/or DNA fragments containing the CD-tag, the tagged proteins can be identified (if a corresponding sequence is present in the current genome databases).

### 1.3 Subcellular location features and feature selection

We have previously developed sets of numerical features (Subcellular Location Features, SLF) to describe protein subcellular location patterns in fluorescence microscope images [10, 11]. The features used include Zernike moment features, morphological features (features derived from morphological processing of the images), and Haralick texture features. Starting with many such features, we have used feature selection methods to create sets that eliminate redundant features. Of the many different methods we have tested, stepwise discriminant analysis (SDA) was found to give the best performance[12].

### 1.4 Classification of 2D images

In order to validate these subcellular location features, several classification methods have been implemented and tested on sets of images of Chinese hamster ovary cells[13] and HeLa cells[10, 11, 14]. The best performance achieved by these classifiers for images without a reference DNA image was 86% accuracy on previously unseen images over 10 major subcellular location patterns using a back-propagation neural network (BPNN) with a single hidden layer of 20 nodes and the SLF-8 feature set[11]. When a reference DNA image is included, the accuracy increases to 88%.

### 1.5 Classification of 3D images

Current fluorescent microscopes can readily provide a stack of 2D images of the same cell to form a 3D image. It is expected that the information content of 3D images is greater than that of 2D images, and one would anticipate that classifiers trained on 3D images combined with a suitable feature set should be more accurate compared to those trained on 2D images. This was found to be the case, since a BPNN classifier achieved an accuracy of 91% on 3D images of the 10 major location classes using only a set of morphological features[15].

### 1.6 Goal of the current project

Although the classifiers we have developed can classify previously unseen examples of major protein location patterns with high accuracy, they cannot recognize patterns not used during their training. When images of large numbers of randomly tagged proteins are considered, it is unlikely that all possible patterns will be included in the set of training classes. Therefore, the ultimate goal of the current project is to develop approaches for constructing an optimal, hierarchical grouping of proteins based on the similarity of their subcellular location patterns. We have previously coined the term Subcellular Location Tree (SLT) to describe such groupings[11] and have applied cluster analysis to group the 10 major location patterns in the 2D images used in our previous work. The goal of the work described here is to find an appropriate approach to generate an SLT for the high resolution 3D protein location patterns of randomly tagged proteins obtained via CD-tagging.

## 2. METHODS

### 2.1 Source of images

Cell lines expressing GFP-tagged proteins previously generated by CD-tagging [9] were used. Cell lines were derived by infection of NIH 3T3 cells with a retrovirus carrying a GFP coding sequence followed by fluorescence microscope examination to identify GFP-expressing clones. Lines were subcloned until they appeared to contain only a single location pattern. In most cases, only a single GFP-containing sequence was detectable by RT-PCR.

3D fluorescence microscope images of 46 tagged clones were taken using a 60x 1.4 NA objective on an Olympus IX50 microscope outfitted with a Nipkow disk confocal imaging system (Solamere Technology Group). The system consisted of a Yokogawa CSU10 Confocal Scanner Unit illuminated via optical fiber from a LaserPhysics Reliant 100s-488 Argon ion laser. Images were captured by a Roper Scientific/Photometrics CoolSnap HQ Cooled CCD Camera, and image capture was controlled by QED software (QED Imaging, Pittsburgh, PA, USA) on an Apple PowerMac G4 computer. The pixel spacing in both directions in the image plane was 0.11 μm and the vertical spacing between adjacent planes (slices) was 0.5 μm. The resulting 1280 x 1024 x 31 3D images each contained from 1 to 3 cells (or partial cells). A total of 933 images containing 989 full cell images were obtained, with each clone having from 16 to 33 cell images.

**2.2 Morphological feature extraction**

Before extraction of morphological features, the raw images were processed as described previously[15] with the exception that single cells in an image were defined using a manually created cropping mask. The background fluorescence was subtracted and all pixels with a value below an automated calculated threshold were set to zero. Fluorescent objects, defined as sets of connected (26-nearest-neighbor) above-threshold pixels, were extracted and used to calculate 14 of the SLF9 features described previously[15]:

3D-SLF9.1 Number of objects in the cell
3D-SLF9.2 Euler number of the cell
3D-SLF9.3 Average object volume (average number of above threshold pixels per object)
3D-SLF9.4 Standard deviation (SD) of object volumes
3D-SLF9.5 Ratio of maximum object volume to minimum object volume
3D-SLF9.6 Average object to center of fluorescence (COF) distance
3D-SLF9.7 SD of object to protein COF distances
3D-SLF9.8 Ratio of maximum to minimum object distance from COF
3D-SLF9.15 – 17: Absolute value of the horizontal (XY plane) component of 3D-SLF9.6 – 8
3D-SLF9.18 – 20: Signed vertical (Z) component of 3D-SLF9.6 – 8

**2.3 Edge feature extraction**

In addition to these features, we defined additional features to create a new set, SLF11 (3D-SLF11.1 through 11.14 were defined to be the 14 3D-SLF9 features listed above). Since we have previously observed that features based on edge detection are useful in classifying 2D images[10], we calculated comparable 3D features. Since the pixel spacing in the vertical direction (0.5 μm) is different from that in the horizontal direction (0.11 μm), the edge features are only calculated along the horizontal dimensions. Briefly, edges were detected in the thresholded cell image by the Sobel method using the `edge` function in the Image Processing toolbox (version 2.2.2 release 12) of Matlab (version 6.0.0.88 release 12) for each slice in a 3D image and used to reconstruct the edges for the 3D image. The following 2 features were then calculated:

3D-SLF11.15 The fraction of above threshold pixels that are along an edge
3D-SLF11.16 The fraction of fluorescence in above threshold pixels that are along an edge

**2.4 Haralick texture feature (3D extension) extraction**

Haralick[16, 17] described a set of 14 statistics based on the gray-level co-occurrence matrix to measure the texture of 2D images, and we have previously included 13 of these among our SLF (we have not used the maximal correlation coefficient due to computational instability). The 3D equivalents of these 13 Haralick texture features were calculated using the algorithms described[16] with the following modifications. In a 3D image, adjacency can occur in each of 13 directions (compared to 4 directions in a 2D image) and there are 13 gray-level co-occurrence matrices accordingly. Therefore, we generated two types of 3D texture features, the average over all 13 directions, and the range between maximum and minimum over all 13 directions:

3D-SLF11.17/30 Average/range of angular second moment
3D-SLF11.18/31 Average/range of contrast
3D-SLF11.19/32 Average/range of correlation
3D-SLF11.20/33 Average/range of sum of squares of variance
3D-SLF11.21/34 Average/range of inverse difference moment
3D-SLF11.22/35 Average/range of sum average
3D-SLF11.23/36 Average/range of sum variance

3D-SLF11.24/37 Average/range of sum entropy
3D-SLF11.25/38 Average/range of entropy
3D-SLF11.26/39 Average/range of difference variance
3D-SLF11.27/40 Average/range of difference entropy
3D-SLF11.28/41 Average/range of info measure of correlation 1
3D-SLF11.29/42 Average/range of info measure of correlation 2

## 2.5 Exclusion of outliers

Significant time (and data storage) is required for the collection of high resolution 3D images, limiting the number of images that can reasonably be collected for a large number of clones. In order to avoid observer bias during collection, interphase cells were chosen essentially at random without filtering to remove unusual patterns. As a result, the image sets for each clone often contained images that were not typical of the set as a whole (these may have corresponded to dead or dying cells, cells just before mitosis or just after cytokinesis, or images with abnormally high intensity pixels due to camera defects). In our previous work in which 50-100 images were collected for each protein, these outliers were outweighed by the large number of normal images (they frequently corresponded to the small percentage of images that were not correctly classified). In the current work, however, only 16-33 images per class were available, and the presence of even a single outlier could dramatically skew calculations of average feature values for a given set. A stringent procedure was therefore used to remove outliers from the set of images for each clone. For each feature calculated, either the Q test (for less than 10 cells per clone), or the simplified t-test (for 11 or more cells per clone, a value was deemed as an outlier when it was more than 3 standard deviations away from the mean) was recursively performed. If for any feature a cell was found to be an outliner, it was excluded from further analysis.

After the exclusion, 660 cells remained in all 46 clones, with from 9 to 22 cells per clone.

### 2.6 Feature selection

The presence in a feature set of irrelevant features or "masking variables,, as defined by Fowlkes and Mallows[18] can dramatically affect the results of cluster analysis. Therefore, only those features which support the discrimination of the clusters should be included in cluster analysis. Based on our previous results, we used stepwise discriminant analysis (SDA) for this purpose. The `stepdisc` function of SAS (SAS Institute, Cary, NC) was used with default parameter values. The input to SDA was the full feature matrix for all non-outlier cells for all clones and the output was a ranked list of features that were considered to contribute to distinguishing the clones at a confidence level of 0.15.

### 2.7 Feature normalization and distance function

In order to cluster the clones based on their location patterns, some measure of the similarity (or distance) between patterns is required. To avoid excessive weighting of features whose absolute values happen to be larger than the other features, we first normalized all features to z-scores (subtracting the mean and dividing by the standard deviation, both calculated across all clones). Euclidean distances were then calculated between each pair of clones using the normalized features.

### 2.8 Clustering algorithms

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm was used to construct a distance tree (dendrogram) based on the calculated distance matrix (using the `linkage` function from version 3.0 release 12 of the Statistics Toolbox of Matlab).

### 2.9 Back Propagation Neural Networks

A single hidden layer BPNN with 20 hidden nodes was implemented using the Netlab toolbox for Matlab as described previously[10], with the following modifications. The images for each clone were randomly divided into three subsets: a test set consisting of one cell for clones with less than 15 cells or two cells otherwise, a training set consisting of two thirds of the remaining cells, and a stop set consisting of the rest of the cells (training was stopped when the error on the stop set reached a minimum). The random division and training process was repeated 20 times and the results averaged over all trials.

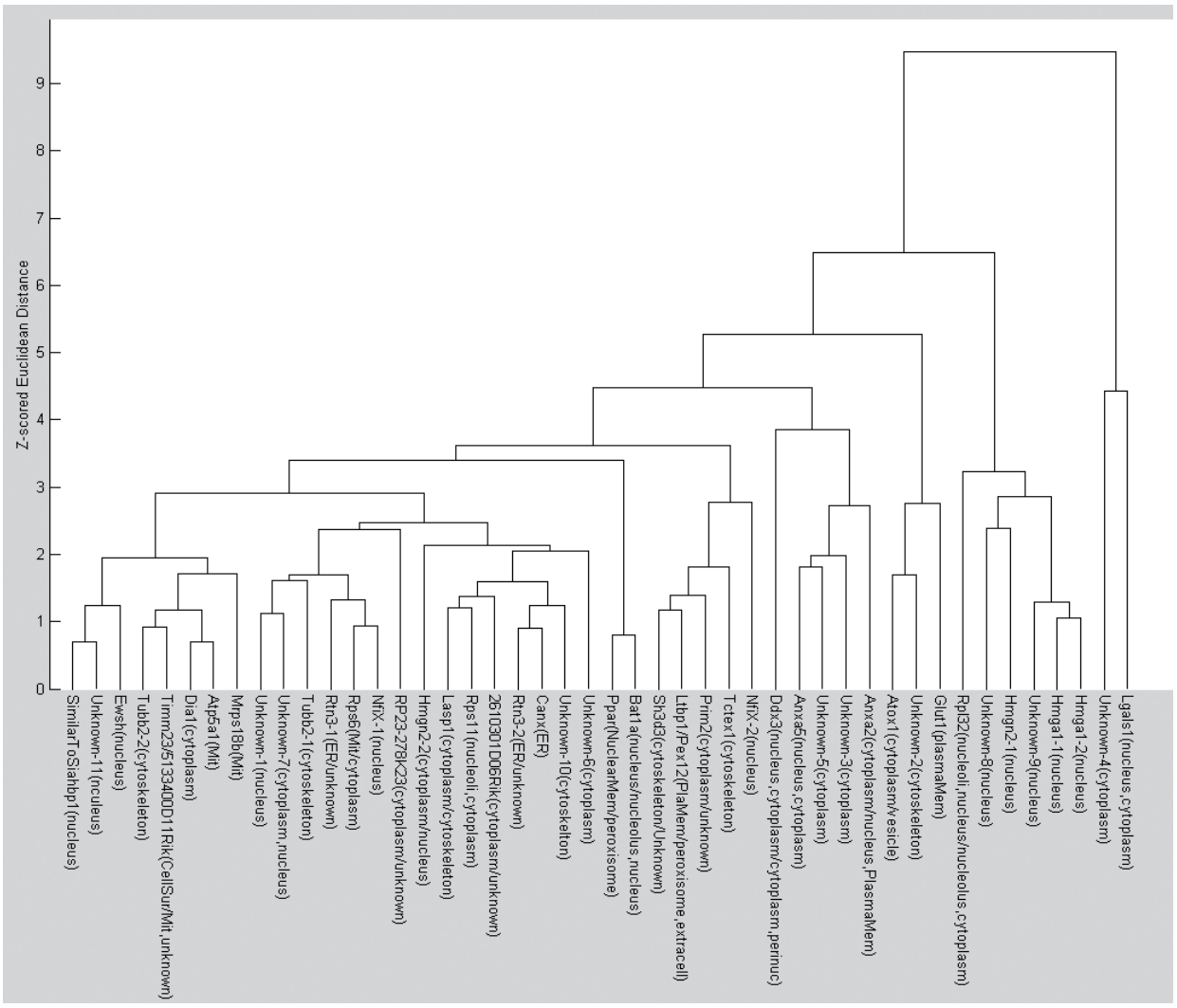


Figure 1: Subcellular Location Tree for the CD-Tagging data set using 14 of the 3D-SLF9 features. The gene name is shown for each clone (if known), followed by the location pattern assigned by human observation, followed by a slash (/) and the location pattern determined from literature searching, if different.

## 3.RESULTS

### 3.1 Clustering based on morphological features

As a first approach to building an SLT for the 46 clones obtained by CD-Tagging, we constructed a dendrogram using the 14 3D-SLF9 features described in section 2.2 and a z-scored Euclidean distance function (Figure 1). In the tree there are two clusters that contain most of the nuclear proteins while the remaining nuclear proteins are distributed among other clusters. We examined images from the two clusters and confirmed that there is a difference in pattern between them. While the cluster consisting of Hmga1-1, Hmga1-2, Unknown-9, Hmgn2, and Unknown-8 contains proteins that are exclusively located within the nucleus, the other cluster (Ewsh/Unknown-11/SimilarToSiahbp1) shows a location pattern consisting of a primarily nuclear pattern combined with some cytoplasmic distribution (example images from each cluster are shown in Figure 2).

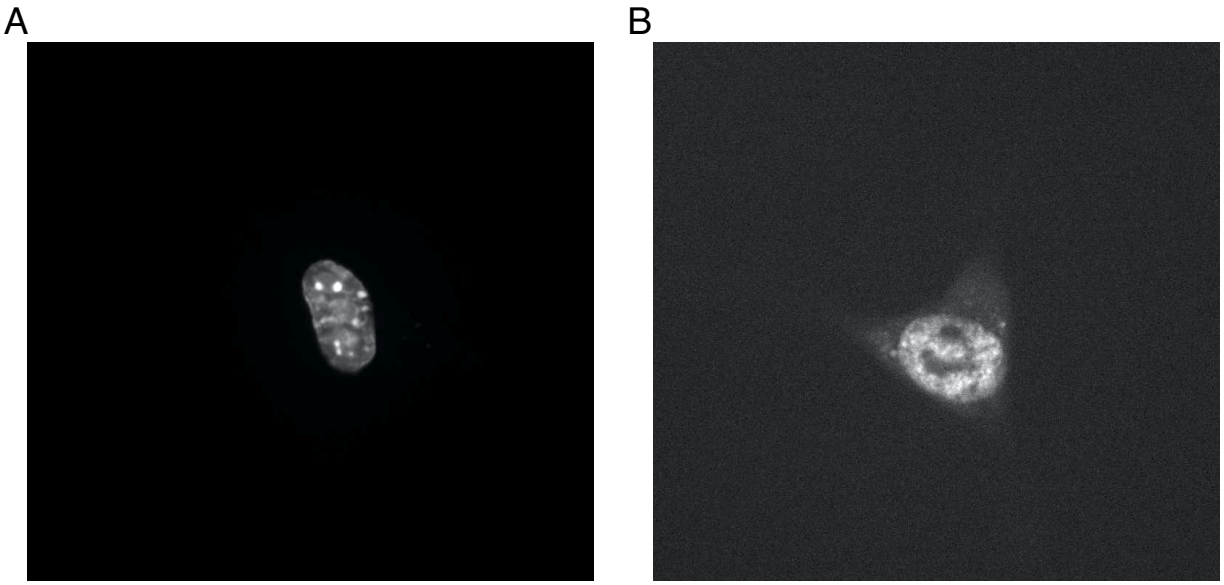

Figure 2: Representative images of clones from the two major clusters of nuclear proteins in Figure 1. (A) Hmga1-1; (B) Unknown-11

A traditional difficulty with interpreting dendrograms such as the one in Figure 1 is in determining which branches of the tree represent equivalent classes. As a criterion for merging such branches we considered whether they could be distinguished by a classifier trained on all clones.

**3.2 Number of inherent clusters**

We therefore trained BPNN classifiers using the same 14 3D-SLF9 features for all 46 clones. An average accuracy of 40% was obtained. This low accuracy reflects the fact that some clones cannot be distinguished from each other. By examining the resulting confusion matrix (not shown), we observed that branches in Figure 1 separated by a z-scored Euclidean distance of 2.8 or less were largely indistinguishable from each other. Using that value, the 46 clones in Figure 1 form 12 clusters. An independent analysis using k-means clustering coupled with the Akaike Information Criterion on the same data set confirmed this estimate of the number of clusters (unpublished data). New BPNN classifiers were therefore trained in which each clone was labeled with its cluster number. The average accuracy improved from 40% for all 46 clones to 71% for the 12 clusters.

### 3.3 Expansion and selection of features

Even with this improvement, the results suggest that the morphological features in 3D-SLF9 do not sufficiently describe the location patterns in the 46 clones. We therefore implemented additional features, 2 edge features and 26 3D texture features, to create the 3D-SLF11 feature set. However, when we trained classifiers for all 46 clones using all 42 of these features, the accuracy dropped dramatically (data not shown), suggesting that the full set contained noisy or uninformative features or that the training had become underdetermined due to the increased number of weights associated with the new features.

The features were therefore ranked by SDA in decreasing order of their ability to distinguish the classes, and BPNN were trained for all 46 classes using various numbers of these ranked features. Figure 3 reveals that the best performance (70%) was achieved using the best 14 of the features (3D-SLF11.16, 3D-SLF11.23, 3D-SLF11.26, 3D-SLF11.33, 3D-SLF11.41, 3D-SLF9.5, 3D-SLF11.42, 3D-SLF11.29, 3D-SLF11.22, 3D-SLF11.18, 3D-SLF11.34, 3D-SLF11.40, 3D-SLF11.38, 3D-SLF11.32). The classifier achieved a comparable accuracy (68%) using as few as the first 10 features.

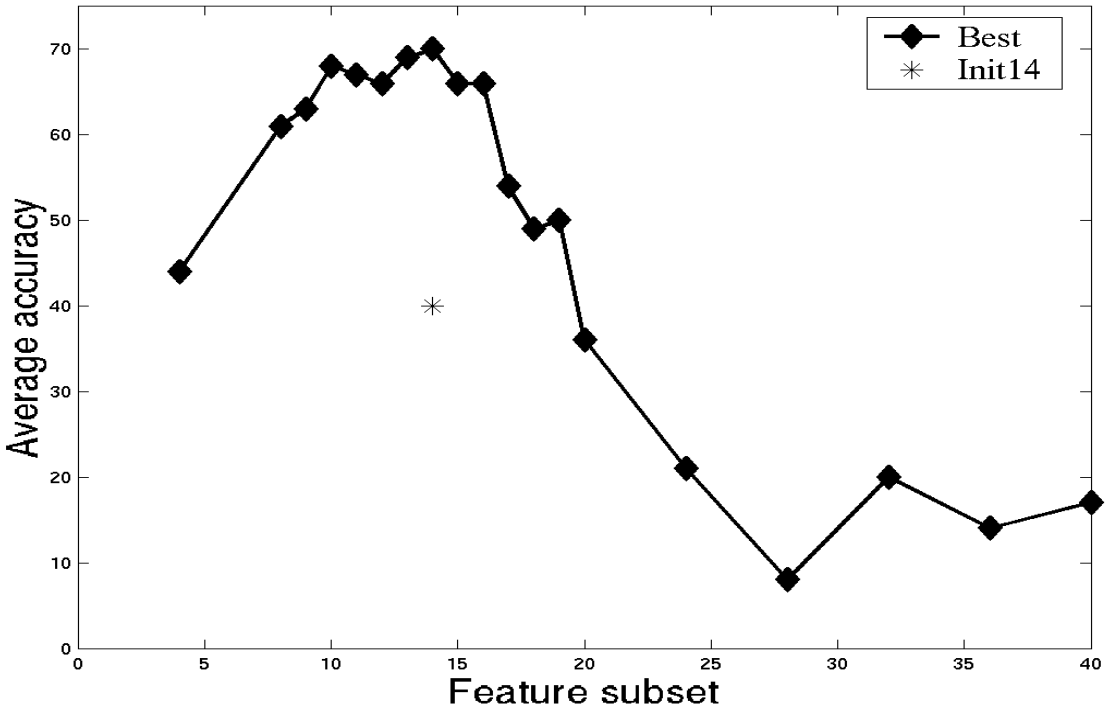


**Figure 3:** Average BPNN performance using different subsets of 3D-SLF11. A BPNN with 20 hidden nodes was trained on all 46 clones and tested on one or two images from each clone. The test was repeated 20 times and the average prediction accuracy over all 46 clones is displayed as a function of the number of ranked features used during training.

Given the strong classification accuracy with only ten features, we calculated a new SLT using this feature set (Figure 4). Again there are two major clusters of nuclear proteins. The inclusion of Unknown-7 (which was classified as both nucleus and cytoplasm by human observation) in Cluster 2 of this SLT further validates the distinction drawn previously between these two clusters.

## 4. DISCUSSION

Our results illustrate the utility of generating Subcellular Location Trees to group subcellular location proteins in large collections of proteins. The approach properly grouped sets of similar nuclear patterns with subtle differences. It also provided a means for objectively describing the patterns of previously unknown proteins relative to known proteins. For example, Unknown-3 was found to be very similar to Lgals1, Unknown-7, Unknown-11 and Unknown-1 to SimilarToSiahbp1, and Unknown-5 to Canx.

The 14 feature and 10 feature subsets that gave the best classification performance contain mostly texture features, with one edge feature and one morphological feature. This is a somewhat surprising outcome since BPNN classifiers trained on morphological features alone achieved 91% accuracy on 3D HeLa images when a parallel DNA image was available and 80% without a parallel DNA image[15]. It suggests that the texture features might provide further improvement in the performance of BPNN on the 3D HeLa data set.

We expect the methods described here to be generally useful for creating a systematic representation of protein location for any database of expressed proteins, such as the recently established collection of yeast protein patterns[19].

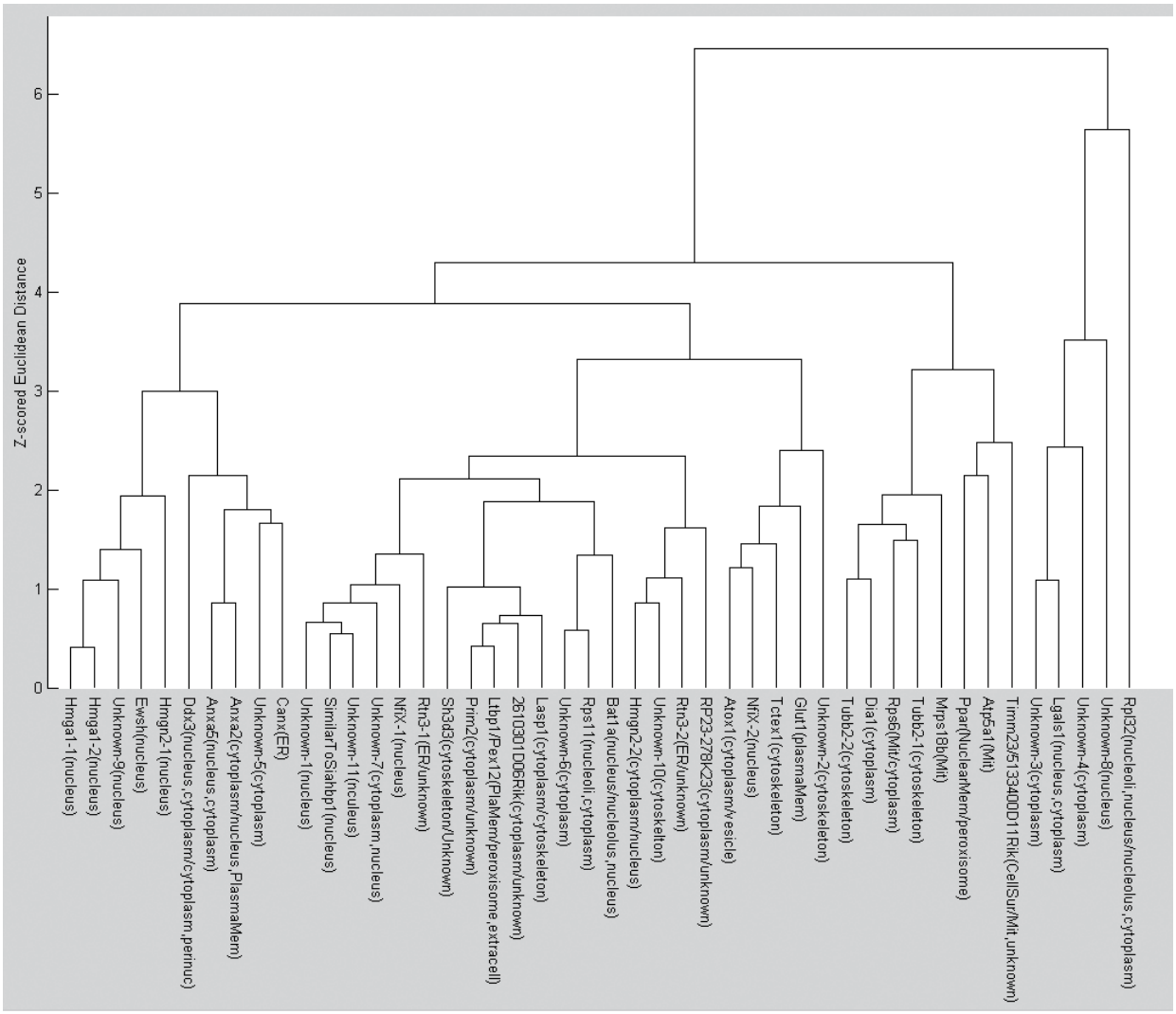

Figure 4: Subcellular Location Tree for the CD-Tagging data set using the SLF11/SDA10 feature set.

## ACKNOWLEDGMENTS

This work was supported in part by NIH grant R33 CA83219 and by a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund. X.C. 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. M.M. Rolls, P.A. Stein, S.S. Taylor, E. Ha, F. McKeon, and T.A. Rapoport, "A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein," *Journal of Cell Biology*, **146**, 29-44, 1999.
2. K. Misawa, T. Nosaka, S. Morita, A. Kaneko, T. Nakahata, S. Asano, and T. Kitamura, "A method to identify cDNAs based on localization of green fluorescent protein fusion products," *Proceedings of the National Academy of Sciences, USA*, **97**, 3062-3066, 2000.

3. S.R. Cutler, D.W. Ehrhardt, J.S. Griffitts, and C.R. Somerville, "Random GFP::cDNA fusions enable visualization of subcellular structures in cells of Arabidopsis at a high frequency," *Proceedings of the National Academy of Sciences, USA*, **97**, 3718-3723, 2000.
4. J.C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann, "Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing," *EMBO Reports*, **1**, 287-292, 2000.
5. P. Tate, M. Lee, S. Tweedie, W.C. Skarnes, and W.A. Bickmore, "Capturing novel mouse genes encoding chromosomal and other nuclear proteins," *Journal of Cell Science*, **111**, 2575-2585, 1998.
6. H.G. Sutherland, G.K. Mumford, K. Newton, L.V. Ford, R. Farrall, G. Dellaire, J.F. Caceres, and W.A. Bickmore, "Large-scale identification of mammalian proteins localized to nuclear sub-compartments," *Human Molecular Genetics*, **10**, 1995-2011, 2001.
7. J.W. Jarvik, S.A. Adler, C.A. Telmer, V. Subramaniam, and A.J. Lopez, "CD-Tagging: A new approach to gene and protein discovery and analysis," *BioTechniques*, **20**, 896-904, 1996.
8. C.A. Telmer, P.B. Berget, B. Ballou, R.F. Murphy, and J.W. Jarvik, "Epitope Tagging Genomic DNA Using a CD-Tagging Tn10 Minitransposon," *BioTechniques*, **32**, 422-430, 2002.
9. J.W. Jarvik, G.W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P.B. Berget, "In vivo functional proteomics: Mammalian genome annotation using CD-tagging," *BioTechniques*, **33**, 852-867, 2002.
10. M.V. Boland and R.F. Murphy, "A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells," *Bioinformatics*, **17**, 1213-1223, 2001.
11. R.F. Murphy, M. Velliste, and G. Porreca, "Robust Classification of Subcellular Location Patterns in Fluorescence Microscope Images," *Proceedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP 12)*, pp. 67-76, 2002.
12. K. Huang, M. Velliste, and R.F. Murphy, "Feature Reduction for Improved Recognition of Subcellular Location Patterns in Fluorescence Microscope Images," *Proceedings of the SPIE*, **4962**, in press, 2003.
13. M.V. Boland, M.K. Markey, and R.F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry*, **33**, 366-375, 1998.
14. R.F. Murphy, M.V. Boland, and M. Velliste, "Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images," *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 251-259, San Diego, 2000.
15. M. Velliste and R.F. Murphy, "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002)*, pp. 867-870, Bethesda, MD, USA, 2002.
16. R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-3**, 610-621, 1973.
17. R.M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, **67**, 786-804, 1979.
18. E. Fowlkes and C. Mallows, "A method for comparing two herarchical clusterings," *Journal of the American Statistical Association*, **78**, 553-584, 1983.
19. G. Habeler, K. Natter, G.G. Thallinger, M.E. Crawford, S.D. Kohlwein, and Z. Trajanoski, "YPL.db: the Yeast Protein Localization database," *Nucleic Acids Research*, **30**, 80-83, 2002.