AUTOMATED INTERPRETATION OF SUBCELLULAR LOCATION PATTERNS

Robert F. Murphy

Departments of Biological Sciences and Biomedical Engineering and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

ABSTRACT

Fluorescence microscopy is widely used to analyze the distribution of proteins within cells. As currently practiced, the assignment of a protein to a particular organelle is done by visual inspection of images or comparison between the distribution of the unknown protein and markers with known location patterns. In order to use fluorescence microscopy for large scale or proteome-wide analysis of protein location, improved approaches that are more automated, objective, and sensitive are needed. Our group has therefore developed automated systems that can recognize the major subcellular location patterns in images of single cultured cells, and has shown that these systems are more sensitive than visual inspection. The foundations of these systems are sets of numerical features that describe the essential characteristics of a subcellular pattern without being overly sensitive to the size, shape and orientation of that cell within the field of view. These features can be used to measure the similarity between protein patterns and therefore for the first time to group proteins in an objective manner based on their high-resolution patterns.

1. INTRODUCTION

The subcellular location (or locations) of each protein is a critical property that provides the context in which that protein carries out its functions. While large scale, comprehensive efforts to catalog other protein properties (such as sequence, structure, binding partners, and enzymatic activities) have been mounted, subcellular location has received far less attention. Until recently, the only information on subcellular location in protein databases was in the form of unstructured text. This made determination of whether two proteins had similar location patterns difficult. A laudable effort by the Genome Ontology (GO) Consortium to create a standardized vocabulary (or ontology) for subcellular location (the GO Cellular Component Ontology) has partially addressed this problem, as databases have added references to GO terms. However, as shown in Table 1, it remains difficult to use these terms to determine the extent of similarity of the subcellular patterns of two proteins. While it is clear that the two proteins are both Golgi proteins, it is unclear whether the differences in terms applied to the two should be interpreted to mean that the two proteins would be expected to reside in distinct regions of the Golgi, whether they would partially overlap, or whether they would be identical. This problem is even greater when trying to determine the similarity of the patterns of proteins found in distinct organelles. For example, based on GO terms, how similar would we expect the pattern of proteins described by the GO term for lysosomes be to the pattern of proteins described by the GO term for endosomes? If comparison of entries for similar proteins suggests the possibility that protein is incorrectly described, is it more likely that its entry describing a pattern as mitochondrial is actually endosomal or lysosomal? Spatial similarity is simply not captured by these terms. What is needed is an approach that can describe the location patterns of all proteins in quantitative rather than qualitative terms.

Table 1. Comparison of terms used to describe subcellular location in protein databases.

Protein	giantin	gpp130
Accession	Swiss-Prot Q14789	TrEMBL O00461
Comments: Subcellular location	Golgi; membrane-associated	(none)
GO Cellular component terms	0000139, Golgi membrane; 0005795, Golgi stack; 0016021 integral to membrane	0030139, endocytic vesicle; 0005801, Golgi cis-face; 0005796, Golgi lumen; 0016021, integral to membrane

2. SUBCELLULAR LOCATION FEATURES

Fluorescence microscopy is the most frequently used method for determining subcellular location, with current practice involving visual examination of images or comparison of an unknown protein to one or more marker proteins (proteins that are known to localize to a particular organelle). The widespread availability of fluorescence microscope systems that produce digital images provides

0-7803-8388-5/04/$20.00

Table 2. Informative features for 2D subcellular pattern classification, organized by feature type. The number of features of each type present in each set is shown. The average accuracy of a classifier trained with each set is also shown (neural network for SLF8 and SLF13, majority-voting ensemble for SLF16).

	2D SLF set			
Feature type	SLF7	SLF8	SLF13	SLF16
Object size	5	4	4	4
Object distance	3	2	2	3
Non-obj. fluorescence	1	1	1	1
Skeleton	5	3	2	2
Edge	5	3	3	3
Hull	3	0	0	0
Zernike	49	7	6	6
Haralick texture	13	11	9	10
Gabor	-	-	-	11
Daubechies	-	-	-	3
DNA	-	-	4	4
No. features	84	31	31	47
Avg. classif. acc.	n/a	86%	88%	92%
Reference	[1]	[1]	[1]	[3]

Table 3. Confusion matrix for 2D HeLa cell images using an optimal majority-voting ensemble classifier and feature set SLF16. The average accuracy is 92%. The probes used to define the classes were directed against DNA (D), an endoplasmic reticulum protein (E), the Golgi proteins giantin (gi) and gpp130 (gp), the lysosomal protein LAMP2 (L), a mitochondrial protein (M), the nucleolar protein nucleolin (N), actin (A), the endosomal protein transferrin receptor (Tf), and tubulin (Tu). Due to rounding, the percentages in each row may not sum to 100. (Data from reference [3].)

	D	E	gi	gp	L	M	N	A	Tf	Tu
D	99	1	0	0	0	0	0	0	0	0
E	0	97	0	0	0	2	0	0	0	1
gi	0	0	91	7	0	0	0	0	2	0
gp	0	0	14	82	0	0	2	0	1	0
L	0	0	1	0	88	1	0	0	10	0
M	0	3	0	0	0	92	0	0	3	3
N	0	0	0	0	0	0	99	0	1	0
A	0	0	0	0	0	0	0	100	0	0
Tf	0	1	0	0	12	2	0	1	81	2
Tu	1	2	0	0	0	1	0	0	1	95

an opportunity to automate the process of determining subcellular location and thereby make it objective and quantitative. Towards this end, our group has developed systems for recognizing the major subcellular structures in images of individual cultured cells [5, 6]. The difficulty of this problem lies in the fact that many cell types show highly variable shape (especially in culture) and that organelles do not have fixed locations within cells. This makes it difficult to design systems that recognize a pattern by direct pixel-by-pixel comparison with either a library of images or a model (either rigid or deformable) of a subcellular pattern. The alternative is to use numerical features to describe the characteristics of a pattern in a shape and orientation-independent manner and use these features to train classifiers. We have therefore described and evaluated a large number of features of various types for this purpose [1, 6, 7]. The evaluation has been done primarily using collections of 2D [6] and 3D [8] images of the patterns of 10 or 11 probes in HeLa cells, which contain 50-100 images of single cells for each probe. By training classifiers with a subset of these images and then measuring the performance on test images not used for training, the utility of specific feature sets (and classification approaches) could be evaluated. To facilitate reference to features found to be useful, we term them Subcellular Location Features and define them in sets used for various purposes. Some of these sets are large general purpose collections (such as SLF7 which contains 84 features) and others are smaller sets that have been selected to discriminate particular sets of patterns (such as SLF8 which was selected from SLF7 using Stepwise Discriminant Analysis to find only features which were useful for classifying ten patterns in 2D images of HeLa cells). The features we have used include ones derived from morphological image processing (including features measuring object size, the distance of objects from some point of reference, the amount of fluorescence in discernible objects, and object skeletonization), edge detection, convex hull finding, decomposition using Zernike polynomials, texture analysis, wavelet transforms, and comparison to a parallel DNA image. Table 2 shows the number of features of various types in the base feature sets for 2D (SLF7) and 3D (SLF9) images, as well as the number of features selected from them and the average accuracy of resulting classifiers. An important conclusion from Table 2 is that the optimal feature sets include features of many types (we have also shown that approaches using a single feature type, such as texture features, do not perform as well).

Results for a neural network classifier using SLF16 are presented in Table 3. The results are in the form of a confusion matrix, in which the value in each cell represents the percentage of test images from the class shown in the row heading that were assigned by the network to the class shown in the column heading. The percentage of each class that is correctly classified is shown along the diagonal. The results show that all major structures can be distinguished with an overall accuracy of 92%, far better than expected at random.

3. COMPARISON WITH VISUAL ANALYSIS

Of particular importance was the finding that the automated system could distinguish two Golgi proteins, Giantin and gpp130, that were originally included in the test because they could not be distinguished by visual inspection. They could be distinguished an average of 75% with a neural network and SLF13 [1], and this improved to 87% using a majority-voting ensemble and SLF16 (Table 3). Examples of the images of these two proteins are shown in Figure 1 to illustrate how similar they are.

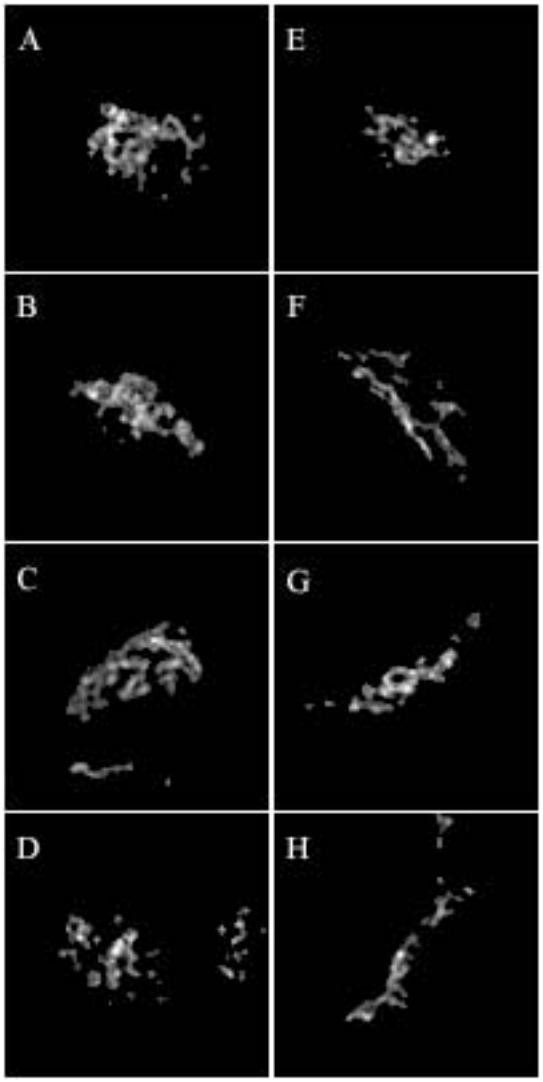

Figure 1. The patterns of the Golgi proteins giantin and gpp130 are visually indistinguishable by fluorescence microscopy. The four most typical images of giantin (A-D) and gpp130 (E-H) were selected from the collection of 2D HeLa cell images using TypIC [2]. (Data from reference [4].)

To test the hypothesis that they could not in fact be distinguished visually, we carried out training and testing using a human subject [1]. The results indicated that while the human observer could correctly classify the other patterns with reasonable accuracy and could identify both giantin and gpp130 images as being from the Golgi, the two proteins could not be distinguished from each other at all (the results were the same as expected for random guesses). A comparison of the accuracy of automated and visual classification is shown in Figure 2. Note that the performances are similar for 7 of the ten classes but that the computer does significantly better not only for the Golgi proteins but also for the lysosomal protein LAMP2.

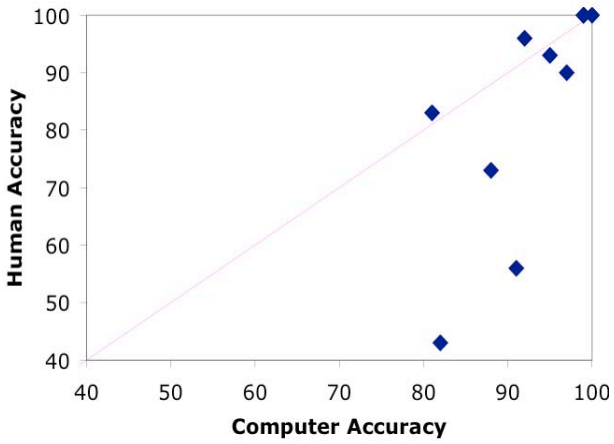

Figure 2. Comparison of classification accuracies from an automated system and from visual examination. Accuracies from Table 3 using SLF16 and a majority-voting ensemble classifier [3] are presented versus the average accuracy for the same images obtained by visual examination [1]. Each symbol represents a different pattern class. In increasing order of human classification accuracy these are: gpp130, giantin, LAMP2, TfR, ER, Tubulin, Mitochondria, nucleolin and DNA (both at 100% for human and 99% for computer accuracy), and actin (100% for both).

The discrimination between the Golgi proteins is even better for 3D images. Giantin and gpp130 can be distinguished an average of 86% with a neural network and SLF9 [8] and this improved to 97% with a majority-voting ensemble and SLF10 [3].

4. SUBCELLULAR LOCATION TREES

The observation that the SLF can be used to train classifiers that not only can recognize the major subcellular patterns but can distinguish subtle differences in proteins within the same organelle raises the possibility that the features can also be used to measure similarity between protein patterns, much as scoring matrices such as PAM250 are used to measure similarity between protein sequences.

The SLF can be used to calculate a multivariate distance beween the average SLF values of any pair of protein patterns. When this is done for all pairs in the 2D HeLa set using a Mahalanobis distance function that adjusts the distance for the presence of correlated features, the distances can be used to construct a hierarchical tree, or dendrogram, in which the vertical distance reflects the distance or dissimilarity between connected clones (Figure 3). The distances agree with expectations in that protein pairs with visually similar proteins have small distances while grossly different proteins have large distances [4]. For example, the two most similar pairs are the two Golgi proteins and the endosomal and lysosomal proteins.

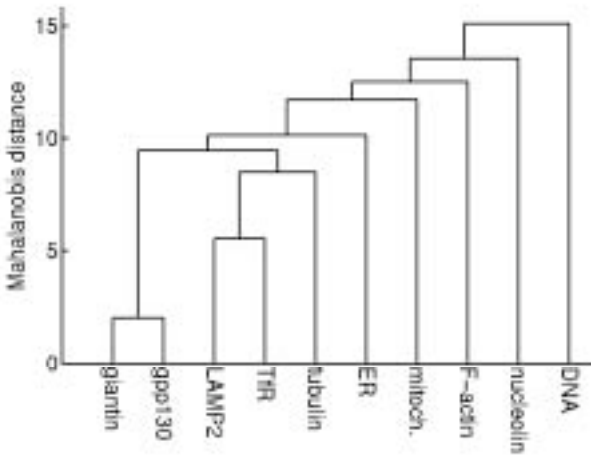

Figure 3. Example Subcellular Location Tree in which the ten patterns in the 2D HeLa dataset are grouped by their similarity as measured by the SLF8 feature set. (From reference [1].)

Therefore, just as scoring matrices can be used to construct trees that group proteins by sequence similarity, so we can expect that the SLFs can be used to build trees that represent the similarity between protein patterns in a systematic and objective manner. As a further illustration of this idea, we have used a collection of 3D images for 46 different clones of 3T3 cells, where each clone expresses a different protein internally fused with GFP [9]. The ten most informative features for these images were selected from the 3D feature set SLF11 and used to calculate the z-scored Euclidean distance between the average feature values of each clone and build a Subcellular Location Tree [10]. We observed that proteins known from the literature to have similar location patterns are grouped together, and proteins whose locations were unknown could have locations assigned by virtue of their similarity to known proteins.

5. CONCLUSIONS

The results reviewed here address the need for automated approaches to the determination and comparison of protein subcellular patterns. In combination with high-throughput microscope systems, the approaches described here can enable a new subfield of proteomics, location proteomics, with the goal of identifying the high-resolution subcellular location patterns of all proteins expressed in a given cell type or organism and systematically organizing them into clusters that share the same pattern. The results of this approach can be merged with other protein databases, which can be expected to greatly aid the discovery of new sequence motifs responsible for protein location.

6. REFERENCES

[1] R.F. Murphy, M. Velliste, and G. Porreca, "Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images," *J VLSI Sig Proc 35:* 311-321, 2003.

[2] M.K. Markey, M.V. Boland, and R.F. Murphy, "Towards Objective Selection of Representative Microscope Images," *Biophys. J. 76:* 2230-2237, 1999.

[3] K. Huang and R.F. Murphy, "Boosting Accuracy of Automated Classification of Fluorescence Microscope Images for Location Proteomics," *submitted*, 2004.

[4] E.J.S. Roques and R.F. Murphy, "Objective Evaluation of Differences in Protein Subcellular Distribution," *Traffic 3:* 61-65, 2002.

[5] 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.

[6] 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.

[7] K. Huang, M. Velliste, and R.F. Murphy, "Feature Reduction for Improved Recognition of Subcellular Location Patterns in Fluorescence Microscope Images," *Proc SPIE 4962:* 307-318, 2003.

[8] M. Velliste and R.F. Murphy, "Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images," in *2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002)*. pp. 867-870, 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] X. Chen, M. Velliste, S. Weinstein, J.W. Jarvik, and R.F. Murphy, "Location Proteomics - Building Subcellular Location Trees from High Resolution 3D Fluorescence Microscope Images of Randomly-Tagged Proteins," *Proc. SPIE 4962:* 298-306, 2003.