

AUTOMATED INTERPRETATION OF MULTIDIMENSIONAL FLUORESCENCE MICROSCOPE IMAGES FOR LOCATION PROTEOMICS

Robert F. Murphy

**Departments of Biological Sciences and Biomedical Engineering and
Center for Automated Learning and Discovery, Carnegie Mellon University
4400 Fifth Avenue, Pittsburgh PA 15213 U.S.A.
E-mail: Murphy@cmu.edu**

KEY WORDS: subcellular location, pattern recognition, image similarity, cluster analysis

Efforts in the growing field of proteomics seek to characterize all expressed proteins in many cell types. Methods for describing proteins in terms of their sequence, structure, and enzymatic activities are well advanced, but there is no current systematic means of adequately describing subcellular location. Fluorescence microscopy is a powerful method for determining subcellular location, but analysis of the resulting images is currently done by visual interpretation. Therefore, current database entries on protein location consist of unstructured text, a set of generic locations from a restricted vocabulary list, or (rarely) an example microscope image. Such entries do not permit the crucial operations that characterize all other biological databases: searching by quantitative similarity and grouping into sets or families that share common attributes. Towards this end, we have developed sets of Subcellular Location Features that capture the essential characteristics of protein patterns in 2D fluorescence microscope images without being overly sensitive to the position, rotation, size or shape of a cell in such images. These features can be used to create automated image classifiers that can recognize the major subcellular structures in a fully automated manner. Importantly, we have shown that the features enable rigorous statistical discrimination between patterns that cannot be distinguished by eye. We have extended this approach to 3D confocal microscope images and demonstrated that such images permit more accurate classification of patterns. This is due in equal parts to the fact that 2D images may not have been collected at the most informative position within the cell and that 3D images inherently yield more informative features. The accuracy of our automated classifiers is over 95% for the major organelle patterns, over 90% for distinguishing two Golgi proteins that cannot be visually distinguished beyond random guessing, and over 99.7% for sets of 9 images drawn from the same slide. Since these results demonstrate that the features capture the essential characteristics of subcellular patterns, they can also be used to measure *similarity* between the protein patterns and to create Subcellular Location Trees that cluster proteins into hierarchies of groups with increasingly similar patterns. We have applied this method to a collection of randomly-tagged proteins to show a systematic representation of the complexity of protein location in a single cell type. The location proteomics methods we describe can be used to capture the manner in which the distribution of every protein (or other macromolecule) changes during the cell cycle, under different environmental conditions, between cell types, and between organisms so that a unified systematics for subcellular location can be created.

This work was supported in part by research grant 99-295 from the Rockefeller Brothers Fund Charles E. Culpeper Biomedical Pilot Initiative, by NIH grants R33 CA83219 and R01 GM068845, and by a research grant from the Commonwealth of Pennsylvania Tobacco Settlement Fund.