 Cytometry Part A 67A:1–3 (2005)

Cytomics: An Entry to Biomedical Cell Systems Biology

Cytomics and Location Proteomics: Automated Interpretation of Subcellular Patterns in Fluorescence Microscope Images

Robert F. Murphy*

Departments of Biological Sciences and Biomedical Engineering, Center for Automated Learning and Discovery, and Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, Pennsylvania

Accepted 8 July 2005

In this column, I briefly reflect on the manner in which automated analysis of the subcellular distribution of proteins (location proteomics) is relevant to the field of cytomics. There are many definitions of cytomics that vary slightly in emphasis. Fundamentally, however, it is the systematic, comprehensive study of at least one cytome, where a cytome is the collection of cell states exhibited by a tissue or organism. We define a cell state as a unique combination of all observable cell behaviors or phenotypes. Different cell types represent different cell states, of course, but the same cell type can exist in more than one state (e.g., activated and quiescent). Clearly, a cell's state is influenced by and reflected in the set of proteins that it expresses.

However, simply knowing how much of a protein is expressed is not sufficient to understanding its contribution to the cell state. It is particularly important to also know its subcellular location because changes in protein subcellular location can cause dramatic effects on cell behavior. Perhaps the most thoroughly studied example of this phenomenon is the changes in protein location associated with apoptosis (1). Changes in location within a cell type may also cause or result from disease, as illustrated by the suspected involvement of the Wnt pathway and β-catenin in a number of cancers (2).

Based on the success of the various genome projects, the feasibility and desirability of undertaking projects to study a single aspect of gene or protein structure or function has become accepted. Many such projects have been initiated, including projects to determine or predict all protein structures and to measure gene and protein expression levels in many cell types and under many conditions. However, subcellular location has received less attention than many other aspects of gene and protein behavior. The major exception is in yeast, in which almost all proteins have been assigned to a set of major subcellular structures (3,4) using fusion of cDNAs with the coding sequence of fluorescent proteins such as the green fluorescent protein. For example, Huh et al (4) used green fluorescent protein tagging of cDNAs and visual examination to assign proteins to 12 categories: cell periphery, bud, bud neck, cytoskeleton, microtubule, cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, vacuole, vacuolar membrane, and punctate. They then used colocalization with red fluorescent protein markers to divide the cytoskeleton class into two classes, actin cytoskeleton and spindle pole, and to add nine new categories: nucleolus, nuclear periphery, golgi apparatus, three types of transport vesicles, endosome, peroxisome, and lipid particle. In all, 4,156 proteins were assigned to these 22 categories in their study.

Pilot projects in mammalian cells have also been described. For example, Simpson et al. (5) used cDNA tagging to localize approximately 100 proteins in a human cell line, and Jarvik et al. (6) used a clever genomic-tagging approach (termed CD-tagging) to localize a similar number of proteins in mouse 3T3 cells. As with the yeast studies, analysis was restricted to assignment of proteins to one of a limited number of major locations.

These results, although useful and illustrative, do not provide location information with sufficient resolution to be useful for understanding and modeling cell behavior. The limited resolution also applies to systems that have been designed for predicting subcellular location from protein sequence. Further, there is an implicit assumption in many prediction schemes or curated protein databases that proteins have a single location regardless of cell type or condition. In contrast, location is not necessarily the same between different cell types, as illustrated by the differ-

*Correspondence to: Robert F. Murphy, Departments of Biological Sciences and Biomedical Engineering, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213.

E-mail: murphy@cmu.edu

Published online 4 August 2005 in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/cyto.a.20179

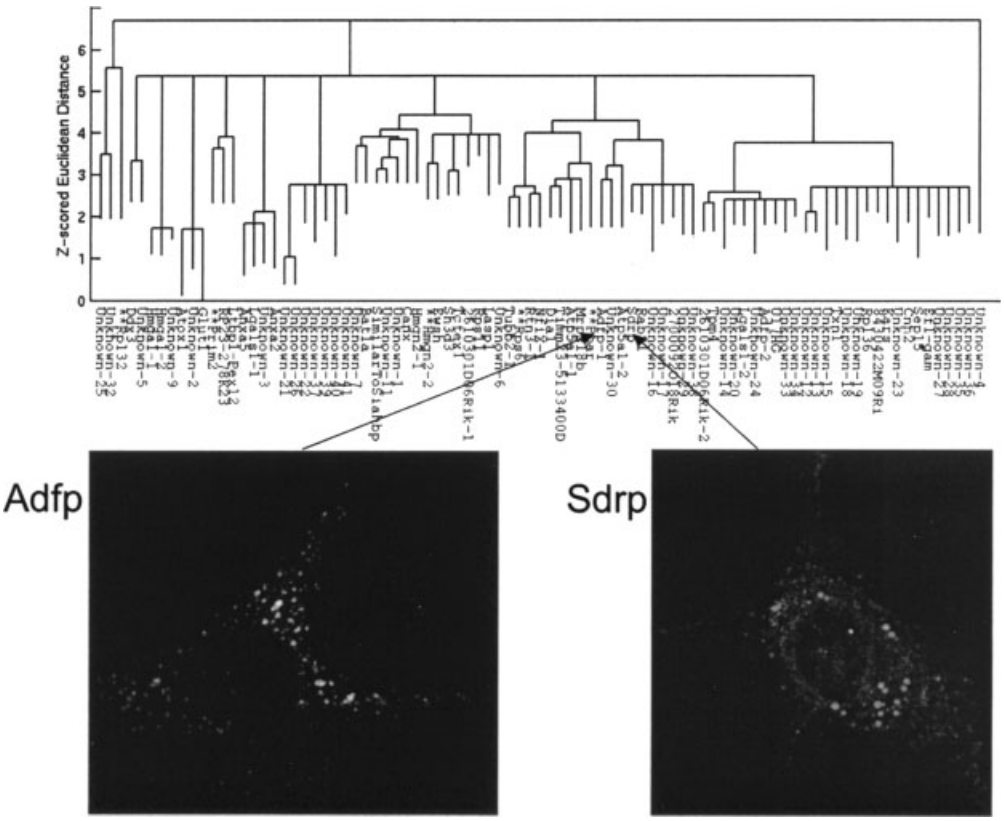

FIG. 1. Consensus subcellular location tree for 87 3T3 cell clones expressed by different CD-tagged proteins. Numerical features were calculated to describe each image and then proteins were grouped into statistically distinguishable groups. An interactive browser (available at http://murphylab.web.cmu.edu/PSLID/tree.html) permits viewing of images for particular proteins in order of the degree to which they are representative of the overall pattern. Examples for two proteins within neighboring but distinct clusters are shown. See Chen and Murphy (13) for more information on the clustering methods used.

ences in subcellular location of viral glycoproteins between cell types that correlate with viral susceptibility (7).

The analysis above demonstrates the need for high-resolution, comprehensive analysis of the subcellular location of proteins in many or all cell types. This demands high-throughput methods for imaging tagged proteins and automated methods for analyzing the resulting images. To meet the latter need, my colleagues and I began applying machine learning methods to subcellular pattern analysis a number of years ago. We initially demonstrated the feasibility of automated classification of subcellular patterns (8) and have extended and refined these results to the point that all major subcellular patterns can be recognized in two- and three-dimensional images of single cultured cells with very high accuracy (9). An important conclusion from this work is that automated classifiers can not only be trained for this task but also can perform better than visual examination (10). More recently, the combination of classification methods with an automated imaging system has been described (11).

Although automated, these classification approaches still have the same limitation as the visual and prediction approaches: they can recognize only the major patterns that they have been trained with. An important alternative therefore is to use unsupervised machine learning (cluster analysis) to group proteins by their high-resolution patterns. We have coined the term "location proteomics" (12) to describe the combination of large-scale protein tagging, high-resolution imaging and clustering by subcellular pattern. The most extensive results of this type described to date are for 90 tagged 3T3 clones that were demonstrated to contain 17 distinct location patterns (Fig. 1) (13). A similar clustering approach has been taken to group drugs by their effects on subcellular patterns (14).

In addition to being critical for bottom-up systems biology efforts to model cell behavior, information that will become available from location proteomics over the next decade can provide important clues to proteins that reflect abnormal cell states. These can then be used with the same automated pattern analysis methods to detect disease or monitor therapy.

LITERATURE CITED

1. Jakobi R. Subcellular targeting regulates the function of caspase-activated protein kinases in apoptosis. Drug Resist Updat 2004;7:11–17.
2. Karim R, Tse G, Putti T, Scolyer R, Lee S. The significance of the Wnt pathway in the pathology of human cancers. Pathology 2004;36:120–128.
3. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, et al. Subcellular localization of the yeast proteome. Genes Dev 2002;16:707–719.
4. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. Global analysis of protein localization in budding yeast. Nature 2003;425:686–691.
5. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep 2000;1:287–292.
6. Jarvik JW, Fisher GW, Shi C, Hennen L, Hauser C, Adler S, Berget PB. In vivo functional proteomics: mammalian genome annotation using CD-tagging. Biotechniques 2002;33:852–867.

7. Fish KN, Soderberg-Naucler C, Nelson JA. Steady-state plasma membrane expression of human cytomegalovirus gB is determined by the phosphorylation state of Ser900. J Virol 1998;72:6657–6664.
8. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 1998;33:366–375.
9. Huang K, Murphy RF. From quantitative microscopy to automated image understanding. J Biomed Optics 2004;9:893–912.
10. Murphy RF, Velliste M, Porreca G. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J VLSI Sig Proc 2003;35:311–321.
11. 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 Res 2004;14:1130–1136.
12. 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.
13. Chen X, Murphy RF. Objective clustering of proteins based on subcellular location patterns. J Biomed Biotechnol 2005;2005:87–95.
14. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science 2004;306:1194–1198.