SYSTEMATIC DESCRIPTION OF SUBCELLULAR LOCATION FOR INTEGRATION WITH PROTEOMICS DATABASES AND SYSTEMS BIOLOGY MODELING

Robert F. Murphy

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

ABSTRACT

Approaches for automatically analyzing subcellular location on a proteome-wide basis have been developed. This permits proteins to be grouped into Subcellular Location Families that share a statistically-indistinguishable pattern. This in turn creates a need for methods to connect these automatically determined locations to existing information in literature and databases, and to communicate the nature of the pattern for each family. The building of generative models is proposed to meet this need and their utility for simulations of cell behavior is discussed.

Index terms -- Location Proteomics, Fluorescence Microscopy, Cluster Analysis, Generative Models

1. INTRODUCTION

A major paradigm in current biological research is the systematic, comprehensive collection of information about a particular biological process. As a result, a number of new fields have been born, each of which focuses on a particular process for one or more types of biological macromolecules. The large field of proteomics is concerned with study of the characteristics of all proteins, and location proteomics [1] is that subfield that focuses on the location of proteins within cells. A critical step towards the development of this field was the demonstration that machine learning methods could be applied to recognize subcellular patterns in fluorescence microscope images [2-5]. In fact, automated classifiers could perform better at distinguishing subtly different subcellular patterns than visual examination [6]. Perhaps most importantly, this work established that features drawn from a range of image analysis approaches could adequately capture the complex, highly variable location patterns displayed by proteins.

Given these results, one possible approach to studying subcellular location on a proteome-wide basis would be to somehow collect images of the distributions of all proteins and then use automated classifiers to assign each each protein to one of the major subcellular patterns. This might be useful for gaining some initial insight into the roles of unknown proteins. However, most proteins are not homogenously distributed throughout a single organelle but rather are often found only in specific regions of an organelle (or in more than one organelle). Thus, the task of fully understanding subcellular location at the proteome level must include the task of identifying the set of possible patterns that proteins may display, with the expectation that some proteins will show novel patterns.

2. COMPREHENSIVE DATA COLLECTION

Dramatic advances in protein-tagging and microscopy technologies have made feasible the determination of the subcellular locations of all proteins expressed in a given cell type. Nonetheless, the rate-limiting step remains acquisition of the necessary images. A number of approaches that could be scaled to the whole proteome level have been described, including both transfection of tagged cDNAs [7] and random tagging of genomic DNA [8]. Large scale projects using different approaches for different organisms and cell types are currently being pursued.

Once images have been collected, they must segmented into single cell regions so that numerical features describing each cell can be calculated. The general task of cell segmentation is a difficult one that is beyond the scope of this discussion, but a number of systems with reasonable performance have been described.

3. IDENTIFICATION OF PROTEIN LOCATION FAMILIES

Given a set of images of the subcellular patterns of a large number of proteins, identifying the set of statistically distinct patterns displayed is an *unsupervised learning* task. Cluster analysis methods can be used for this type of task, and the feasibility of applying these methods to subcellular patterns has been demonstrated [1, 9]. A critical issue in using these methods is establishing the criteria by which clustering results are to be evaluated. One criterion that is widely used is that results agree with all well-accepted prior knowledge. An alternative that does not require prior knowledge is that clusters formed from different samplings of available data agree with each other. We have obtained good results using this *consensus clustering* approach [10].

We are currently using automated microscopy to collect images for thousands of tagged proteins in NIH 3T3 cells and plan to apply consensus clustering methods to identify *Subcellular Location Families* [10]. Of course, if this process is repeated for a different cell type (or the same cell type under different conditions, such as in the presence of a drug), we may expect that the some proteins may be found in different clusters or that new clusters may appear. Thus the definition of Subcellular Location Families can be either *local* to a specific cell type and condition or *global*. Determining the global families will require data collection on an unprecedented scale. However, we can expect that results from a well-chosen set of cell types/tissues and conditions can provide a reasonable approximation to the global families.

4. REPRESENTATION OF PROTEIN LOCATION FAMILIES AT MULTIPLE RESOLUTIONS

Regardless of whether a grouping of proteins into location families is local or global, an important issue is how to relate them to results from other sources. Comprehensive experiments that could provide data for clustering are typically taken under well-defined conditions so that images of different proteins can be readily compared. These may be different between large scale experiments for different cell types. To address this problem, we have obtained preliminary encouraging results on using log-linear transforms to map numerical features between different cell types and conditions, assuming that a group of calibration proteins is present in both sets (as would normally be the case) [11]. However, data from smaller scale projects or individual papers in the literature are likely to have been acquired under widely varying conditions and calibration data would not typically be available for each source. In particular, such data will typically vary in spatial resolution (i.e., magnification). Reconciling image-derived locations with annotations in databases is even more difficult, since they will typically use vocabularies that do not capture the full range of possible location patterns.

One approach to this problem is to create a hierarchical tree that shows location clusters for images of increasing resolution (magnification). Starting from high-resolution images, such a tree can be generated by cycles of two-fold downsampling and clustering followed by linking the clusters containing the same proteins [11]. An example is shown in Figure 1. Note that there is an implicit root node at the top of the tree since at some point downsampling would make all patterns appear the same. Note also that the topology of this tree does not have to match that of hierarchies such as the Genome Ontology Consortium's Cellular Component Ontology. This is because the former links organelle patterns that appear to be the same at a particular magnification (e.g., lysosomes and mitochondria might appear the same at low magnification), while the latter links organelles that are specializations of the same root organelle (which lysosomes and mitochondria are not).

5. GENERATIVE MODELS FOR REPRESENTING PROTEIN LOCATION FAMILIES

We next consider how we can communicate results on protein patterns for automatically determined subcellular location families. One approach, of course, is to simply report the number of the node in a hierarchical tree to which that protein is assigned. If queried about the nature of the pattern at that node, one possible response is the names of all other proteins at that node - hoping that one of them will be known to the questioner! Another is to return one or more images (such as the most typical image from that node). It is difficult by this approach, however, to communicate the often extreme variability within one pattern, or to emphasize what differentiates a particular pattern from the patterns of proteins at other nodes. This requires some *model* of the pattern, created either automatically or by hand. Such a model could be purely *descriptive*, which would enable recognition of future examples of the pattern, or *generative*, which would enable synthesis of new images of the pattern. Statements from descriptive models might include statements about the amount of edge content in a pattern or that it has objects of approximately 1 micron in diameter, while generative models would contain parameters describing shape distributions learned from training images.

We have obtained encouraging results on learning generative models of subcellular distribution from images of

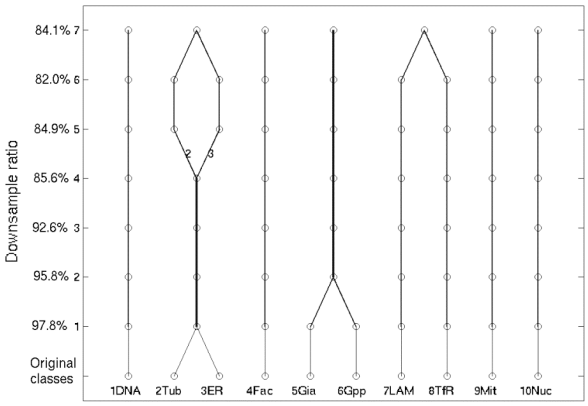

Figure 1. Coalescence of subcellular patterns upon downsampling. The labels across the bottom show the classes of tagged proteins in the 3D HeLa dataset. Note that seven patterns can still be distinguished even after seven-fold downsampling of the original images. From [11].

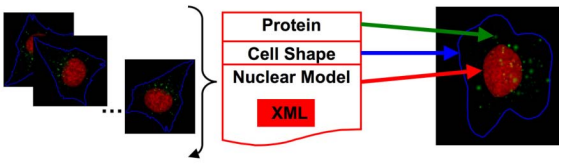

Figure 2. Creation of generative model. The parameters of a three-component model can be learned from a set of images and captured in an XML file. This can be used to synthesize new images in three steps.

HeLa cells (Zhao, T., and Murphy, R.F., in preparation). The starting point is a three-color image in which one channel reflects DNA content, one channel reflects total protein content, and one channel reflects the amount of a specific protein of interest. We use the DNA channel to build a model that captures the shape and chromatin texture of the nucleus, and then build a cell shape model that is conditional on the nuclear model. (This ensures that the nucleus is always contained with the cell.) Lastly, we learn models of the shape of each protein-containing organelle and their distribution in space relative to the nucleus and cell membrane. The parameters of the generative model can be stored in a compact XML file. The principle is illustrated in Figure 2. Such XML files are built in a fully automated manner from each cluster and can be readily downloaded to permit local viewing of generated images and other uses.

Generative models are useful not only for capturing and communicating the essence of a location family, but also for use in simulations of cell behavior. A number of sophisticated systems have been described for such simulations, such as Virtual Cell [12]. Such systems can utilize an empirical compartment geometry (typically in the form of a binary image for each compartment in the model) to provide simulation results that take subcellular location into account. The generative models of subcellular location we have developed can be used to create many such compartment geometries in a manner that reflects the statistical variation observed in the original image data. This allows simulation conclusions to be based on the full range of observed distributions (Figure 3).

Generative models can also potentially be combined so that simulations can include accurate distributions for many proteins, even more than could simultaneously be imaged. This requires some mechanism for specifying the correlation between the synthesized distributions for different proteins. One approach would be to obtain experimental data on the colocalization of all pairs of proteins (or enough pairs that the remaining correlations could be inferred). This would appear impractical for tens of thousands of proteins using

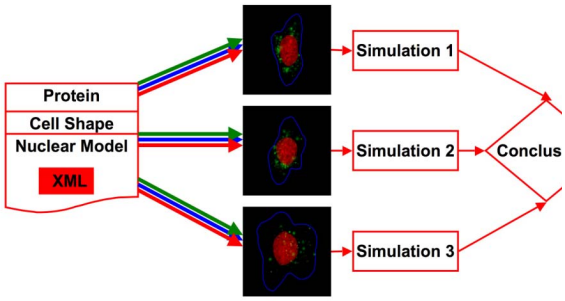

Figure 3. A generative model can be used to synthesize a collection of images reflecting variation in the underlying pattern for subsequent use in simulations.

conventional microscopy. While technology for cyclical imaging of as many as a hundred proteins in the same sample has been described [13], it cannot be applied to living cells. We therefore propose the operational definition that all proteins in an automatically determined subcellular location family have effectively perfect correlation between their distributions. This permits the models for different proteins to be combined into a single simulation (Figure 4). The assumption of perfect correlation can be relaxed by including noise in the generation process.

6. CONCLUSION

The combination of approaches for proteome-wide tagging and high throughput microscopy is expected to provide detailed information on the subcellular distribution of proteins. Automated methods can be used to group proteins by their subcellular location pattern. Generative models can then be used as an effective means of communicating the patterns displayed by these groups and using them in systems biology simulations. This promises to permit for the

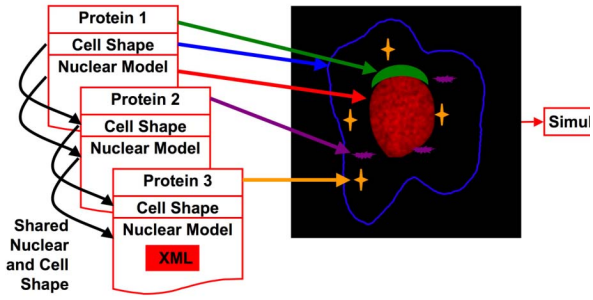

Figure 4. A set of generative models learned from separate images of multiple proteins can be used to synthesize an image containing all of them. This assumes that the correlations can be inferred.

first time the creation of simulations in which thousands of proteins are properly localized inside a virtual cell (or cells) and observed variation in their locations can be accounted for in the simulations.

7. ACKNOWLEDGEMENT

This original research reviewed here was supported in part by NIH grants R01 GM078622-01 and. NSF grant EF-0331657.

8. REFERENCES

[1] 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*, vol. 4962, pp. 298-306, 2003.

[2] M. V. Boland, M. K. Markey, and R. F. Murphy, "Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy," *19th Annu. Intl. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 594-597, 1997.

[3] M. V. Boland, M. K. Markey, and R. F. Murphy, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry*, vol. 33, pp. 366-375, 1998.

[4] 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*, vol. 17, pp. 1213-1223, 2001.

[5] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes, "Automated recognition of intracellular organelles in confocal microscope images," *Traffic*, vol. 3, pp. 66-73, 2002.

[6] 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*, vol. 35, pp. 311-321, 2003.

[7] 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 Rep*, vol. 1, pp. 287-92, 2000.

[8] 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*, vol. 33, pp. 852-867, 2002.

[9] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, "Multidimensional Drug Profiling by Automated Microscopy," *Science*, vol. 306, pp. 1194-1198, 2004.

[10] X. Chen and R. F. Murphy, "Objective clustering of proteins based on subcellular location patterns," *J Biomed Biotechnol*, vol. 2005, pp. 87-95, 2005.

[11] X. Chen and R. Murphy, "Interpretation of Protein Subcellular Location Patterns in 3D Images Across Cell Types and Resolutions," *Lecture Notes in Computer Science*, vol. 4414, pp. 328-342, 2007.

[12] Moraru, II, J. C. Schaff, B. M. Slepchenko, and L. M. Loew, "The virtual cell: an integrated modeling environment for experimental and computational cell biology," *Ann N Y Acad Sci*, vol. 971, pp. 595-6, 2002.

[13] W. Schubert, "Exploring molecular networks directly in the cell," *Cytometry A*, vol. 69, pp. 109-12, 2006.