Location proteomics: a systems approach to subcellular location

R.F. Murphy[1]
Departments of Biological Sciences and Biomedical Engineering, Center for Automated Learning and Discovery and Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Abstract

Systems Biology requires comprehensive systematic data on all aspects and levels of biological organization and function. In addition to information on the sequence, structure, activities and binding interactions of all biological macromolecules, the creation of accurate predictive models of cell behaviour will require detailed information on the distribution of those molecules within cells and the ways in which those distributions change over the cell cycle and in response to mutations or external stimuli. Current information on subcellular location in protein databases is limited to unstructured text descriptions or sets of terms assigned by human curators. These entries do not permit basic operations that are common to other biological databases, such as measurement of the degree of similarity between the distributions of two proteins, and they are not able to fully capture the complexity of protein patterns that can be observed. The field of location proteomics seeks to provide automated, objective high-resolution descriptions of protein location patterns within cells. Methods have been developed to group proteins into statistically indistinguishable location patterns using automated analysis of fluorescence microscope images. The resulting clusters, or location families, are analogous to clusters found for other domains, such as protein sequence families. Preliminary work suggests the feasibility of expressing each unique pattern as a generative model that can be incorporated into comprehensive models of cell behaviour.

Introduction

Systems Biology requires comprehensive, systematic data on all aspects and levels of biological organization and function. The genomics revolution led to a new paradigm for obtaining such data: acquiring data on a specific aspect of structure or function across a complete (or nearly complete) set of biological entities. The complete set of these entities can be referred to using the suffix 'ome' appended to the name of the entity to be studied (e.g. genome, transcriptome and proteome) and its study referred to using the suffix 'omics' (e.g. proteomics). The aspect to be studied can then be used as a modifier. Thus proteomics is the study of all proteins (in a given cell or tissue, each of which may have a distinct proteome), and interaction proteomics is the study of the interactions between the proteins in a proteome. This approach to terminology has advantages over alternatives that define an 'ome' [1] for each possible aspect or collection of things that can be studied, even if those things are not tangible. For example, the disadvantage of using the alternative 'interactome' is that, although it is clear that the aspect being studied in a systematic way is interactions, the nature of the things that are interacting is unspecified.

Following the logic laid out above, we have coined the term location proteomics [2] to describe the systematic study of the location of proteins within cells. Location proteomics requires one or more methods for determining or predicting the location, combined with a systematic means of organizing and referring to the set of possible locations in which a protein may be found.

Word = 0.001 picture

To date, the latter has been best provided using the terms of the cellular component ontology created by the Gene Ontology Consortium [3]. However, there are at least two significant problems with using this vocabulary-based approach. The first is the often arbitrary manner in which terms are assigned to a protein (and the resulting difficulty of reasoning from them). For example, we may ask why one protein (human gpp130/golph4, SwissProt accession number O00461) is assigned the term 'Golgi lumen' and another (human giantin/golgb1, SwissProt accession number Q14789) is assigned 'Golgi stack.' The second problem is the inability of the vocabulary to capture complex locations, such as 'just the rims of the *cis*- and *medial*-Golgi cisternae'.

Thus ways in which the Gene Ontology terms can be combined are both too ambiguous and too numerous to ensure that the same unique combination is always assigned to a given protein pattern, and are insufficient to capture the significant changes in the patterns observed in cells.

Since experimental determination of the location has been performed only for a small fraction of all the proteins, location prediction (reviewed in [4]) would appear to be a

Key words: fluorescence microscopy, location proteomics, machine learning, protein tagging, subcellular location, 3T3 cell.
Abbreviations used: SLF, subcellular location feature.
[1]To whom correspondence should be addressed (email murphy@cmu.edu).

valuable approach (as it has been, to some extent, in structural proteomics). Unfortunately, the absence of a well-defined framework for assigning known proteins to a high-resolution location limits the ability of researchers to train location predictors with sufficient complexity to be ultimately useful for predicting locations for unknown proteins. Knowledge of the high-resolution location patterns of proteins is required for meaningful systems modelling. A particular area of uncertainty is whether location predictors, especially those based on general properties such as amino acid composition, are capable of predicting the ways in which location will change for specific amino acid changes (e.g. for proteins encoded by mutant alleles).

Automated determination of subcellular location using fluorescence microscopy

There is a strong need for the collection of new data providing information on the high-resolution patterns of proteins, but there is also a need for tools that can interpret that data automatically and assign proteins to systematically organized locations. The most common and practical means of obtaining data on subcellular location is to collect images of tagged proteins by fluorescence microscopy. The development of tools to interpret such images has been the focus of research by my group for a number of years.

Before discussing these tools, it is worth discussing various approaches to tagging of proteins for determining their location. These can be divided into those that tag native, unmodified proteins and those that modify the protein (usually by manipulating the encoding DNA) to introduce a means of visualizing it. The primary approach to tagging native proteins is immunofluorescence, which uses antibodies, but fluorescent probes that bind to specific proteins (e.g. phalloidin binding to F-actin) are also used. This approach has two major disadvantages: it cannot be used on live cells and it depends on the availability of antibodies or probes of sufficient specificity. The latter is especially a problem when considering the determination of location on a proteomewide basis.

Tagging of proteins by manipulating their encoding DNA sequence does not suffer from these limitations, but introduces the concern that the location of the tagged protein may be altered by the presence of the tag. This concern cannot be eliminated, but it can be lowered in some circumstances (as discussed below). Protein tagging can be accomplished either by manipulating a specific isolated coding sequence (either genomic or cDNA) and then introducing the manipulated gene into cells or by tagging a coding sequence (either randomly or specifically) within the genome. A particularly powerful version of the latter approach is CD tagging, which utilizes a retroviral vector to insert a green fluorescent protein-encoding exon into genomic DNA [5]. Tagging of genomic DNA has the advantage that normal regulatory sequences are preserved so that the levels of the tagged protein can be expected to be similar to those of the unmodified protein. In contrast, expression of tagged cDNAs is usually accomplished using a strong promoter so that levels are typically much higher than those for the endogenous proteins. This can lead to a concern beyond that resulting from the tagging itself, since overexpression may saturate targeting mechanisms and lead to mislocalization of the tagged protein.

Many projects describing the analysis of significant numbers of tagged proteins have been reported (reviewed in [5]). Most of these projects have involved cDNA fusions.

Given a means of generating images of tagged proteins, conclusions about the subcellular locations of those proteins can be drawn by visual inspection or by automated interpretation. To demonstrate the feasibility of the latter approach, my group carried out initial work on automated classification of protein patterns for proteins whose subcellular location is known [6,7]. Using two-dimensional images of HeLa cells showing the distributions of nine proteins (with a parallel image of DNA), we showed that the patterns of all major organelles could be recognized. This recognition was accomplished using sets of numerical features to describe the pattern in each image combined with an automated classifier, such as a neural network. As we have implemented additional features and more robust classifiers over the past few years, the overall accuracy of recognition has improved from an initial value of 84% for a neural network classifier using a set of 37 features [7] to the best current value of 92% for a majority-voting ensemble classifier using a set of 47 features [8]. Perhaps the most important conclusion from the present study was not only that automated recognition of subcellular patterns was feasible, but also that it could be more accurate than visual inspection! This was demonstrated by measuring human accuracy on the same images. The results showed an overall accuracy of 83%, but, most importantly, showed that the patterns of two Golgi proteins that can be distinguished with an average accuracy of >86% by the most recent automated system could not be distinguished beyond random guessing by visual classification [9].

Since cells are three-dimensional (and usually have significant thickness compared with the axial resolution of a microscope, at least at high magnification), we next examined whether the use of three-dimensional images would improve the classification accuracy of the HeLa patterns [10]. This was indeed the case, with the most accurate current classifier being able to distinguish the same patterns with an average accuracy of 98% [11]. Table 1 shows a confusion matrix for these patterns, where rows of the matrix depict the protein which was actually labelled in an image, and the columns depict the pattern assigned to that image by the automated classifier, and the values in the matrix represent the fraction of the images in a given row that were classified in a given column. Perfect performance would result in 100% along the diagonal and 0% elsewhere, while random performance would yield approx. 10% in each position. The results are nearly perfect and the ability to distinguish the two Golgi proteins, as well as the very similar lysosomal and endosomal patterns, can be clearly seen.

The sets of numerical features used to obtain these results were drawn from a number of categories that capture distinct

Table 1 | Results of an automated classification of three-dimensional images of HeLa cells

Values shown are the percentage of images from the class shown in the row that were classified as belonging to the class shown in the column. The average rate of correct classification was 98%. Taken from [11]. ER, endoplasmic reticulum; LAMP2, lysosome-associated membrane protein 2; Mit., mitochondria.

	Output of the classifier									
True class	**DNA**	**ER**	**Giantin**	**Gpp130**	**LAMP2**	**Mit.**	**Nucleolin**	**Actin**	**TfR**	**Tubulin**
DNA	98	2	0	0	0	0	0	0	0	0
ER	0	100	0	0	0	0	0	0	0	0
Giantin	0	0	100	0	0	0	0	0	0	0
Gpp130	0	0	0	96	4	0	0	0	0	0
LAMP2	0	0	0	4	95	0	0	0	0	2
Mitochondria	0	0	2	0	0	96	0	2	0	0
Nucleolin	0	0	0	0	0	0	100	0	0	0
Actin	0	0	0	0	0	0	0	100	0	0
TfR	0	0	0	0	2	0	0	0	96	2
Tubulin	0	2	0	0	0	0	0	0	0	98

information about an image. For example, morphological features capture information about the properties of fluorescent objects in an image (such as individual vesicles), while texture features reflect the probability that a pixel of one intensity is found adjacent to a pixel of another intensity (e.g. reflecting whether the pattern is random or organized into elements like stripes or circles). All features were implemented in such a manner that they are at least largely unaffected by the position, rotation, size or shape of a given cell in a field. A standard nomenclature for the features and feature sets has been defined. Each set of features is referred to by a number and the prefix SLF (subcellular location feature), and each individual feature is referred to by the set name followed by a period and the number of that feature within the set (e.g. SLF3.2 is the second feature within set SLF3). Once a set of features has been constructed, the critical next step is deciding which features are useful for the purposes of classification. We have compared a number of methods that can be used for this feature reduction step [12]. The best results were obtained with one of the classical methods, stepwise discriminant analysis.

The demonstration that the SLF can be used to represent protein location patterns sufficiently so that they can be recognized with very high accuracy leads naturally to a new use for the features: calculating the similarity between two location patterns. This can be done, for example, to compare two cell images of the same protein or the average feature values from many images of two different proteins. The use of the SLF in this manner creates, for the first time, an objective way of measuring the degree of similarity between the location patterns of different proteins, a capability that has been critical in other domains (such as sequence comparison using programs such as BLAST). Similarity of location for two proteins could previously be measured by comparing the Gene Ontology terms assigned to them; however, as discussed above, this similarity measure relies on subjective assignment of Gene Ontology terms to the patterns.

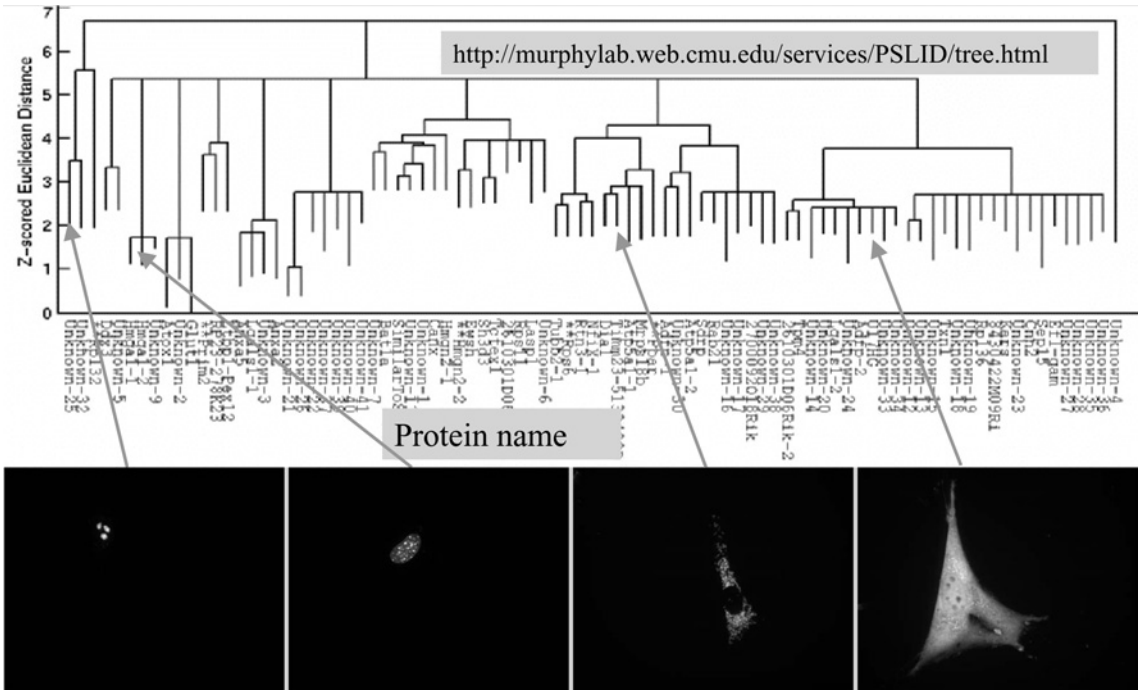
The next step towards a systematic framework for representing subcellular location is the use of similarity measured using SLFs to construct an objective grouping (or clustering) of proteins based on their location. This is an unsupervised learning task (as opposed to the supervised learning task of classification), and it is therefore critical to have one or more explicit criteria for evaluating the final groupings of proteins. There are many methods that can be used to perform this clustering, and many distance functions that can be used. A very analogous situation arises in the analysis of transcript levels using microarrays, but the case of subcellular patterns may be considered more challenging because the variation in pattern from cell to cell within the same population is usually greater than the variation in relative RNA level from duplicate samples.

We have explored using different clustering methods and different distance functions to cluster three-dimensional images of different clones of 3T3 cells obtained by Dr J. Jarvik and Dr P. Berget using CD tagging (as discussed above). We proposed a measure of the degree to which clusterings by different methods agree as a criterion for choosing a distance function, and described a method for rejecting outlier images and choosing a features set. When this approach was applied to images of the first 90 clones of 3T3 cells, we obtained a consensus cluster tree that agrees with assignments by visual inspection (and, where available, information on location from protein databases), but which subdivides patterns further than can be done reliably by eye. Figure 1 shows such a tree with representative images from a few of the clusters. The entire tree (and the supporting images) can be viewed through an interactive browser at http://murphylab.web.cmu.edu/services/PSLID/tree.html.

Of course, to build a comprehensive tree for subcellular location requires images of far more than 90 proteins. A major challenge for the field will be to collect images for tens of thousands of tagged cell lines (just within one cell type!). Tools for incorporating images from more than one

Figure 1 | Subcellular location tree for three-dimensional images of randomly tagged proteins in 3T3 cells

CD-tagged clones were generated and images collected by spinning disc confocal microscopy as described in [2] and clustered as described in the text and in [13].



cell type will also be required, and we have made some preliminary progress on combining images from two cell types into a single tree.

Location proteomics meets systems biology

The final task required to provide systems biology with sufficient information on subcellular location so that accurate cell (and tissue) simulations can be made will be to convert the protein patterns discovered by the methods described above into generative models of protein distribution. Such models will enable the generation of synthetic cell images that are statistically compatible with the underlying images and, hopefully, in a manner that will allow the distribution of more than one protein to be simulated in a single cell. We have obtained very preliminary encouraging results for building such models, but the full realization of this goal will be a major challenge over the next few years. Ultimately, we anticipate the availability of detailed knowledge of the distribution of all proteins within all the major cell types, and of models that predict how these distributions change during development and disease.

References

1 Greenbaum, D., Luscomb, N.M., Jansen, R., Qian, J. and Gerstein, M. (2001) Genome Res. **11**, 1463–1468
2 Chen, X., Velliste, M., Weinstein, S., Jarvik, J.W. and Murphy, R.F. (2003) Proc. SPIE Int. Soc. Opt. Eng. **4962**, 298–306
3 Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B. and Mungall, C. (2004) Nucleic Acids Res. **32**, D258–D261
4 Feng, Z.P. (2002) In Silico Biol. **2**, 291–303
5 Jarvik, J.W., Fisher, G.W., Shi, C., Hennen, L., Hauser, C., Adler, S. and Berget, P.B. (2002) Biotechniques **33**, 852–867
6 Boland, M.V., Markey, M.K. and Murphy, R.F. (1998) Cytometry **33**, 366–375
7 Boland, M.V. and Murphy, R.F. (2001) Bioinformatics **17**, 1213–1223
8 Huang, K. and Murphy, R.F. (2004) BMC Bioinformat. **5**, 78
9 Roques, E.J.S. and Murphy, R.F. (2002) Traffic **3**, 61–65
10 Velliste, M. and Murphy, R.F. (2002) in 2002 IEEE International Symposium on Biomedical Imaging (ISBI-2002), Bethesda, MD, pp. 867–870
11 Chen, X. and Murphy, R.F. (2004) in 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, CA, pp. 1632–1635
12 Huang, K., Velliste, M. and Murphy, R.F. (2003) Proc. SPIE Int. Soc. Opt. Eng. **4962**, 307–318
13 Chen, X. and Murphy, R.F. (2005) J. Biomed. Biotechnol., in the press

Received 15 March 2005