Putting proteins on the map

Robert F Murphy

A combination of microscope technology and statistical analysis enables the identification of proteins that share subcellular location patterns.

The emergence of systems biology has led to a demand for quantitative, comprehensive data on all aspects of biological phenomena. Some aspects, such as RNA expression and protein structure, have received much attention, while others, such as protein location, have lagged behind. In this issue, Schubert *et al.*[1] describe a new approach for mapping the subcellular location of hundreds of proteins and identifying proteins that share subcellular location patterns in single cells or tissue sections.

Detailed knowledge of the spatiotemporal distribution of all proteins expressed in a given cell type is critical to understanding how that cell type behaves and to building computational models that simulate cell behavior. Acquiring this information for even a single cell type is daunting, and collecting it for all cell types in all their developmental, environmental or disease states seems beyond reach. An additional, potentially circular complication is that simply determining how many cell types in an organism must be understood is a challenge that may require knowledge of subtle differences in protein expression and location.

The top-down approach to systems biology posits that it is not necessary to know all the details of the ways in which cells work in order to make useful predictions about how organisms fail and how those failures may be prevented or corrected[2]. Though this vision is attractive, the bottom-up approach, in which one starts from the properties of individual molecules and builds models at increasingly higher scales, may be necessary for tackling some complex diseases and is one of the most important goals of biological research.

Robert F. Murphy is Professor of Biological Sciences, Biomedical Engineering and Machine Learning, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA.
e-mail: murphy@andrew.cmu.edu

Bottom-up cell modeling requires subcellular location data on tens of thousands of proteins for on the order of 100 cell types and 100 conditions. For each combination of protein, cell type and condition, data are also needed for all the temporal scales that can affect protein location, from smaller than seconds up to years. If only one image for each combination of protein, cell type, condition and time were sufficient, a brute-force approach would require on the order of a 100 billion images. Accounting for variation in pattern within a cell type would require at least 10–100 times that number. Furthermore, data on the spatial correlation among proteins are also needed, and the variation in shape and organelle position from cell to cell makes it difficult to determine whether the pattern of one protein is the same as that of another unless they are measured simultaneously.

At the same time, the brute-force approach would probably generate an enormous amount of redundant information: for example, all of the integral ribosomal proteins might show the same pattern under most conditions. We are therefore left with developing a solution to a complex optimization problem.

Such a solution may be found in combining a range of methods providing different types of data along with sophisticated techniques to identify what new information must be acquired. For subcellular location, some systematic methods have previously been worked out. Approaches for large-scale fluorescent tagging of proteins have been described[3,4] and automated microscopes for collecting images have been developed[5,6]. Much of the analysis of the resulting images has been by visual inspection, but automated methods are becoming well established[7]. A combination of random protein tagging and automated analysis was recently used to identify a large number of proteins whose location pattern changes during the cell cycle[8].

An important strategy for reducing the complexity of the location problem is the grouping of proteins or conditions with similar or identical location patterns. One such approach is to collect many separate images of each protein and then use cluster analysis to group proteins into statistically different location patterns (or location families). When this was done for a demonstration set of images of randomly tagged proteins, the groupings agreed with and provided finer distinctions than visual inspection[9]. Cluster analysis has also been applied to group drugs by their effects on a small number of markers[10].

The underlying assumption of these approaches is that proteins that show statistically indistinguishable patterns in separate images are in fact colocalized (or that drugs that show statistically indistinguishable effects separately would show the same effect if added together). Two-color labeling could be used to directly test colocalization, but imaging all pairwise combinations of proteins would square the number of images needed. Using more than two fluorescent dyes can reduce this number (while increasing the complexity of the process), but the maximum number of distinct fluorophores that can currently be distinguished is around ten.

This is where the work of Schubert *et al.* comes in. The authors have developed a robotic immunofluorescence microscopy system, called multi-epitope ligand cartography (MELC), that can sequentially image as many as 100 distinct antibodies in the same sample (**Fig. 1**). This is accomplished using only one fluorophore by repeated rounds of staining (with different antibodies), imaging and photobleaching. After appropriate registration of a sequence of images, an image of dozens to hundreds of proteins in the same tissue or cell can be obtained. Their paper not only presents this exciting technology, but also contains an approach for partially analyzing the enormous

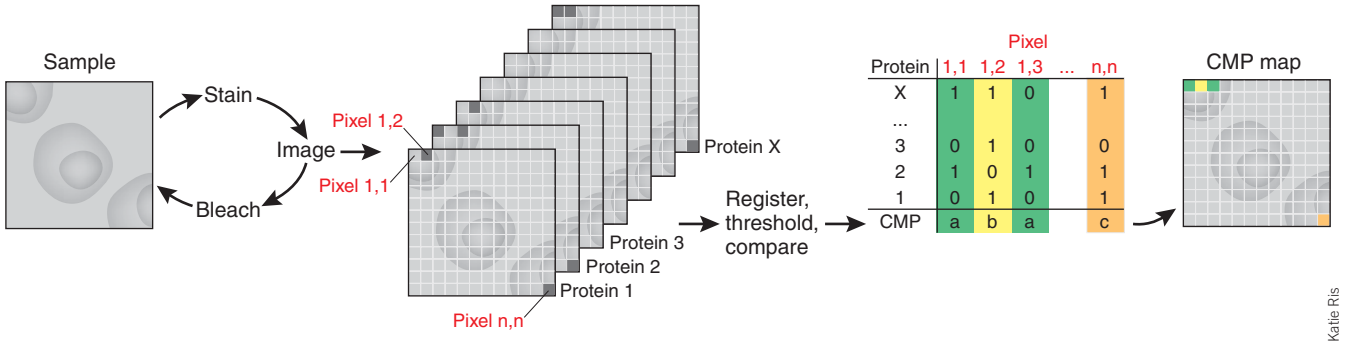

Figure 1 Multi-epitope ligand cartography (MELC) technology. MELC technology starts with a fixed and mounted tissue or cell sample and subjects it to fully automated cycles of fluorescent staining, imaging and photobleaching. Each cycle can use an antibody specific for a different protein (or any other fluorescent probe) so that the result is a set of images of the distributions of many different proteins for the same field. After thresholding each of these images, each pixel can be represented as a vector of binary values (a row in the example table) indicating whether it is positive or negative for a particular protein. Each unique combination defines a combinatorial molecular phenotype (CMP), and the spatial pattern formed by all the pixels with the same CMP can be identified.

complexity of pattern information present in the acquired images.

As an initial step, the authors sought to identify unique combinations of proteins that were found in a single pixel. They did so by using a threshold to decide whether each pixel was positive or negative for each protein, and then creating a binary vector of length equal to the number of proteins analyzed (**Fig. 1**). They call these vectors combinatorial molecular phenotypes (CMPs). Thus each pixel in an image can be described in terms of which unique combination of proteins it contains. Not surprisingly, they did not observe all possible combinations (for 48 proteins this would be 2^{48} CMPs), but the number observed (in the hundreds of thousands for 48 proteins) suggests a higher degree of complexity of protein location than had previously been appreciated.

Even more significantly, the presence or absence of particular CMPs may reflect different cell or tissue states. For example, the authors identified CMPs in skin images that can distinguish between patients with psoriasis, patients with atopic dermatitis and healthy individuals.

Although the CMP analysis just scratches the surface of the information content in the MELC images, it reveals the value of this new technology both for top-down detection of altered states and for bottom-up understanding of protein location families. By permitting the acquisition of correlated images of a large number of proteins, the MELC technique can provide an important complement to current approaches. For fixed cells or tissue sections, it provides the multiplexing that other methods lack. Other fluorescence microscopy methods, by contrast, can provide temporal information and high spatial resolution in live cells (without requiring isolation of specific antibodies).

An additional complementary approach to determination of location is the prediction of location from sequence. The higher-complexity, higher-resolution information coming from automated microscopy will allow training of substantially improved prediction systems, which in turn can be used to guide determination experiments. I anticipate that all of these tools can be used in a data-driven planning process to learn everything we need to know about the locations of proteins.

1. Schubert, W. *et al. Nat. Biotechnol.* **24**, 1270–1278 (2006).
2. Ideker, T., Galitski, T. & Hood, L. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
3. Jarvik, J.W. *et al. Biotechniques* **33**, 852–867 (2002).
4. Huh, W-K. *et al. Nature* **425**, 686–691 (2003).
5. Taylor, D.L., Woo, E.S. & Giuliano, K.A. *Curr. Opin. Biotechnol.* **12**, 75–81 (2001).
6. Price, J.H. *et al. J. Cell. Biochem.* **39** (Suppl.), 194–210 (2002).
7. Chen, X. & Murphy, R.F. *Int. Rev. Cytol.* **249**, 193–227 (2006).
8. Sigal, A. *et al. Nat. Methods* **3**, 525–531 (2006).
9. Chen, X., Velliste, M., Weinstein, S., Jarvik, J.W. & Murphy, R.F. *Proc. SPIE* **4962**, 298–306 (2003).
10. Perlman, Z.E. *et al. Science* **306**, 1194–1198 (2004).

Complementary therapies for inflammation

Yi Wang

Monoclonal antibodies against the human C5a receptor offer new therapeutic prospects for complement inhibition.

As one of the most potent mediators of inflammation, the complement component C5a is a promising target for anti-inflammatory drugs. However, a drawback of using animal models to develop antagonists of complement activation is that interspecies differences in receptors and ligands may complicate drug development and delay the evaluation of their efficacy in humans. In this issue, Lee *et al.*[1] generate mice in which the native C5a receptor has been replaced by its human counterpart and use these animals to produce high-affinity monoclonal antibodies (mAbs) capable of preventing and reversing inflammation in a mouse model of rheumatoid arthritis. This strategy may find broader application in raising antibodies to cell-surface targets and facilitating more meaningful preclinical trials.

The activation of complement is central to defense against infection and harmful stimuli. However, as activated complement components do not discriminate between self and non-self, inappropriate complement

Yi Wang is at Alexion Pharmaceuticals, Inc., 352 Knotter Drive, Cheshire, Connecticut 06410, USA
e-mail: wangy@alxn.com