Proteins on Location

Location proteomics is the product of a happy marriage between high-resolution imaging and sophisticated computational tools. By systematically relating function to location, the young field could provide keys to understanding the function of proteins in cells.

By Jen Crebs

Knowing where people spend their time is an important factor in knowing something about what they do. Similarly, identifying a protein's cellular address or location helps shed light on what that protein's doing in a cell. Devising tools and refining methods to achieve this in a high-throughput and systematic way is the objective of a young field called location proteomics.

The proteome of every cell is comprised of thousands of protein types featuring different and distinct interactions, biochemistry, and locations. Cellular processes can either cause or result from variations in the proteome, making the study of protein localization integral in determining function. But because proteins don't stick around at one cellular address for long, getting a handle on their travels — within cells, across cell types, and at different time points — requires sophisticated data gathering and analysis tools.

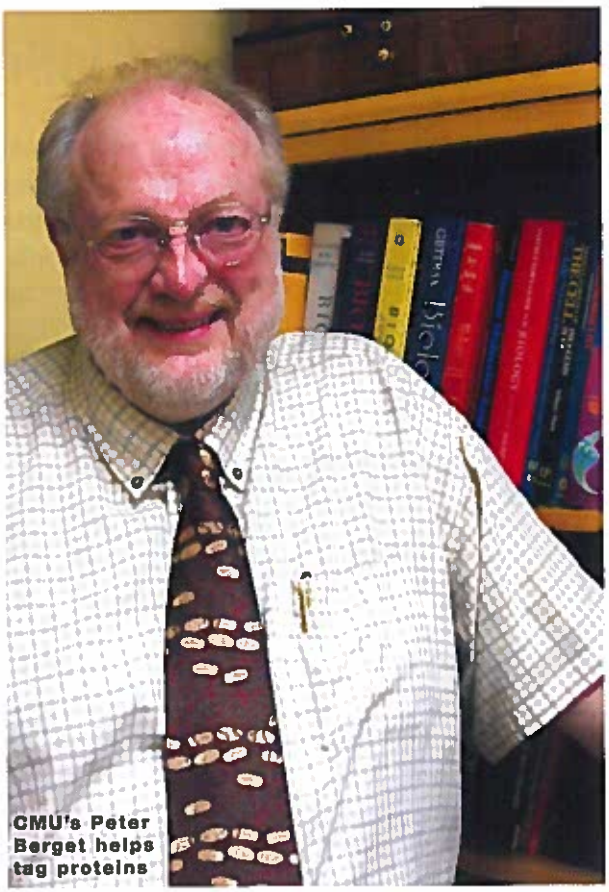
Location proteomics is concerned with systematically describing and relating protein location within cells, and researchers interested in this problem use a host of experimental and computational methods to accomplish it. "The vision behind what we're trying to go toward in location proteomics is to [perform] analyses of protein patterns systematically and automatically," says Robert Murphy, professor of biological sciences and biomedical engineering and director of the Center for Bioimage Informatics at Carnegie Mellon University.

Building systems-level frameworks to sort through protein localization patterns is Murphy's specialty. To acquire data on those patterns in the first place, and to do so on a large scale in living cells, is a process that has been refined by two of Murphy's colleagues, Peter Berget and Jon Jarvik, also professors at Carnegie Mellon. Together, the three have cast a wide net in capturing proteins, describing them in objective terms, and devising computational tools to cope with the resulting mass of data.

The field may have been baptized in Pittsburgh, but the analysis of protein patterns is being approached from various quarters. Groups are looking at how protein distributions change over time, during development, in disease states, and in the presence or absence of drugs. Murphy notes that location proteomics is "where the [Human] Genome Project was 20 years ago or more" in that data are still needed on "every protein for most cell types under the most important conditions."

TAG AND CAPTURE

Data collection in location proteomics



CMU's Peter Berget helps tag proteins

requires methods to visualize and capture images of proteins. Comprehensively detecting proteins is a question of devising the right tagging protocol, after which sets of images can be generated by high-resolution, confocal microscopy.

"Proteins can be detected in any number of ways that basically divide into two approaches: gene tagging and antibody-based approaches," Murphy says. Gene tagging involves introducing a tag or epitope into all of the expression genes of a particular cell type, with the result that all proteins in the live cell are tagged. Antibody-based methods rely on making specific antibodies for every protein in the cell; typically, antibody-based approaches require that the cell be fixed prior to imaging.

Berget and Jarvik developed a protein tagging protocol that has the advantage of maintaining live cells and native regulation. Jarvik, who initially developed the method, named it Central Dogma-tagging because it tags all molecular classes (DNA, RNA, and protein) referred to in the Central Dogma of molecular biology. In CD-tagging, a DNA sequence encoding a fluorescent protein is inserted into the intron of a target gene. Unlike epitope tagging, CD-tagging targets genomic DNA, so that the tagged gene retains the sequences needed for natural gene expression.

The idea for tagging proteins this way grew out of Jarvik's earlier work on the green algae *Chlamydomonas reinhardii*, when he was faced with the problem of trying to study an organelle that was difficult to purify. Since the organism featured "plenty of introns, and the technique targets introns, it came to me that we could do this," he says. "We could put a tag, in principle, in any intron whose gene contained one or more protein in the coding sequence. And that's most genes." Bigger genomes containing more introns are particularly well suited to CD-tagging, "so it just sort of became irresistible not to take it into the mammalian system."

According to Berget, "The original GFP CD-tagging vector was capable of placing a tag into the middle of a protein for only one particular class of introns in mammalian genes, and there are three classes of introns." It was a natural move for Berget to make two new vectors to tag all three types of introns, which paved the way to target more genes.

Thanks to the advent of 3D microscopy and new tagging techniques, researchers are "looking at many, many tens of thousands of proteins in parallel in dozens of cells."

Since then, Berget has further modified the tool by incorporating different fluorescent protein modules and smaller peptide tags into the CD-tagging vectors. "One of the benefits of that is it allows you to tag proteins with reagents other than standard GFP," he says, allowing for the use of "different types of fluorescent tags or chemical tags to be attached to proteins after they're synthesized in cells."

"We are also trying to develop vectors that'll deliver tags that ... after you've studied them, you can reverse them so that the tags are no longer present," Berget says. Making use of the Cre-lox recombination system, Berget's designer vectors allow him to tag a cell and then treat it with a certain reagent so that the tag disappears, in order to make sure the cell behaves normally before and after tagging as a control for future research. On the vector front, he is also currently working on "developing a partial mirror set of vectors using lentiviruses ... because they can infect a much broader range of cell types."

"Certainly what we've found [is] that the tagged proteins almost always localize where they should," Jarvik says. "There's always a concern that the tag will interfere with a phenomenon like alternative splicing. But the system deals with that. The average human gene has about eight introns in it — eight different places where our tag can be. If you see similar localization for the protein, whether it's tagged in any of those eight places, that is itself sort of a control. It suggests you're not seeing anything artifactual."

Fluorescence microscopy has long been the method of choice to track proteins at static points in their cellular commutes. "In the beginning, we would look at a single cell and a single protein, and do so for a series of cells," explains Roland Eils, a computational biologist the University of Heidelberg. Thanks to the advent of 3D microscopy and new tagging techniques, researchers like Eils are "looking at many, many tens of thousands of proteins in parallel in dozens of cells."

To get a close look at the cell with the expressed tagged protein in it, spinning disc confocal microscopy is used. "It's not like electron microscopy in terms of resolution," Jarvik says, "but for what you can get out of light microscopy, it delivers very high-resolution localization for that protein in the cell." Since live cells can be made visible with CD-tagging techniques, Jarvik and others

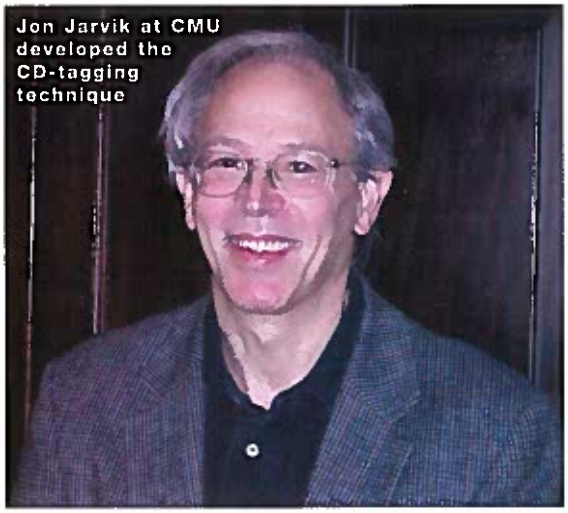
have been using time-lapse imaging techniques by which proteins can be observed continuously moving about and changing distribution during cell cycle events.

IMAGE ANALYSIS

Categorizing locations for different proteins in a fully automated way requires the development of high-throughput data-mining tools to sort through the mass of images. This involves choosing image features — typically in large numbers — and feeding them into a software program capable of recognizing protein distribution patterns. The analysis tools involved combine pattern recognition with machine-learning methods in order to objectively describe protein location.

Carnegie Mellon's Murphy is particularly interested "in the fact that there are far more subtle differences in the distribution of proteins in cells than can be appreciated by eye, and can be described in a limited sense, like the names of organelles." Although one can certainly get information on a protein by looking it up in a repository such as Entrez or SwissProt, he says, existing terms used to refer to protein location are inherently limited in describing previously unknown patterns or complex locations. For example, he points out that "with Gene Ontology Consortium terms, it's very difficult to say 'Well, this protein is present on the rims of the cis-Golgi cisternae.'"

One of the goals of Murphy's computational work is to extract more information on location patterns by implementing systems that can automatically identify what's important in a particular sample. That is, "to learn what features of a distribution, what patterns in a culture are indicative of whatever conditions have been imposed on that sample, whether that's drugs or some expression of various genes... our goal is to support movement toward higher-resolution imaging — to increase that content even further," he says.


Jon Jarvik at CMU developed the CD-tagging technique

The approaches Murphy has developed rely on standardized features that describe a protein's location in a cell image. These subcellular location features measure qualities such as a protein's shape, texture, wavelength decompositions, and edge qualities. Using a set of established features, Murphy's freely accessible software can identify and learn how to group "subtle differences in protein pattern" found in "images of varying quality," he says.

Clustering proteins based on subcellular location features brings in the use of machine-learning tools to construct objective groups of proteins based on their location. Murphy, Jarvik, and Berget collaborated on a project to explore clustering methods using images of NIH 3T3 clones obtained via CD-tagging. Automating the clustering of proteins by their fluorescent images outperformed visual methods to locate proteins, the team found.

"Rather than trying to restrict the analysis to particular protein patterns or particular organelles that are known to exist, we allow the computational analysis to find the patterns that are present," Murphy says of the clustering approach. "This allows us to not be restricted to just a small set of compartments or labels that we can put on each protein, but rather be able to truly identify the groups of proteins that are sharing a single location."

Another development in Murphy's work concerns separating two subcategories within a pattern, an approach he calls "unmixing." He finds that subcellular location features can be used to compare two cell images of the same protein or to estimate average feature values from many images of two different proteins. This leads to an objective way of measuring degrees of similarity in location patterns between proteins, a critical step toward overcoming inherent limitations of vocabulary-based classification schemes.

FURTHER HURDLES

Moving forward, there are plenty of areas of improvement. "One of the largest challenges on the computational side is the sheer amount of data we have to handle," says Roland Eils. He estimates that his lab's experiments produce "three to five terabytes of data every month." Storage is not the problem, but handling and processing the data is a tremendous challenge, he says.

Murphy says that a major area of development in location proteomics is learning how to compare results from one cell type to another. The clustering of proteins reveals patterns specific to a particular cell type. "You can do the same process in another cell type and end up with another set of clusters, and the question to ask is how can you link those two things by virtue of the underlying protein," he says. Murphy's group has made some progress in training classifiers to recognize patterns across cell types, but the data is still preliminary.

Eils agrees with Murphy. "Almost all of our studies are based on artificial cell systems," like HeLa lines or other immortalized cell cultures. "We all hope that this will help us understand the function of the protein ... in artificial cell systems, but everybody knows that ultimately we have to do these kinds of studies in primary cells. This is not at all easy." GT