Meeting report

Assembling a jigsaw puzzle with 20,000 parts

Edward M Marcotte

Address: Department of Chemistry and Biochemistry and Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, Austin, TX 78712, USA. E-mail: marcotte@icmb.utexas.edu

Published: 23 May 2003

Genome ***Biology*** 2003, **4:**323

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/6/323

A report on the Keystone Symposium 'Proteomics: Technologies and Applications', Keystone, USA, 25-30 March 2003.

The Keystone Symposium 'Proteomics: Technologies and Applications' might have passed as a joint convention of puzzle fanatics and auto mechanics. Analytical chemists, computational biologists, and geneticists rubbed shoulders as they presented the 'bleeding edge' of proteomics, a field at whose heart is a central question: how do we take the parts lists of genes from genome-sequencing projects and learn how the thousands of encoded proteins work and interact in the cell? Speakers at the symposium presented ever more sophisticated approaches to tease out subtle relationships among proteins, always with an eye towards understanding how to cope with the copious data generated by high-throughput approaches.

Tony Pawson (Mount Sinai Hospital, Toronto, Canada) gave the opening address, posing the question, which would recur during the conference, of whether the prevalence of interacting protein domains allows cells to create new signaling pathways more easily. To illustrate this idea, Pawson described the creation of a functional signaling pathway that does not exist naturally. In normal mice, receptor tyrosine kinases are involved in transmitting signals for cell growth, whereas death receptors provide the signal for apoptosis. Pawson described 're-routing' receptor tyrosine kinases, by changing their intracellular interaction domains, leading to a swap in the receptors' signaling targets and thereby inducing apoptosis from a receptor that normally signals growth. This notion of reprogramming signaling pathways by changing interaction domains was later reinforced by Alissa Resch (University of California, Los Angeles, USA), who presented tantalizing evidence that one function of alternate gene splicing might be to activate different subnetworks of interacting proteins by selectively adding or removing protein interaction domains.

Pawson also described in some detail an approach for creating transgenic mice that appears extremely promising for scaling up functional studies of mouse genes. Using *in vivo* expression of plasmid-based short hairpin RNA, Pawson demonstrated that he could down-regulate the GTPase-activating protein RasGAP p120 in embryonic stem cells. Subjecting the stem cells to tetraploid aggregation and subsequent embryogenesis leads to the formation of aberrant mouse embryos; this process is much quicker than the traditional route to generating knockout mice using homologous recombination, taking only about 3 weeks rather than months. The talk set the stage well for the conference to come: although it focused strongly on the underlying biology, it shifted - inevitably - to the technology required to study these ideas on a genome-wide scale.

Brute-force proteomics

As a field, proteomics is unashamedly high-throughput, and a clear trend at the symposium was the high-throughput direct assault on cellular systems, an approach exemplified by the talk given by Ruedi Aebersold (Institute for Systems Biology, Seattle, USA). He described the increasingly popular conception of proteomics not simply as the construction of a second parts list - the list of expressed proteins - but as an assay system. According to this idea, quantitative measurement of the proteome serves as a 'phenotype' for the cell and requires accurate, rapid, and replicable measurement of the concentration of each expressed protein. Ironically, although experimental technology such as mass spectrometry is currently capable of meeting these demands, the ability to interpret the data fully lags behind. In Aebersold's estimate, the technology has matured to the point where the experiment itself requires perhaps only around 15% of the researcher's time; approximately 10% of the time is spent analyzing the data to identify and quantify proteins while, in the major bottleneck at this stage, around 70% of the time is spent verifying that the protein assignments are correct, often by manual checking of the raw data. This need for manual analysis was echoed by Bruno Domon (Celera Genomics

Group, Rockville, USA), who described in detail the extensive manual data checking performed to verify mass spectrometry peptide assignments, which is aided in ambiguous cases by the addition of synthetic peptides to samples in order to confirm the identities of important proteins.

Pavel Pevzner (University of California, San Diego, USA) described computational efforts to sidestep this demanding manual examination of raw data. He presented the development of algorithms that can better identify peptides from experimental peptide fragmentation spectra, even in the presence of up to two post-translational modifications. Pevzner pointed out that one difficulty in this area of research is the dearth of publicly available mass spectrometry proteomics data, which complicates the development and testing of algorithms. A public repository of proteomics data could therefore potentially spur a great deal of research in this area, much as has happened for DNA microarrays and genome sequence data. A number of other talks addressed the need for quantitative benchmarking of proteomics results, including the talk by Josh Elias (Harvard Medical School, Boston, USA), who described how to interpret mass spectra by learning quantitative trends in the types of peptide fragmentation ions actually observed.

Fitting the pieces together

Along with large-scale efforts to characterize protein expression, efforts to define protein-interaction networks were also well represented. Stanley Fields (University of Washington, Seattle, USA) described an effort to develop a 'yeast two-hybrid-like' system specifically for membrane proteins, which have tended to be omitted from high-throughput assays. Fields' approach is to use a split-ubiquitin assay, in which the amino- and carboxy-terminal halves of ubiquitin are each fused to a set of approximately 700 yeast integral membrane proteins; the carboxy-terminal fusions also encode a LexA/Vp16 transcription factor. Interaction between a pair of membrane proteins reconstitutes ubiquitin, triggering proteolysis by ubiquitin-specific proteases, which releases the transcription factor and consequently stimulates transcription of a reporter gene. Early results show that the method successfully finds interacting membrane proteins, albeit with a reasonably high false-positive rate. Steven Michnick (Université de Montréal, Canada) described measuring protein interactions by using protein-fragment complementation assays together with a fluorescent assay. By reconstituting green fluorescent protein or dihydrofolate reductase, the latter detected via binding of a fluorescent substrate, the cellular location of the protein interactions could be observed, suggesting that, in addition to cataloging protein interactions, it may also soon be possible to follow their spatial and temporal dynamics in the cell.

Scaling up two-hybrid screens to measure protein interactions across a complete proteome presents numerous problems, including that of cloning and expressing the thousands of necessary proteins. Marc Vidal (Dana Farber Cancer Institute, Boston, USA) described the current state of all-versus-all protein-interaction screening in the nematode *Caenorhabditis elegans*. More than 12,000 of the approximately 20,000 predicted open reading frames in *C. elegans* have now been successfully cloned from a cDNA library into recombinational cloning vectors. Vidal reported that by transferring clones into yeast two-hybrid expression vectors, his group has identified roughly 1,500 interactions between more than 1,200 proteins.

A highlight of the protein-interaction talks was the presentation by Michael Snyder (Yale University, New Haven, USA), who detailed the comprehensive biochemical assaying of yeast proteins. Having cloned 5,800 yeast proteins and expressed them as fusion proteins, then affinity purified the products using the fused tag, the proteins have been printed onto microarray slides. Although a large fraction of the proteins are no doubt improperly folded, the resulting array had a sufficient fraction of properly expressed and folded proteins that they could be assayed for biochemical or binding activities. For example, screening with calmodulin revealed 33 new binding targets, while screening with 14-3-3 proteins revealed 140 new binding partners, several of which were independently verified by co-immunoprecipitation.

The power of such protein microarrays for systematic assays was reiterated by Dolores Cahill (Max Planck Institute for Molecular Genetics, Berlin, Germany), who described creating microarrays with around 37,000 human proteins expressed from cDNA expression clones in *Escherichia coli*. Cahill screened the arrays with antibody-containing sera in order to characterize the spectrum of binding specificities within each serum sample. She speculated that in this way it would be possible to rapidly screen sera to detect autoimmune disorders or otherwise characterize the state of the immune system. Gavin MacBeath (Harvard University, Cambridge, USA) described the use of protein arrays for high-throughput screening of small-molecule inhibitors that disrupt protein interactions. He described how to scale up the search for active inhibitors and simultaneously characterize inhibitor specificity by constructing the arrays directly in the wells of microtiter dishes.

But what do the proteins do?

Teasing out the precise functions of proteins on such a large scale would seem to be an even more daunting task than following their expression or interactions. But a number of clever approaches were described that gave some indications as to how this problem will ultimately be approached. Andrew Fraser (Wellcome Trust Sanger Institute, Cambridge, UK) described the use of RNA interference (RNAi) to attempt to inactivate around 17,000 *C. elegans* genes. Through a curious trick of fate, when *E. coli* expressing

C. elegans genes are eaten by worms, the worms respond by suppressing expression of the corresponding genes. Fraser and colleagues cloned and expressed each worm gene in *E. coli*, fed the bacteria to worms, and assayed the worms for knockout phenotypes, detecting phenotypes for around 1,700 genes. Fraser described systematic analyses using these data of relationships between gene function and position, including the strong selection against essential genes appearing on the X chromosome and the tendency for neighboring genes to share RNAi phenotypes. Interestingly, Fraser noted that genes that are in the same SL2 ribonucleoprotein-dependent *trans*-spliced 'operons', the worm equivalent of bacterial operons as recently characterized by Stuart Kim and colleagues, often do not appear to share RNAi phenotypes, raising the question of whether worm genes in the same operon will generally be functionally related, as is the case for bacterial genes encoded in the same operon.

A simple, but powerful, functional assay was described by Benjamin Cravatt (Scripps Research Institute, La Jolla, USA) for profiling the set of enzymes with a given general type of catalytic activity. In this activity-based profiling approach, enzyme substrates were synthesized with a reactive group and a fluorophore, such that the substrate cross-links to the general class of enzyme acting upon this type of substrate. By adding these hybrid substrate reagents to cells grown under different conditions, the subpopulation of active enzymes with the appropriate specificity could be monitored or purified. Using this approach, Cravatt described the design of a potential analgesic to inhibit fatty acid amide hydrolase (FAAH), the serine hydrolase that regulates signaling by the endocannabinoid CB1. The active component of marijuana, tetrahydrocannabinol (THC), is a CB1 agonist, so Cravatt speculated that blocking FAAH might promote analgesia while avoiding the other side effects of THC. Current FAAH inhibitors cross react with many FAAH-like serine hydrolases, but by using activity-based profiling, Cravatt could screen FAAH inhibitors for their *in vivo* selectivity, and in this manner was able to identify a specific FAAH inhibitor. Interestingly, during this process, a selective inhibitor was developed for a second FAAH-like protease that has yet to be characterized, meaning that a specific inhibitor was developed for an enzyme before ever knowing the enzyme's natural substrate.

Three-dimensional puzzles

Another highlight of the symposium was a presentation by Joachim Frank (Wadsworth Center, State University of New York, Albany, USA), who described the current state of knowledge of the *E. coli* ribosome structure obtained using single-particle reconstruction by cryoelectron microscopy. More than 110,000 individual ribosome images have now been integrated into Frank's model structure, which has been fitted with the recently solved X-ray crystal structures of ribosomal proteins to create a hybrid atomic-resolution/cryo-EM ribosome structure. Frank described the complex movements taking place in the ribosome during the processes of peptide-bond formation and ribosome stalling, which were revealed in beautiful modeling movies. Perhaps most striking was the extent to which the ribosomal proteins appeared to shift positions, suggesting they might play active roles in the mechanical movement of the ribosome.

A similarly detailed picture of the yeast nuclear pore complex is not yet available, but Andrej Šali (University of California, San Francisco, USA) and Brian Chait (Rockefeller University, New York, USA) described progress in this area. By combining Chait's experimental data on the distribution of proteins of the nuclear pore complex with other constraints such as estimated protein sizes and known protein interactions, Šali described the computational construction of a nuclear pore complex model that satisfied all of the available constraints, including the known symmetry of the complex. The model is a work in progress, but the approach taken allows for the inclusion of many diverse forms of experimental data, suggesting that it can only become more accurate as additional data are added.

Finally, starting from a protein's structure, David Baker (University of Washington, Seattle, USA) described how one might begin to design protein-protein interactions. Starting with an algorithm that finds amino-acid sequences that pack well into a fixed protein structure, Baker described how the approach could be generalized to create a new interface between two DNA-binding domains of known structure. Experimental construction of the designed proteins confirmed that they do in fact interact and bind to a novel DNA sequence.

In summary, progress in proteomics has been rapid, and the sheer scale of the projects presented makes for news in itself. The field is still quite far from solving the proteomics jigsaw puzzle, however, and researchers are grappling with understanding the emergent complexity. In the closing words of conference organizer Aebersold, "We don't really know what it all means, but I'm sure it's interesting".