

# The protein microscope: incorporating mass spectrometry into cell biology

Alexander W Bell, Tommy Nilsson, Robert E Kearney & John J M Bergeron

Mass spectrometry has come into its own as an extremely powerful tool for the study of whole proteomes. So why are not more cell biologists embracing it with open arms?

Mass spectrometry-based proteomics has been at the core of several noteworthy advances in cell biology, particularly in the assignment of the protein composition of cellular organelles<sup>1,2</sup> and the study of protein interaction networks<sup>3</sup>. The combined strength of being able to both identify and quantify the molecular players in a given process provides essential information that no other technology can now parallel.

For example, through mass spectrometry we now know the composition of the nuclear pore as well as its evolutionary conservation across species<sup>4</sup>. This constitutes a complete, accurate and permanent (CAP) proteome, a goal-oriented definition coined by Sir Sidney Brenner (see acknowledgment in ref. 1). Several near-CAP proteomes have also been achieved recently (for reviews of the nuclear pore CAP proteome and other near-CAP proteomes, see refs. 1,2), including components of the secretory pathway, synaptic vesicles, clathrin-coated vesicles and the nucleolus. To make the leap from near-CAP to CAP status will take considerable effort but is achievable in the long term. If done properly and with quantitative data, such CAP proteomes could serve as the foundation for future biology.

The human body has some 20,500 protein-coding genes<sup>5</sup>. Knowing when, where and how much of these are expressed at the protein level is an important goal in the post-genomic era. Furthermore, insight into states of post-translational modifications and protein interactions can also be obtained through mass spectrometry analysis, providing a molecular basis for additional analysis by complementary approaches such as imaging, genetics and molecular biology.

Given the enormous potential of mass spectrometry-based proteomics, it is perhaps surprising that so few cell biologists have fully embraced this technique. Is it because proteomics so far has had a very limited impact in hypothesis-driven research, because it is not an easily accessible technique, or is it because it is simply lacking in charm compared to, for example, live-cell video microscopy? With imaging, we can observe dynamic processes—something that proteomics has a hard time rivaling. This does not suggest, however, that one is more or less useful than the other. Rather, imaging and proteomics technologies are complementary tools, and mass spectrometry can lead to biological insights, especially when imaging proves difficult. For example, vesicular transport carriers such as the clathrin-derived, COPI or COPII vesicles are virtually impossible to observe in living cells. This is not because of their small size, but because they are usually close to much brighter membranes and, like orbiting planets of distant stars, very difficult to discern. Electron microscopy has provided us with firm evidence of the existence of these vesicles, and because they can be biochemically purified, they have been excellent material for proteomics studies<sup>6–8</sup>. These studies have already provided important insights

into their nature and function, and together with future studies should help answer some long-standing questions: what do they contain, how are they formed, transported and targeted?

Sample preparation is perhaps another main reason why cell biologists view proteomics with some caution. This is well-founded, and we go as far as to state that many proteomics efforts suffer from a lack of rigor<sup>9,10</sup>. Perhaps surprisingly, rather than acknowledging this, the proteomics community has instead attempted to circumvent the need for rigorous sample preparation by developing advanced bioinformatics tools. But this is only a very limited substitute for proper sample preparation, and when combined with problems of data presentation that effectively prohibit critical assessment, such efforts have had a limited impact in the cell biology community. The proteomics field needs to address these issues to earn much-needed credibility, not only from cell biologists but from the larger scientific community as well.

Present limitations of the mass spectrometry instrumentation itself and the issue of dynamic range are other general problems. To some extent, the dynamic range can be boosted by additional sample preparation strategies including biochemical extraction or biological refinement. For example—by a high-salt wash to remove peripheral proteins, followed by detergent extraction and detergent-phase separation to separate integral membrane proteins from soluble proteins—the proteome of the endoplasmic reticulum was extended from about 1,237 proteins to 1,700 proteins with three orders of magnitude between the most abundant and least

Alexander W. Bell and John J.M. Bergeron are at the Department of Anatomy and Cell Biology, McGill University, Montreal, Quebec H3A 2B2, Canada.

Tommy Nilsson is at the Department of Medical and Clinical Genetics, Institute of Biomedicine and the Proteomics Centre at the Sahlgrenska Academy, Göteborg University, 413 90 Göteborg, Sweden.

Robert E. Kearney is at the Department of Biomedical Engineering, McGill University, Montreal, Quebec H3A 2B4, Canada. Alexander W. Bell and Tommy Nilsson contributed equally to this work. e-mail: john.bergeron@mcgill.ca

abundant protein<sup>7</sup>. This success is a good example of how the proteomics community could incorporate the more than half a century of accumulated know-how and tradition that already exists in the biochemistry and cell biology fields to improve sample preparation for proteomics studies.

Yet another source of confusion for cell biologists is that proteomics is not an exact technique, even though it is often portrayed this way. Although high-resolution mass spectrometry yields mass measurements of peptide fragments with unprecedented accuracy, it is still a probability-based technique. Most proteomic platforms characterize, by tandem mass spectrometry, peptides resulting from trypsin digestion of protein mixtures. The high mass accuracy<sup>11</sup> of high-resolution tandem mass spectrometers assures a good probability (>95%) of unambiguously assigning about half the mass spectra to primary sequences<sup>7</sup>. Most platforms also assess false positive rates of assignment, which are usually <2% (ref. 12). The output is a list of peptides, which are then grouped into protein sequences by database-matching algorithms.

All presently available databases are fraught with redundancies, inconsistencies in nomenclature, fused genes, inappropriately translated introns and so on. In many cases, nondescriptive (for example, a hypothetical protein or open reading frame) or multiple names are erroneously assigned. This introduces another uncertainty factor apart from the fact that the investigator must mount a Herculean effort to sort this out. As such, proteomics cannot presently be considered a stand-alone and absolute method. Rather, it serves as an efficient tool to generate a molecular framework for additional testing and validation, at least until the bioinformatics community provides more reliable databases.

It is the cell biologist's role to assure that protein assignments are accurate. For protein families in which individual proteins may share >90% sequence identity, a BLAST-like alignment tool (BLAT) alignment of the back-translated protein sequence to the genome may be necessary to distinguish between homologs and isoforms. Protein entries in the protein database corresponding to fused genes and related gene-prediction problems (for example in- and out-of-reading-frame sequencing errors, or incomplete editing of intronic sequences) are normally recognized by BLAST analysis and must be corrected manually.

In an attempt to solve the problem of data presentation and to help the cell biologist sort through a morass of data, we have established a workable framework understandable to cell biologists, called CellMapBase<sup>13</sup> (Supplementary Fig. 1 online). A key element in its design is that it is comprehensive as it incorporates information defining the biological origin, the details of protein separation and in-gel tryptic digestion, mass spectrometry raw data and peak lists, peptide identification and subsequent protein identification, and automatic and manual annotation of the results. It also indexes proteins according to their primary sequence and provides cross-references to the nonredundant protein sequences (nr) database from the US National Center for Biotechnology Information (NCBI), the Universal Protein Resource (UniProt) and the International Protein Index (IPI). This helps to overcome the constant changes in public sequence databases. CellMapBase is, however, only a step toward an intuitive framework and interface that ultimately should enable biologists to explore and extract reliable and meaningful data with the same ease as getting a cup of coffee.

The relational aspect of the CellMapBase makes it possible to treat data from many experimental conditions as a single set of data. This facilitates the establishment of a common set of proteins for the purpose of comparison by grouping the data into supersets according to biologically relevant features while retaining all the underlying data. These supersets can then be broken down or subgrouped according to some aspect of the experimental design. Additionally, CellMapBase also retains information about each peptide in each group or subgroup so that estimating relative protein abundance is possible<sup>7</sup>, which can provide an initial quantitative overview of protein distribution<sup>14</sup> and the degree of contamination. For example, a mitochondrial protein is unlikely to belong to a purified Golgi fraction. Nevertheless, cell biologists are aware that purified Golgi fractions invariably contain contaminating membranes and proteins such as those from mitochondria. By calculating the relative abundance of mitochondrial proteins among different subcellular fractions, expressing the degree of contamination as a percentage becomes straightforward<sup>7</sup>.

The ultimate goal of developing a spatial and quantitative map of all protein members in all tissues and subcellular

compartments cannot be reached unless cell biologists, bioinformaticians and mass spectrometrists work together to address the problems and limitations that exist today. Too many proteomics efforts run the risk of becoming obsolete if no biologist can or will use the data. Also hindering proteomics efforts is a lack of standards and of a unifying framework.

No strategy presently exists to bring data together nor to ensure that they have a meaning tomorrow. Who should do this? We think that the cell biologists, having some experience in bridging different cultures—like yeast genetics with biochemistry and developmental biology—should have an active role in addressing this challenge. Funding agencies also have a responsibility to ensure that public and private funds are spent in a manner that ensures some degree of permanence and reliability in proteomics-based projects, and to help pave the road to the ultimate goal of mapping the entire human proteome. This will require a coordinated international effort, a challenge that perhaps the Human Proteome Organization is willing to take on—or at least coordinate.

Note: Supplementary information is available on the Nature Methods website.

#### ACKNOWLEDGMENTS

We gratefully acknowledge support by the Canadian Institutes for Health Research, Canada Foundation for Innovation, Natural Sciences and Engineering Research Council of Canada, Genome Quebec and Genome Canada, McGill University, Göteborg University, the Swedish Research Council and the Knut and Alice Wallenberg foundation. We thank A. Gilchrist, C. Au and Z. Bencsath-Makkai for help with this manuscript.

1. Au, C.E. *et al. Curr. Opin. Cell Biol.* **19**, 376–385 (2007).
2. Yates, J.R. III, Gilchrist, A., Howell, K.E. & Bergeron, J.J. *Nat. Rev. Mol. Cell Biol.* **6**, 702–714 (2005).
3. Collins, S.R. *et al. Mol. Cell. Proteomics* **6**, 439–450 (2007).
4. Cronshaw, J.M., Krutchinsky, A.N., Zhang, W., Chait, B.T. & Matunis, M.J. *J. Cell Biol.* **158**, 915–927 (2002).
5. Pennisi, E. *Science* **316**, 1113 (2007).
6. Blondeau, F. *et al. Proc. Natl. Acad. Sci. USA* **101**, 3833–3838 (2004).
7. Gilchrist, A. *et al. Cell* **127**, 1265–1281 (2006).
8. Takamori, S. *et al. Cell* **127**, 831–846 (2006).
9. Simpson, J.C. & Pepperkok, R. *Genome Biol.* **7**, 222 (2006).
10. Zhang, Z. & Chan, D.W. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2283–2286 (2005).
11. Zubarev, R. & Mann, M. *Mol. Cell. Proteomics* **6**, 377–381 (2007).
12. Elias, J.E., Haas, W., Faherty, B.K. & Gygi, S.P. *Nat. Methods* **2**, 667–675 (2005).
13. Bencsath-Makkai, Z. *et al. Conf. Proc. IEEE Eng. Biol. Soc.* **4**, 3567–3570 (2003).
14. Bergeron, J.J. & Hallett, M. *Nat. Biotechnol.* **25**, 61–62 (2007).