The need for a public proteomics repository

John T Prince, Mark W Carlson, Rong Wang, Peng Lu & Edward M Marcotte

Large-scale protein expression measurements promise to open exciting areas of research, ranging from systematic investigations of post-transcriptional regulation, to numerical models of protein synthesis and decay, to detailed functional analyses of the many proteins expressed by a cell. This promise is even greater when such protein expression data are combined with DNA microarray or protein interaction data—such combinations offer the tantalizing promise of quantitative models of cellular events. Unfortunately, the field of proteomics has been slow to embrace one of the most important lessons from DNA microarrays: open availability of raw data. The availability of DNA microarray data, coupled with public genome sequence data, is arguably one of the primary forces driving computational research in functional genomics.

A shortage of public data

The availability of protein expression data currently lags far behind microarray data. This is due to both technical and historical reasons. From a technical standpoint, protein expression profiling via two-dimensional gels is a challenging approach (even more so if quantification is desired[1]), whereas protein expression profiling by mass spectrometry (reviewed in ref. 2) is still a young technology in its early stages of adoption. Historically, protein mass spectrometrists have not distributed their raw protein or peptide mass spectra to the community, a puzzling omission given that many other types of mass spectra are available.

Much of our rich understanding of global gene expression patterns comes not only from the power of DNA microarrays to generate those data, but also from the fact that experimenters deposited their raw data into the public domain (*e.g.*, as in the Stanford Microarray Database (SMD)). The easy availability of mRNA expression data has led to countless computational analyses of the same data sets, each teasing out ever more subtle trends in the data. Examples abound of mining such data for new insights long after publication: evidence for post-transcriptional gene regulation was found by comparing newly measured protein expression levels to previously published mRNA expression levels[3]; a considerable source of variation among cell-cycle controlled genes' mRNA transcription[4] was attributed to the day upon which each microarray was analyzed[5]; microarray data were later reinterpreted to reveal cell cycle defects among the original cell populations[6]. Perhaps the strongest example of the long-lived utility of these data has been in the comparisons of mRNA expression patterns across hundreds of microarray experiments[7,8] to discover coexpressed systems of genes.

Astonishingly, in spite of the significant progress over the past decade in high-throughput protein expression profiling—notable examples include development of the isotope-coded affinity tags (ICAT) technology[9] and stable isotope incorporation[10,11] for quantitative proteomics, the multidimensional protein identification technology (MudPIT) for analyzing complex proteomes[12], high-throughput expression profiling of yeast proteins[13,14] and semi-quantitative expression profiling of *Plasmodium* proteins[15,16]—there are probably fewer than 20 mass spectrometry–based protein expression data sets in the public domain, none of which is stored in a central location, and even fewer raw data sets of protein mass spectra.

For example, the annual *Nucleic Acids Research* database issue (2004) lists 39 distinct databases dedicated to mRNA expression data, summarizing more than (roughly) 10^8 measurements of mRNA expression, as compared to two databases of two-dimensional gel electrophoresis protein expression data, with perhaps 10^3–10^4 expression measurements, and none with mass spectrometry–based protein expression profiles. As a consequence, interpretation of the original mass spectrometry data by anyone other than the original experimenters has been negligible, and proteomics, unlike genomics, has yet to see the many benefits gained by reanalysis of the data by computational and statistical researchers.

The way forward?

In principle, proteomics data can be analyzed by many of the same techniques as microarray data. Thus, the computational analyses proven so powerful for microarrays, such as clustering of genes and samples by their expression patterns, analysis of gene coexpression across experiments, and all manner of differential and comparative expression analyses, should be directly applicable to proteomics data. As the correlation between protein and mRNA expression levels seems to be relatively poor in eukaryotes[3,13,14], the promise is great for proteomics data to reveal many new trends among the genes. However, where proteomics really stands to gain is not at such high-level analyses, but in the actual interpretation of the mass spectra themselves.

Mass spectrometry proteomics data sets are currently analyzed with essentially the same algorithms developed 5–10 years ago to interpret mass spectra (*e.g.*, the SEQUEST[17] or Mascot[18] algorithms). While this longevity attests to their usefulness, the lack of competition is no doubt partly because of limited access to the mass spectra by the statistical community. In a typical high-throughput mass spectrometry experiment, far less than half of the spectra are ever satisfactorily interpreted (*e.g.*, only 17% of 162,000 mass spectra could be interpreted in

John Prince, Mark Carlson, Rong Wang, Peng Lu and Edward M. Marcotte are at the Center for Systems and Synthetic Biology & Institute for Cellular & Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA.
e-mail: marcotte@icmb.utexas.edu

a recent large scale analysis of the yeast proteome[19]), underlining the need for improved algorithms for interpreting these raw data, and for recognizing when proteins have been post-translationally modified and interpreting the mass spectra appropriately[20]. This is not to argue that advances haven't been made (see ref. 21), but clever statisticians who might have better interpretations of mass spectral data have a strong barrier to entering the field—there's virtually no public data. Again, this is in direct contrast to the DNA microarray field, where free release of the raw data has allowed many nonexperimentalists to step in and contribute. As a consequence, microarray hybridization statistics are now far better worked out (*e.g.*, see refs. 22,23), and more importantly, are transparent in a way that proteomics statistics often are not (as pointed out in ref. 2).

Such public data and transparent statistics are a major goal of numerous proteomics groups, and data standards for exchanging protein mass spectra are being developed (*e.g.*, PEDRo[24], PSI-MS XML[25], mzXML (http://sashimi.sourceforge.net)), led in part by a parent organization (The Human Proteome Organization (HUPO), Montreal, Canada) dedicated to pushing such standards forward. Ironically, in spite of a small number of individual investigators who have deposited protein expression data at various sites on the internet (links are given at the internet site listed below), the field now has multiple standards established for mass spectrometry–based proteomics data, yet little public data. In the spirit of open databases, such as Genbank, SwissProt, Pfam or SMD, each of which contributes greatly toward enabling research in computational biology, it would seem the time is ripe for a centralized proteomics database.

A centralized database

Mass spectra are accumulating at a phenomenal rate—ThermoFinnegan (Waltham, MA, USA) alone boasts sales of hundreds of ion trap mass spectrometers specifically for MudPIT-style proteomics experiments. Thus, we believe a centralized database would quickly be populated. As a gesture toward initiating a public repository and with the hopes of encouraging computational analyses of proteomics data, we've deposited a number of protein mass spectrometry data sets into the public domain in an Open Proteomics Database (OPD) at http://bioinformatics.icmb.utexas.edu/OPD.

The data residing in OPD represent diverse proteomics samples—some interpreted, some uninterpreted, some on simple but defined samples to be used for training algorithms, and some on highly complex samples, such as whole-cell lysates from differing organisms. In all, proteomics data from *Escherichia coli*, *Mycobacterium smegmatis*, *Saccharomyces cerevisiae* and *Homo sapiens* are represented, with roughly 400,000 total mass spectra, cataloging the expression of several thousand proteins overall. All data are freely accessible, with the intent that computational groups interested in studying the many computational problems posed by proteomics will have a source of protein mass spectra and expression data. OPD currently contains data from our own research group, with hyperlinks to other public protein expression data sets. We invite submissions of additional data from the community, particularly mass spectrometry analyses of whole proteomes or organelles, so as to begin building up the critical mass of public proteomics data.

1. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. & Aebersold, R. *Proc. Natl. Acad. Sci. USA* **97**, 9390–9395 (2000).
2. Aebersold, R. & Mann, M. *Nature* **422**, 198–207 (2003).
3. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
4. Spellman, P.T. *et al. Mol. Biol. Cell* **9**, 3273–3297 (1998).
5. Alter, O., Brown, P.O. & Botstein, D. in *Microarrays: Optical Technologies and Informatics*, vol. 4266 (eds. Bittner, M.L., Chen, Y., Dorsel, A.N. & Dougherty, E.R.) 171–186 (International Society for Optical Engineers, Bellingham, WA, 2001).
6. Lu, P., Nakorchevskiy, A. & Marcotte, E.M. *Proc. Natl. Acad. Sci. USA* **100**, 10370–10375 (2003).
7. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. *Science* **302**, 249–255 (2003).
8. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
9. Gygi, S.P. *et al. Nat. Biotechnol.* **17**, 994–999 (1999).
10. Jensen, P.K. *et al. Electrophoresis* **21**, 1372–1380 (2000).
11. Ong, S.E. *et al. Mol. Cell Proteomics* **1**, 376–386 (2002).
12. Washburn, M.P., Wolters, D. & Yates, J.R. 3rd. *Nat. Biotechnol.* **19**, 242–247 (2001).
13. Griffin, T.J. *et al. Mol. Cell Proteomics* **1**, 323–333 (2002).
14. Washburn, M.P. *et al. Proc. Natl. Acad. Sci. USA* **100**, 3107–3112 (2003).
15. Lasonder, E. *et al. Nature* **419**, 537–542 (2002).
16. Florens, L. *et al. Nature* **419**, 520–526 (2002).
17. Eng, J., McCormack, A.L. & Yates, J.R. 3rd. *J. Am. Soc. Mass. Spectrom.* **5**, 976–989 (1994).
18. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. *Electrophoresis* **20**, 3551–3567 (1999).
19. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J. & Gygi, S.P. *J. Proteome Res.* **2**, 43–50 (2003).
20. Pevzner, P.A., Mulyukov, Z., Dancik, V. & Tang, C.L. *Genome Res.* **11**, 290–299 (2001).
21. Boguski, M.S. & McIntosh, M.W. *Nature* **422**, 233–237 (2003).
22. Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. & Wong, W.H. *Nucleic Acids Res.* **29**, 2549–2557 (2001).
23. Yang, Y.H. *et al. Nucleic Acids Res.* **30**, e15 (2002).
24. Taylor, C.F. *et al. Nat. Biotechnol.* **21**, 247–254 (2003).
25. Hermjakob, H. *et al. Nat. Biotechnol.* **22**, 177–183 (2004).