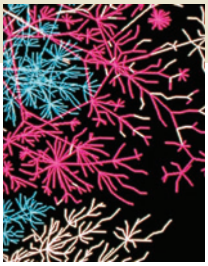

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

Toward data standards for proteomics

Veerasamy Ravichandran & Ram D Sriram

Biologists have traditionally worked with relatively small data sets and shared results only with others working on similar biological systems. With the advent of genomics, and now proteomics, they are faced with exponentially growing sets of highly interrelated, heterogeneous, complex and rapidly evolving types of data. The lack of cohesion between heterogeneous proteomics data, resulting from the diverse structure and organization of independently produced data sets, poses a serious problem for progress in the field. We argue here that the adoption of consensus standards for the interpretation, handling and dissemination of specific types of protein data should be a priority for the proteomics community.

Problems posed by proteins

With the completion of the Human Genome Project, the current trend in research is to focus on gene products, primarily proteins, and the overall biological systems in which they act, creating the emerging fields of proteomics and systems biology[1]. Characterizing the protein content of cells—proteomics—poses many challenges that genomic analyses does not.

The chemical properties of the nucleic acid bases are very similar, so separation and purification is relatively easy compared with protein separations, where proteins can have very diverse chemical properties, complicating handling, separations and identification. In addition, many proteins exist in very small quantities in cells and tissues, making them difficult to identify and analyze. Vanishingly small quantities of nucleic acid sequences can be amplified using PCR for accurate detection, whereas comparable amplification methods are not available for proteins. Even when the primary amino acid sequence of a protein is known, deducing its structure, function and/or interacting partners is far from straightforward. The behavior of proteins is determined by their tertiary structure, so an assay that is based on protein binding depends on maintaining the native conformation of the protein. This places constraints on the systems that are used to capture protein targets in affinity-based assays.

Furthermore, protein quantity is not necessarily correlated with function. Proteins can undergo a number of post-translational modifications that affect their activities and cellular location, such as metal binding, prosthetic group binding, glycosylation, phosphorylation and protease clipping, among others. RNA splicing can also produce a number of similar proteins that differ in function. Thus, a complete proteomics analysis must not only measure cellular protein level, but also determine how the proteins interact with one another and how they are modified.

Problems posed by proteomics

A brief survey of the proteomics literature highlights the problem in comparing results from different investigators, even when they are working on the same or similar problem. Very often, a large number of proteomics experiments have to be done, and even for the same investigator, repeating exactly a given result in a given experiment under a given set of conditions may be problematic, even though the same protocol, analytical instruments and controls are used. Many of the parameters in a proteomics experiment are hard to control, describe completely or replicate. Along with sample preparation, proteome characterization presents various technical challenges[2].

There are also many ways to represent a proteomics experiment and its associated results, which can lead to several data-interoperability problems. For mass spectrometry–derived data, for example, the file format is easier to standardize, but the criteria that are used to statistically 'identify' a protein differ greatly and are software dependent[3]. For example, in a typical liquid chromatography tandem mass spectrometry experiment, ~1,000 collision-induced dissociation (CID) spectra can be acquired per hour. Even with the overoptimistic assumption that every one of these spectra leads to the successful identification of a peptide, it would take considerable time to analyze complete proteomes. In an effort to ensure that significant data of high quality are entering the proteomics literature, the journal *Molecular and Cellular Proteomics* is introducing guidelines for authors submitting manuscripts containing large numbers of proteins identified primarily by liquid chromatography–coupled tandem mass spectrometry[4].

With the emergence of mass spectrometry and several other high-throughput platforms for rapidly analyzing proteins on a large scale, a wealth of computational tools has been developed to analyze the resulting data[5]. But because there is such a plethora of software tools available, the task of choosing the most suitable analytical platform for a particular data set is becoming ever more challenging. Despite the combined efforts of biologists, computer scientists, biostatisticians and software engineers, no one-size-fits-all solution is available for the analysis and interpretation of complex proteomics data, whether from two-dimensional gel electrophoresis, mass spectrometry, yeast two-hybrid studies or other analytical platform. The lack of cohesion between heterogeneous scientific data, resulting from the diverse structure and organization of independently produced data sets, is creating an impractical situation for data interoperability and integration[6].

Standards as solutions

Proteomics data require systematic mining, reformatting, annotating, standardizing and

Veerasamy Ravichandran and Ram D. Sriram are at the Manufacturing Metrology and Standards for the Health Care Enterprise Program, Manufacturing Systems Integration Division, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899-8263, USA.
e-mail: vravi@cme.nist.gov

integration in a unified computational framework. In the context of time-to-completion pressures and the sheer volume of data being generated, the proteomics community therefore needs to rapidly reach consensus in several areas (some general types of standard are summarized in **Box 1**).

The first area that needs to be addressed is best practice in terms of controls for specific types of proteomics experiment (e.g., mass spectrometry, two-dimensional gels) and how those results are presented in the literature and in databases. Second, the community needs to come up with certified models that can derive 'best' recommended values from critically evaluated experimental data and validated benchmarked predictive methods for any real or proposed measurement. Lastly, effective data management standards and techniques (e.g., quality, traceability or uncertainty estimates) are required for gathering, integrating and maintaining information about disparate types of proteomics data accessed from diverse sources.

Already, some steps are underway to address the first problem. Last year, an editorial in *Molecular and Cellular Proteomics*, for example, proposed draft guidelines to improve the quality and statistical robustness of peptide and protein identification data associated with mass spectrometry papers[4]. These include documenting the search engine used and the way protein assignments were made using a particular algorithm, defining how peptides should be counted toward protein identification, increasing the stringency of information required to use single-peptide identification for protein assignment, and minimizing redundancy by ensuring that a protein with different names and accession numbers in different databases is reported only once.

The second problem requires the development of standard reference data (SRD), which can be used to calibrate data and standardize procedures for conducting proteomics experiments. The National Institute of Standards and Technology (NIST) provides SRD (http://www.nist.gove/srd/) in a variety of domains that could have a significant role in providing SRD for proteomics, in addition to the protein data bank. For example, output from data-mining algorithms or mass spectrometry studies can be compared against appropriate NIST SRD.

In terms of assessing data quality, correctness and completeness, proteomics data present a particular challenge. As the sensitivity and power of analytical technology progresses, it is likely to become apparent that existing data sets may be incomplete or of insufficient quality. For example, a mitochondrial preparation can contain nuclear proteins as contaminants, which can subsequently be entered into databases and erroneously annotated as proteins of mitochondrial origin. How can such data be subsequently purged and corrected?

To optimize the management and sharing of proteomics data, common standards and ontologies must be adopted. These controlled vocabularies also need to change with time, as new information and insights into proteomics data are obtained. Even if there were a perfect vocabulary that 'got it right the first time,' it would have to change with the evolution of scientific knowledge. All too often, vocabularies change in ways that are convenient for creators but wreak havoc with users.

In certain respects, the proteomics community can draw on precedents in other fields of biology that have faced similar problems in annotating and unifying diverse and heterogeneous data sets. Use of controlled vocabulary is already facilitating analysis of several types of high-throughput data, including the Health Level 7 (HL-7) standard exchange format for sharing clinical data[7], the macromolecular Crystallographic Information File (mmCIF) for sharing macromolecule crystallographic data[8] and the Microarray Gene Expression Database (MGED) Group's MIAME (minimal information for a microarray experiment) model for sharing DNA microarray data[9].

Box 1 Types of data standards

Data standards provide a well-defined syntax with precise definitions and examples. They employ ontologies to define the basic terms and relationships comprising the vocabulary of a topic area, as well as the rules for combining terms and relationships to define extensions to the vocabulary. In addition, standards incorporate data relationships, data types, range restrictions, allowed values, interdependencies, exclusivity, units and methods. They are generally required when excessive diversity, as in proteomics data, creates inefficiencies or impedes effectiveness. Standardized data should have the following characteristics: first, completeness, comprehensiveness, consistency, reliability, and timeliness; second, accessibility and availability of effective tools for displaying data in a user-friendly manner; and third, availability across system boundaries in an interchangeable format.

A standard can take many forms, but essentially it comprises a set of rules and definitions that specify how to carry out a process or produce a product[13]. For the purpose of this article, we adopt the definition that standards are documented agreements containing technical guidelines to ensure that materials, products, processes, representations and services are fit for their purpose. Under this definition, there are four broad types of standards.

The first type is the measure or metric standard. This is a standard against which all comparable quantities are measured. For example, a test result may be expressed in two different units (grams/liter and milligrams/milliliter) that are mathematically identical but visually different. Slightly more complex is the case where the units are different, and not mathematically equivalent, for the same test. An example might be grams/deciliter and milligrams/milliliter. A familiar example is the loss of the $125 million Mars Climate Orbiter due to inconsistency in the units used.

The second type of standard is process-oriented or prescriptive, where descriptions of activities and processes are standardized. This type of standard provides the methodology to perform tests and processes in a consistent and repeatable way. For example, calibration, validation and standardization of different instruments in different platforms that perform the same proteomics analysis (e.g., mass spectrometry) are critical for analyzing and comparing data.

The third type of standard is performance, rather than process, based. These standards are often based on product experience. For example, analysis and comparison of diverse proteomics data are performance based.

The fourth standard type is based on interoperability among systems. In this type, process and performance are not explicitly determined, but a fixed format is specified. The goal of this type of standard is to ensure smooth operation between systems that use the same physical entity or data.

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

Box 2 The development of standards

Standards are formulated in several ways. First, they may be implemented by a single vendor that controls a sufficient portion of the market to make its product the market standard (e.g., Microsoft's Windows application). Second, the community can agree on an available standard specification (e.g., exchange format for macromolecular data exchange[8]). Third, a group of volunteers representing interested parties can work in an open process to create a standard (e.g., data exchange formats for microarray experiments, MIAME[9]). And fourth, government agencies, such as the National Institute of Standards and Technology (NIST, Gaithersburg, MD, USA) can coordinate the creation of consensus standards (e.g., physical and data standards).

The process of creating a standard by a community proceeds through several stages. It begins with an 'identification stage,' during which someone becomes aware that there exists a need for a standard in some area and that technology has reached a level that can support such a standard. If the time for a standard is ripe, then several appropriate individuals can be identified and organized to help with the 'conceptualization stage,' in which the characteristics of the standard are defined. What must the standard do? What is the scope of the standard? And what will be its format?

In the ensuing 'discussion stage,' participants begin to create an outline that defines content, to identify critical issues and to produce a time line for production of the standard. In the discussion, the pros and cons of the various concepts are discussed. Usually, a few dedicated individuals draft the initial standard; other experts then review the draft. Most standards-writing groups have adopted an open policy; anyone can join the process and can be heard. A draft standard is made available to all interested parties, inviting comments and recommendations. A standard will generally go through several versions on its path to maturity, and a critical stage is early implementation. This process is influenced by accredited standards bodies, by the federal government, by major vendors and the marketplace.

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

Comprehensive synonyms for standard vocabularies are a critical requirement for query selection of proteomics data. In controlled vocabulary parlance, redundancy is the condition in which the same information can be stated in two different ways. Synonymy is a type of redundancy that is desirable: it helps people recognize the terms they associate with a particular concept and because the synonyms map to the same concept, then the coding of the information is not redundant. A list of synonyms that are internally mapped to the same annotation entry can solve the problem of unmatched synonyms. For example, there are many ways to search for T lymphocyte (T-Lymphocyte, T cell, etc.).

To deduce possible clues about the action and interaction of proteins in the cell, one must classify them into meaningful categories that are collectively linked to existing biological knowledge. There have been many attempts to classify proteins into groups of related function, localization, industrial interest and structural similarities[10]. A proteomics strategy of increasing importance involves the localization of proteins in cells as a necessary first step toward understanding protein function in complex cellular networks. A classification of all the proteins according to their function, for example, is necessary for an overview of the functional repertoire of the protein complement of an organism of interest.

The annotation of data elements also requires that all of the related data records within a file are consistent and properly integrated across the group of files. Ideally, annotation would be largely automated and the information that accompanies high-throughput data would be seamlessly integrated into annotation forms submitted with the data[11].

To accommodate advances in the proteomics field, the community will also need to do a certain amount of crystal gazing when formulating annotation structure. What information (that is, data fields) will be necessary to ensure proteomics data are useful 3, 5 or even 10 years from now? What information will be necessary to make comparisons among measurement methods (e.g., surface-enhanced laser desorption ionization (SELDI), isotope-coded affinity tags (ICAT), multidimensional protein identification technology (MudPIT), microarrays or protein chips)? What is the balance between requesting too much information and achieving sufficiently characterized data? How much data should be captured? In what format should the data be stored? Who will take responsibility for the data?

The last question may be one of the most pressing. The sheer profusion of sites hosting proteomics data and databases containing similar types of data is hampering progress in the field. In our view, the collection of available content into standard, centralized and robustly indexed databases by national or internationally recognized entities would be a significant step forward.

Conclusions

Proteomics data are complex and difficult to process with existing tools. Data cannot be interchanged easily among different hardware, software, operating systems or application platforms. Metadata describing the content, format, interpretation and historical evolution of the proteomics data are not available to either end users or application designers. In addition, not all proteomic data are definitive; for example, identification of a single peptide does not automatically indicate the exact protein or protein isoform that it is derived from. Data collection at a volume and quality that is consistent with the use of statistical methods is a significant limitation of proteomics today. The analysis and interpretation of the enormous volumes of proteomic data remains an unsolved challenge, particularly for gel-free approaches. All this necessitates the development of semantically rich standards for proteomic data.

Data standards are essential to permit cooperative interchanges and querying between diverse and perhaps dissociated proteomics databases. By adopting a strong, clear set of consensus standards for the interpretation, handling and dissemination of specific types of protein data, the proteomics community can spur innovation by codifying accumulated technological experience and forming a baseline from which new proteomics technologies can emerge. Research, development and regulatory activities will be much more productive if provided with a wider range of critically evaluated data, virtual measurement methods and new methods for managing the dramatic increase in research data. The economic benefits of data interoperability standards are also immediate and obvious.

How should the community go about formulating these standards (see **Box 2**)? Already, some progress is being made in this direction. The Human Proteome Organization's (HUPO; Washington, DC) Protein Standards Initiative (PSI) aims to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification[12]. Currently, PSI is focusing on developing standards for two key areas of proteomics: mass spectrometry and protein-protein interaction data. In addition, as part of its 'Roadmap' initiative, the US National Institutes of Health (NIH, Bethesda, MD, USA) is also looking into ways of developing a community-based plan for the consistent analysis, representation, dissemination and publication of proteomics data (http://nihroadmap.nih.gov/buildingblocks/proteomics/).

Even so, several key questions remain. What is the scope of the standards required? Should they deal only with the exchange of experimental data? Should the scope be expanded to include other types of data exchange? Or should all these be separate efforts?

The contents of this article do not necessarily reflect the views or policies of the National Institute of Standards and Technology. Any mention of commercial products within this article is for information only and does not imply endorsement by NIST.

ACKNOWLEDGEMENTS
We thank Eugene Major (NINDS, NIH) and Bob Allen (The Johns Hopkins University) for their helpful comments.

1. Tyers, M. & Mann, M. *Nature* **422**, 193–197 (2003).
2. Verma, M. *et al. Nat. Rev. Cancer* **3**, 789–795 (2003)
3. Aebersold, R. & Mann, M. *Nature* **422**, 198–207 (2003).
4. Carr, S. *et al. Mol. Cell. Proteomics* **3**, 531–533 (2004).
5. Galperin, M.Y. *Nucleic Acids Res.* **32**, D3–D22 (2004).
6. Ravichandran, V. *et al. Electrophoresis* **25**, 297–308 (2004).
7. Berman, J.J., Edgerton, M.E. & Friedman, B.A. *BMC Med. Inform. Decis. Mak.* **3**, 1–9 (2003).
8. Westbrook, J.D. & Bourne, P.E. *Bioinformatics* **16**, 159–168 (2000).
9. Ball, C.A. *et al. Nat. Biotechnol.* **22**, 1179–1183 (2004).
10. Lyman, J.A. *et al. AMIA Annu. Symp. Proc.* **920** (2003).
11. Ouzounis, C.A. *et al. Nat. Rev. Genet.* **4**, 508–519 (2003).
12. Hermjakob, H. *et al. Nat. Biotechnol.* **22**, 177–183 (2004).
13. Chute, C.G. *AMIA Annu. Symp. Proc.* 68–73 (1998).

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology