Informatics and data management in proteomics

David Fenyö and Ronald C. Beavis

Proteomics has become dominated by large amounts of experimental data and interpreted results. This experimental data cannot be effectively used without understanding the fundamental structure of its information content and representing that information in such a way that knowledge can be extracted from it. This review explores the structure of this information with regard to three fundamental issues: the extraction of relevant information from raw data, the scale of the projects involved and the statistical significance of protein identification results.

The study of the proteins expressed by an organism (proteomics) is based on the idea that it is possible to effectively identify and track these proteins in time or space. This idea has moved from the realm of pure speculation to an achievable reality by the availability of high-speed, low-cost methods of identifying and quantifying proteins based on chromatography and mass spectrometry, as well as complementary techniques for the analytical display of mRNA populations using cDNA-array technology [1]. This review will deal with a selection of the information-handling challenges associated with protein identification experiments and the manipulation of these results into a form that allows them to be assessed in terms of biological problems.

The largest general problem associated with handling proteomics' information is that the basic experimental protocols involve multiple steps of fluid handling, electrophoresis, chromatography, mass spectrometry and computer database searching. The data obtained from each step must be linked to the data associated with the subsequent procedures, allowing for the possibility that some steps may or may not be performed. In addition, there could be sequences of steps to be performed iteratively or alternatively. The data from each step can be recorded electronically either automatically or by manual intervention. The electronic data formats for each type of instrumentation is different and it is usually in a form that can only be read by one vendor's software. Information can be recorded in a wide variety of forms, but the current goal is to store it in some form of laboratory information management system via a combination of the manually entered data and some meta-data obtained from vendor-specific software that captures important features of the original raw data. These systems are based on commercial relational database platforms and they have been successful at capturing experimental results in analytical laboratories that require strict auditing of procedures and results.

Data versus meta-data

Proteomics experimental information is derived from the output of different types of analytical instrumentation from multiple vendors. Even though there have been efforts made in the past to standardize the electronic data-representation format across analytical instrumentation platforms, these efforts have largely failed. The failure of vendors to agree on any common formats has led to a profusion of different types of electronic data binary files. The information necessary to read and write these formats is usually a proprietary secret of a particular vendor, protected by applicable local and international laws and regulations, although the vendor might supply a software utility or component that allows some type of limited access to the data.

The difficulty of dealing with this profusion of binary electronic formats has led to the design of proteomics information gathering systems that are based on meta-data formats, rather than on the original raw data. These meta-data are usually some form of 'feature table': a limited representation of the original data obtained by converting the raw electronic binary data into a discrete set of locator and intensity coordinates. These simple formats use ASCII characters, with a series of numbers separated by some combination of spaces, commas or tabs, using combinations of line feeds and carriage returns as terminators. These data formats have the advantage that they can be read by third-party software as well as common spreadsheet applications (e.g. Microsoft's Excel). Viewed from an informatics perspective, the creation of this meta-data is a form of lossy compression [2].

The major drawback to these meta-data files is that they have no agreed-upon format and they do not represent any type of standard. The original format is usually based on a vendor's text-export function, which is subsequently adapted by other vendors and users for their own use. For example, the so-called 'DTA' meta-data format is used to transfer tandem mass spectrometry data to most

David Fenyö*
Ronald C. Beavis
Proteometrics LLC,
PO Box 984, New York,
NY 10014, USA.
*e-mail:
dfenyo@proteome.ca

0167-7799/02/$ – see front matter PII: S0167-7799(02)02094-2

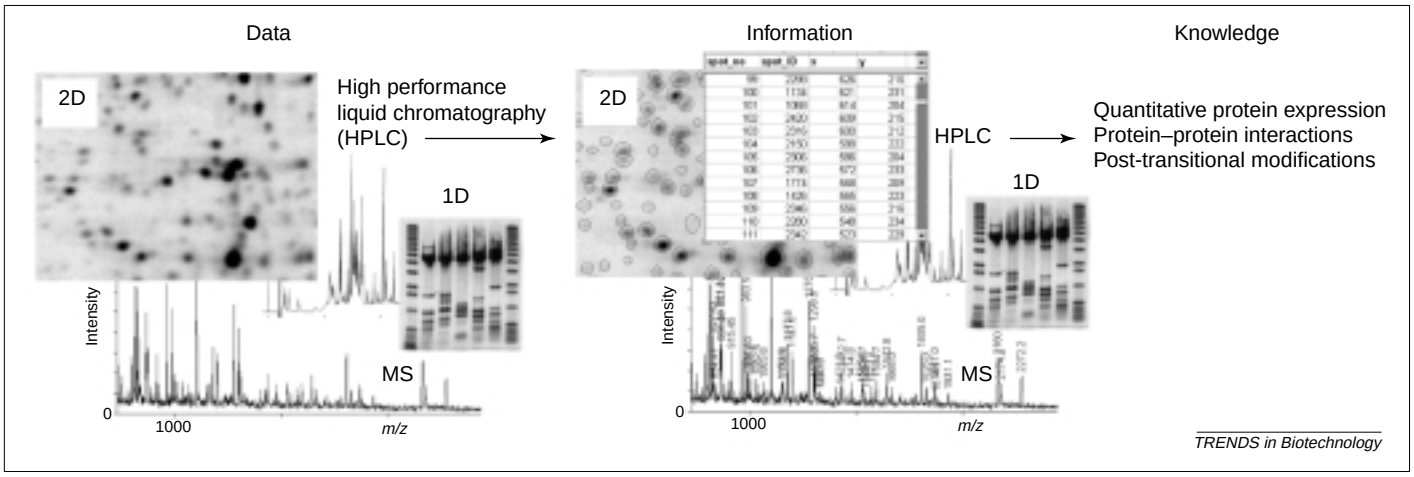

Figure 1. The informatics structure of a simple proteomics experiment involving multiple analytical techniques.

The 'data > information > knowledge' hierarchy of formal information design is useful in thinking about these experiments. 'Data' is the raw binary information extracted from instrumentation, present in large amounts but not very useful by itself. 'Information' is the processed meta-data obtained from manual or automated feature extraction (examples of such features include peak spot coordinates) and sequence databases. 'Knowledge' is what we want from the experiment: sequence identifications and quantitative results.

protein identification search engines. The original format was generated by a Thermo-Finnigan utility, but it has been altered to include at least six commonly used variations in internal formatting through the substitution of different sets of symbols to represent column separators and terminators. Surprisingly, none of the tandem mass spectrometry meta-data file-types contain any information that describes what is recorded or how the data was obtained. They also cannot store information that is vital to their interpretation, such as the mass of the fragment ions, their charge or any indication as to how the m/z-intensity pairs stored in the file were calculated.

As a common example of how these feature tables are used in practice, take the simple example of an experiment in which 2D-gel electrophoresis is used to separate and quantify proteins and a liquid-chromatograph coupled to an ion trap tandem mass spectrometer is used to identify the proteins found in gel plugs. The stained gel is analyzed using 'spot-picking' software, which determines the x–y coordinates of the center of the spot, its area and the intensity of the staining. These values are placed into a feature table file. The file is then loaded (either automatically or manually) into a gel plug removal robot, which removes the appropriate spot and loads the gel plug into a 96-well plate, where it is digested and the peptides extracted. The results of the robot's fluid handling processes are placed in a feature table, recording the success or failure of the various operations. This feature table is then loaded (either automatically or manually) into the mass spectrometer's software, which drives a high-performance liquid chromatograph auto-sampler. The resulting tandem mass spectra are each stored in individual feature tables that record the parent ion mass and charge, as well as the m/z and intensity of the fragment ions (up to 1000 individual feature table files per gel spot). A limited amount of chromatographic information is encoded into these files' path names. Each one of these files is submitted individually to a search engine that attempts to match the file data to a peptide sequence and a further set of feature table files is returned that record a list of potential matches and a set of arbitrary scores that can be used to assess the quality of the resulting correlation between a spectrum and a peptide sequence. A realistic estimate of the number of individual feature tables generated in such an experiment (per 96-well plate) is from 4000–60 000, depending on the details of the table formats. It is worth noting that most of these files represent contaminants and noise, but they are generated and recorded in an effort to obtain the highest sensitivity possible. The information content of hundreds of megabytes of storage can frequently be compressed into the name of a single, commonly identified protein, such as 'cytokeratin' [3].

The creation, logging, management and translation of these arbitrarily formatted, automatically generated feature-table files has become the central role of the first generation of proteomics information-handling systems. The much larger volume of vendor-specific binary data might also be managed in some way. The only role of the original data file(s) is often as a record of an instrument's settings, but it is retained in its entirety in the unlikely event that it will be required in the future. It should be stressed that these binary data are not used in any way for the final interpretation of the results: the meta-data in the feature tables are used exclusively for all analysis. Naïve users often regard this meta-data as a set of trivial temporary files, but they are truly the only record of what information was actually used in the next step of the process. It might also be impossible to recreate the meta-data at a later date from the original data, because version changes

to the vendor-specific software that reads the binary data could unpredictably affect their output when reading obsolete binary file formats. Storing this information for later retrieval and reuse is an important part of any well-structured proteomics informatics system.

Small, medium or grande?

Proteomics experiments can be characterized by the amount of experimental data generated during the experiments and the resulting informatics models [4] (see Fig. 1). Small-scale experiments produce relatively small amounts of data, such as identifying a subset of the proteins in a one- or two-dimensional gel [5]. The experiments themselves are designed around some very specific sets of hypotheses, which are tested in detail by the experimental results. Analysis involves a considerable amount of intuitive iteration and re-testing by trained observers and thus the process cannot be easily or profitably automated. Therefore, the importance of the information's structure is reduced: the observers will each have their own methods of extracting information and knowledge from the dataset.

Medium-scale proteomics uses many of the same methods as small-scale experiments, except that the dataset becomes sufficiently large that it is practically impossible for trained observers to evaluate it all [6]. The goal of these experiments is still to extract the maximum amount of knowledge from the dataset, but the volume of information required becomes a real impediment to the iterative, intuitive approach used in small-scale experiments. Observers are used to screen the information to determine whether it has been properly constructed from the data. The informatics model used for this type of project is necessarily relatively rigid, reducing the reporting flexibility inherent in small-scale experiments. Data and information storage become form-filling exercises, with the goal being to populate a relational database that can be used to examine the information on a broad scale.

Large-scale proteomics projects use fully automated systems to acquire and analyze the data for it to become information [7]. Observer intervention is not required, other than as a quality control measure. These projects are constructed around abstract hypotheses and the extraction of knowledge is difficult: a huge amount of data and a correspondingly large amount of information is generated in the belief that it will give rise to emergent patterns which will result in unanticipated benefits even if the initial hypotheses were vaguely formulated. This approach is central to the 'systems biology' approach [1]. The informatics model used must be completely rigid, following a set of external rules imposed by the experimental design [8].

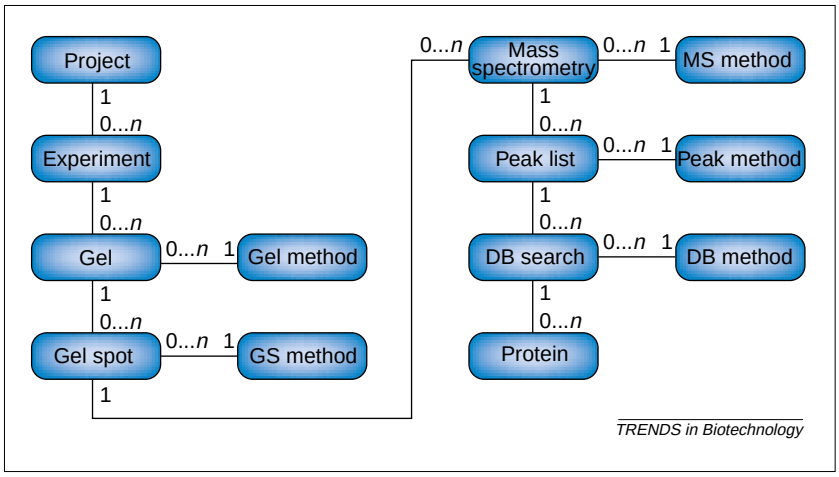

Figure 2. A generic database model for a proteomics project, displayed in standard database notation.

This type of notation permits the compact representation of very complex webs of interrelations between database relations (tables). Each of the boxes represents a relation that contains information about a particular object. Each relation is linked to other relations: the numbers adjacent to the lines indicate how many links can be made to a particular relation. For example, a single (1) 'project' relation can be linked to 0 to *n* 'experiment' relations. Each single (1) 'experiment' relation can be linked to 0 to *n* 'gel' relations, and each single (1) 'gel relation' can be linked to 0 to *n* 'gel spot' relations. In addition, a 'gel' relation is linked to a 'gel Method' relation, which can be linked to 0 to *n* 'gel' relations (and so on). Establishing an appropriate set of relations and linkages is the most important part of any informatics design that requires the use of a relational database.

These three informatics models are in different states of development. The small-scale models are well-tested and easy to understand. They involve the type of *ad hoc* information structure in common use in research laboratories and depend on the in-depth review of the data at all scales to come to conclusions. Both the medium- and large-scale models are currently being tested and refined. Their value will depend on the standardization of reporting and database formats, so that independent evaluation of the results and conclusions of these studies can be performed.

There is currently no standard set of database models for proteomics experiments. As an example, Fig. 2 shows a practical, generic database model for a protein identification project. The systems in current use are either proprietary or designed for special purpose applications. The relational databases have been shaped by the format of the meta-data presented to the designers. Therefore, the models take on the shape of the experimental data, rather than focusing on the information that is truly desired by the end-user of the database. Dozens of projects have been started around inappropriate experimental metaphors that reflect long-held desires of particular analysts, such as presenting all of the experimental results through an interface that looks like a stained 2D-electrophoresis gel. This type of metaphor-based interface design attempts to encapsulate all of the normal functions of a laboratory information management system database into a very

limited graphical representation, making the final results unnecessarily difficult to evaluate. The only real exception to this trend is the Biomolecular Interaction Network Database (BIND) project. It uses a publicly available database model based on the US National Center for Bioinformatics database standard, ASN.1 [8]. The BIND data model has been used successfully to store and render the information from at least one large-scale protein identification project [7].

What does it mean to identify a protein?

The question of evaluating the validity of protein identification results is still a matter for active research and it has not been solved satisfactorily [9,10]. Several statistical approaches have been proposed [11–13]. These studies have looked at the problem from different viewpoints, but they have not provided conclusions that can be directly applied to the common problem of estimating the confidence of experimental protein identifications in a simple and interpretable manner. The current software tools and scoring algorithms leave interpreting the results up to the user. Many practitioners have developed their own, pseudo-statistical tests for validating results based on the comparison of results from different scoring algorithms. This type of approach should be used with great caution without a detailed knowledge of the statistical distributions produced by the various scoring algorithms. A consensus similar to that used for evaluating sequence similarity scores must be reached for protein identification to become a truly mature technique [14,15]. Currently, the authors favor adopting the combination of relative probabilities and expectation values for reporting sequence identifications [16–20], because of their simple and intuitive interpretation. These standard statistical measures require an understanding of the underlying statistical distributions and an empirical method of estimating them, but they have the advantage of being equally applicable to identifications based on either single mass spectrometer or tandem mass spectrometer measurements. These measures also have the advantage of being relatively independent of the details of an underlying scoring system: they are properties of the overall distribution of scores, rather than the absolute values of the scores themselves.

Acknowledgments

The authors would like to thank Manitoba Centre for Proteomics, the Canadian Institutes for Health Research and the Boards of Canada for their assistance, support and inspiration.

References

1 Griffin, T.J. *et al.* (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 1, 323–333
2 Sayood, K. (1996) *Introduction to Data Compression*, pp. 1–12, Morgan Kaufmann
3 Pierce, J.R. (1980) *An Introduction to Information Theory: Symbols, Signals and Noise, 2nd Edition*, pp. 107–144, Dover Publications
4 Shedrof, N. (1999) Information interaction design: a unified field theory of design. In *Information Design* (Jacobson, R., ed.), pp. 267–292, MIT Press
5 Andersen, J.S. *et al.* (2002) Directed proteomic analysis of the human nucleolus. *Curr. Biol.* 12, 1–11
6 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
7 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
8 Bader, G.D. and Hogue, C.W. (2000) BIND – a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 16, 465–477
9 Rappsilber, J. and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* 27, 74–78
10 Mann, M. *et al.* (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473
11 Perkins, D.N *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20, 3551–3567
12 Pevzner, P.A. *et al.* (2001) Efficiency of database search for identification of mutated and modified proteins via mass. *Genome Res.* 11, 290–299
13 Parker, K.C. (2002) Scoring methods in MALDI peptide mass fingerprinting: ChemScore and the ChemApplex program. *J. Am. Soc. Mass Spectrom.* 13, 22–39
14 Karlin, S. and Altschul, S. (1990) Methods for assessing the statistical significance of molecular sequence features using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* 87, 2264–2268
15 Durbin, R. *et al.* (1998) *Biological Sequence Analysis*, pp. 36–45, Cambridge University Press
16 Eriksson, J. *et al.* (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* 72, 999–1005
17 Field, H.I. *et al.* (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification and archives data in a relational database. *Proteomics* 2, 36–47
18 Fenyö, D. and Beavis, R.C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* (in press)
19 Keller, A. *et al.* (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* 6, 208–212
20 Keller, A. *et al.* Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* (in press)