BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 2. 174–188
Advance Access publication February 15, 2008

doi:10.1093/bib/bbn004

Information quality in proteomics

David A. Stead, Norman W. Paton, Paolo Missier, Suzanne M. Embury, Cornelia Hedeler, Binling Jin, Alistair J. P. Brown and Alun Preece

Submitted: 12th July 2007; Received (in revised form): 11th January 2008

Abstract

Proteomics, the study of the protein complement of a biological system, is generating increasing quantities of data from rapidly developing technologies employed in a variety of different experimental workflows. Experimental processes, e.g. for comparative 2D gel studies or LC-MS/MS analyses of complex protein mixtures, involve a number of steps: from experimental design, through wet and dry lab operations, to publication of data in repositories and finally to data annotation and maintenance. The presence of inaccuracies throughout the processing pipeline, however, results in data that can be untrustworthy, thus offsetting the benefits of high-throughput technology. While researchers and practitioners are generally aware of some of the information quality issues associated with public proteomics data, there are few accepted criteria and guidelines for dealing with them. In this article, we highlight factors that impact on the quality of experimental data and review current approaches to information quality management in proteomics. Data quality issues are considered throughout the lifecycle of a proteomics experiment, from experiment design and technique selection, through data analysis, to archiving and sharing.

Keywords: *information quality; proteomics; standards; quality assessment; information management*

INTRODUCTION

Proteomics can be defined as the study of the protein complement of a biological system, for example an organism, cell, or tissue. The gathering of information about a proteome, indeed about any biological component or system, requires both experimental observation ('wet lab' procedures) and data analysis ('dry lab' procedures). The quality of the final output depends on several considerations including experimental design, control of biological and analytical variability, the recording of descriptive information about the experiment ('metadata') along with the results themselves, and the appropriate use of bioinformatics tools and statistical significance tests for data analysis [1–3].

Proteomics is generating increasing quantities of data from rapidly developing technologies employed in a variety of different experimental workflows. Large-scale proteomics experiments have relied mainly on the technologies of two-dimensional gel electrophoresis (2DE) [4] and liquid chromatography-tandem mass spectrometry (LC-MS/MS) [5].

Corresponding author. David A. Stead, School of Medical Sciences, University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen, AB25 2ZD, UK. Tel: +44 (0) 1224 555804; Fax: +44 (0) 1224 555844; E-mail: d.stead@abdn.ac.uk

David Stead is a Senior Research Fellow in the School of Medical Sciences, University of Aberdeen. Based in the Aberdeen Proteomics facility, he specialises in protein/peptide mass spectrometry.

Norman W. Paton is a Professor of Computer Science at the University of Manchester. He works on data integration and query processing, and on their applications in the life sciences.

Paolo Missier is a Research Fellow at the School of Computer Science, University of Manchester, where he has been a member of the Qurator project.

Suzanne M. Embury is a Lecturer in the School of Computer Science at the University of Manchester. Her main areas of interest are in information quality, information integration and software evolution.

Cornelia Hedeler is a Research Associate in the School of Computer Science at the University of Manchester. Her areas of interest are in bioinformatics, genome data management and data integration.

Binling Jin obtained her PhD from Manchester University in 2004, and for the past 3 years has worked as a Research Fellow at Aberdeen, looking at information quality in e-Science.

Al J. P. Brown is a Professor of Microbiology at the University of Aberdeen and Director of the Aberdeen Proteomics facility. His main research interests are in the pathogenomics of *Candida albicans*.

Alun Preece obtained his PhD from University of Wales in 1989. He is currently a Reader in Computing Science at Aberdeen. His research includes application of ontologies in e-Science.

The high-throughput nature of some of these experiments presents the scientist with significant data handling challenges, not least of these being the problem of false-positive results [6]. If proteomics data is to be stored in public repositories for re-use by other scientists or combined with other data sets to inform systems biology studies [7], then the quality of that data has wider importance still. Ways must be found to measure, annotate and make accessible the quality of proteomics data sets so that researchers can decide whether they are suitable for their particular application.

A framework for discussion of data quality issues in proteomics

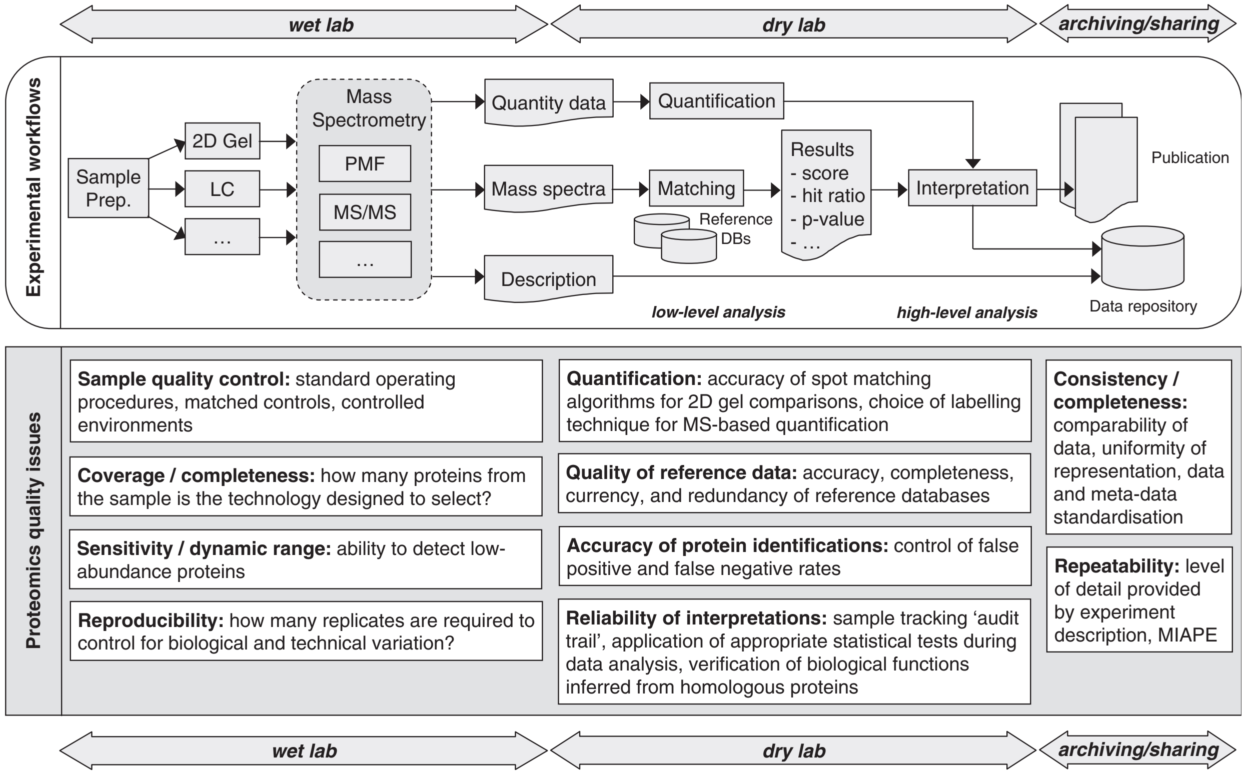
A typical proteomics experiment involves a wet lab and a dry lab portion, as well as a final phase during which the results are published and shared with the scientific community.

We have identified a number of quality issues that pertain to each of these phases, as illustrated in Figure 1. Some of these issues relate to the determination of which proteins are relevant to the experimental hypothesis (qualitative proteomics) or concern the measurement of how much of these proteins are present in the sample (quantitative proteomics). These two are inter-linked, however, because quantification of individual proteins in the sample is a necessary prelude to the selection of the subset of proteins whose expression levels are considered to be regulated as a result of the experimental challenge. Other issues may only become apparent once the experimental data have been published.

In this survey we present an analysis of how technology and other factors, such as the adoption of standardised descriptions of the experiments, affect the quality of the outcome in the different phases of the experiment, as well as the ability of the scientific community to exploit those results. The separation into technologies and phases within the workflow facilitates this analysis and is reflected in the organisation of the article.

In the following section, Quality issues in experimental data generation, we discuss problems affecting the quality of the data produced in the lab, based on the following considerations:

- A number of factors may affect the *quality of the sample*, including the environment in which the experiment takes place and the adoption of standard operating procedures.



Sample quality control: standard operating procedures, matched controls, controlled environments

Coverage / completeness: how many proteins from the sample is the technology designed to select?

Sensitivity / dynamic range: ability to detect low-abundance proteins

Reproducibility: how many replicates are required to control for biological and technical variation?

Quantification: accuracy of spot matching algorithms for 2D gel comparisons, choice of labelling technique for MS-based quantification

Quality of reference data: accuracy, completeness, currency, and redundancy of reference databases

Accuracy of protein identifications: control of false positive and false negative rates

Reliability of interpretations: sample tracking 'audit trail', application of appropriate statistical tests during data analysis, verification of biological functions inferred from homologous proteins

Consistency / completeness: comparability of data, uniformity of representation, data and meta-data standardisation

Repeatability: level of detail provided by experiment description, MIAPE

Figure 1: A framework illustrating where information quality issues (lower panel) can arise during typical proteomics workflows (upper panel).

- The choice of experimental technology affects the number of proteins that can be successfully identified, i.e. the *completeness* of the result, which includes the ability to detect low-abundance proteins (*sensitivity*).
- *Reproducibility* takes account of the biological and technical variations inherent in the experiment by including a number of replicates, and using matched controls to minimise the number of simultaneous biological variables.

Also, two main factors that affect errors in protein identification, namely:

- The *accuracy* of the matches obtained using algorithms (inaccurate matches may result in false positives).
- The *quality of the reference protein databases* used for the match: an *incomplete* or *inaccurate* database may result in false negatives, i.e. by failing to identify proteins that are present in the sample.

Then, in the section on Archiving and sharing proteomics data, we focus on the issues that affect the potential exploitation of the result by the community, namely:

- *Uniformity of representation*, i.e. the standardisation of the data formats, as well as of the metadata that describes the experiment.
- *Accuracy* and *level of detail* of the experiment description, which is necessary to allow the entire experiment to be repeated.

Finally, we will conclude by noting how software environments for recording and annotating proteomics experiments may alleviate some of the problems in the last two phases of the workflow by providing views of the data that make some of its quality characteristics explicit to the user.

QUALITY ISSUES IN EXPERIMENTAL DATA GENERATION

The quality of the final output from proteomics studies intrinsically depends on the wet lab phase of the workflow. We begin a discussion of the importance of controlling the biological and technical variables within proteomics experiments by considering the quality of the biological sample.

Sample quality control

As with all lab-based studies, the importance of minimising variation by strict adherence to robust experimental protocols, use of single batches of reagents and use of matched controls should not be underestimated in proteomics. Moreover, the authors of a recent commentary state that 'many proteomics efforts suffer from a lack of rigor' and focus on sample preparation as an area that requires more critical attention [8].

The experimental design should closely reflect the scientific hypothesis being tested and it may be important to control factors such as the genetic background of microbial strains, the age and gender profiles of patient populations, the cell type composition of tissue extracts and the environmental variables that might affect sample quality (temperature, constituents of growth media, timing of collection, etc.). Checking sample quality at an early stage and replacing poor-quality samples where possible is preferable to having to deal with the resulting quality issues during subsequent data analysis steps. For example, 1D gels can be used to test the quality of samples destined for 2DE by revealing the effects of proteolytic degradation.

Simply recording, in a prescribed way, the details of how the various experimental steps were carried out may promote the use of standard operating procedures, and is an argument for the development of proteomics standards (see Standardisation of experiment description section). Standard operating procedures tend to be developed by individual laboratories for a particular type of sample [9–11] and may not be particularly successful when transferred to other applications.

Wet lab quality issues for 2DE

The design of 2DE-based experiments is an important factor determining the quality of the output from such studies. The aims should be to minimise biological and analytical variation, and to avoid introducing bias between sample groups. It is desirable to develop and use standard operating procedures for protein extraction/solubilisation, electrophoresis and gel image analysis, and to determine the number of biological and technical replicates necessary to detect a given difference in protein expression level between sample groups [12, 13].

Such a 'power analysis' is rarely, if ever, carried out in proteomics because of the cost involved in setting up a pilot study. 'How many replicate gels should be run?' is an often-asked question. Unfortunately, there is no simple answer—it depends on the variability of the biological material, reproducibility of the technique and the degree of difference between sample groups that is to be detected (other things being equal, fewer replicates would be required to confidently detect a 3-fold change in spot volume than a 1.5-fold change). For samples generated under well-controlled conditions, it may be reasonable to run four (biological) replicate gels per sample group and to include spots that are detected and matched in at least three gels per set when using software that allows for missing values. For samples known to be variable, such as tissue, it may be advantageous to run more replicates or pool material within sample groups prior to 2DE.

A major advance in the reproducibility of 2DE has been provided by immobilised pH gradient gels (IPG strips) for the first dimension separation coupled with pre-cast second dimension gels [14], as used in the IPGphor and Ettan DALT systems (GE Healthcare) [15]. 2DE remains one of the most important technologies for separating complex protein mixtures and continues to be developed.

The problem of under-representation of certain classes of proteins in 2DE is well-known, for example for membrane proteins (very hydrophobic), DNA-binding proteins (very alkaline) and signalling proteins (low abundance), and standard 2DE methods may not be suitable for proteins having such characteristics [4].

Quantification in 2DE depends on visualising the resolved protein spots with a suitable stain, scanning the gel to produce a bitmapped image and using software to select spots showing reproducible changes in intensity [16]. Choice of staining technique can markedly affect the sensitivity of 2DE, with silver-stains and fluorescent dyes better able to detect low-abundance proteins than the commonly used Coomassie blue. However, silver-staining is less suitable for quantitative studies because the reaction is non-stoichiometric and the end-point is subjective [4].

Success of the subsequent image analysis is strongly dependent on the similarity of the replicate 2D gels and also on the performance of the software algorithms used for spot matching or gel image alignment. Difference gel electrophoresis (DIGE) technology, in which two samples are labelled with different CyDyes and mixed prior to the electrophoretic separation, reduces the number of gels required and overcomes some of the problems associated with gel-to-gel variability [17].

Dry lab quality issues for 2DE

The accuracy of comparative studies using 2D gels depends on the success of matching spots within and between groups of replicate gel images from different experimental conditions. Recent developments in 2DE analysis software such as Progenesis SameSpots (Nonlinear Dynamics) or Melanie 6.0 [Geneva Bioinformatics (GeneBio) SA] focus on reducing the amount of user editing (subjective input) and the application of multivariate statistical methods for analysing the spot data [18, 19]. For example, a principal components analysis (PCA) plot can provide a quality check showing that replicate gels from a particular experimental condition group together, and are separated from those generated under different conditions. By way of illustration, Figure 2 shows how the replicate 2DE gel images from an experiment involving the growth of *Candida albicans* in media containing different carbon sources grouped together in a PCA plot, confirming that the experimental challenge (changing the carbon source) had a greater impact on the proteome than biological and analytical variance. Some programs, e.g. Progenesis SameSpots, depend on warping the images so that the corresponding spots can be superimposed—volume data is collected from every replicate whether a protein spot is visible or not and therefore missing values do not compromise the use of statistical tests.

Wet lab quality issues for peptide mass fingerprinting (MALDI-TOF MS)

Experimental design is equally important in MS-based proteomics and has been discussed elsewhere [1, 20]. Some technologies are less suitable for detecting low-abundance proteins, as we have already noted for 2D gel staining. Peptide mass fingerprinting (MALDI-TOF MS) has typically been employed for the identification of relatively abundant proteins in 2D gel spots. It does not have the power to identify proteins in more complex mixtures because suppression effects in the ionisation source limit the number of peptides that can be simultaneously analysed by MALDI-TOF MS, and success depends on obtaining good sequence coverage from a reasonable number of peptides (at least 4–6, preferably more) matched to each protein.

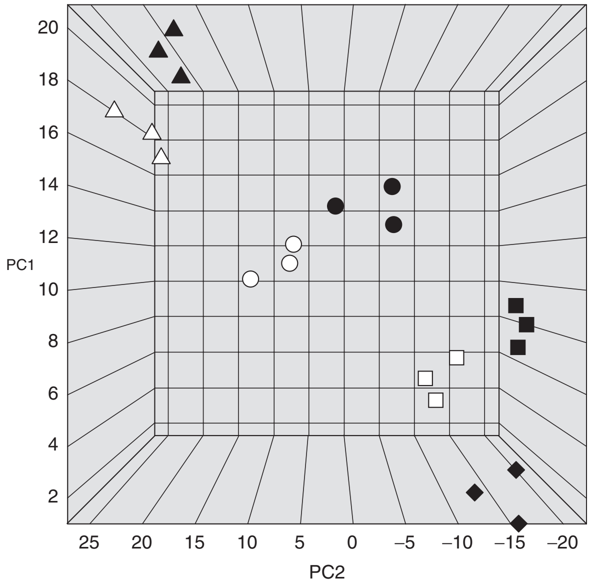


Figure 2: PCA plot for replicate 2D gels from different samples groups. *Candida albicans* cells were grown on various carbon sources (closed diamonds – glucose; open squares – casamino acids; closed squares – casamino acids plus glucose; open circles – oleic acid; closed circles – oleic acid plus glucose; open triangles – lactate; closed triangles – lactate plus glucose), soluble proteins were extracted and separated by 2DE. The gel images from three biological replicates per condition were subjected to a principal components analysis using SIMCA-P (Umetrics).

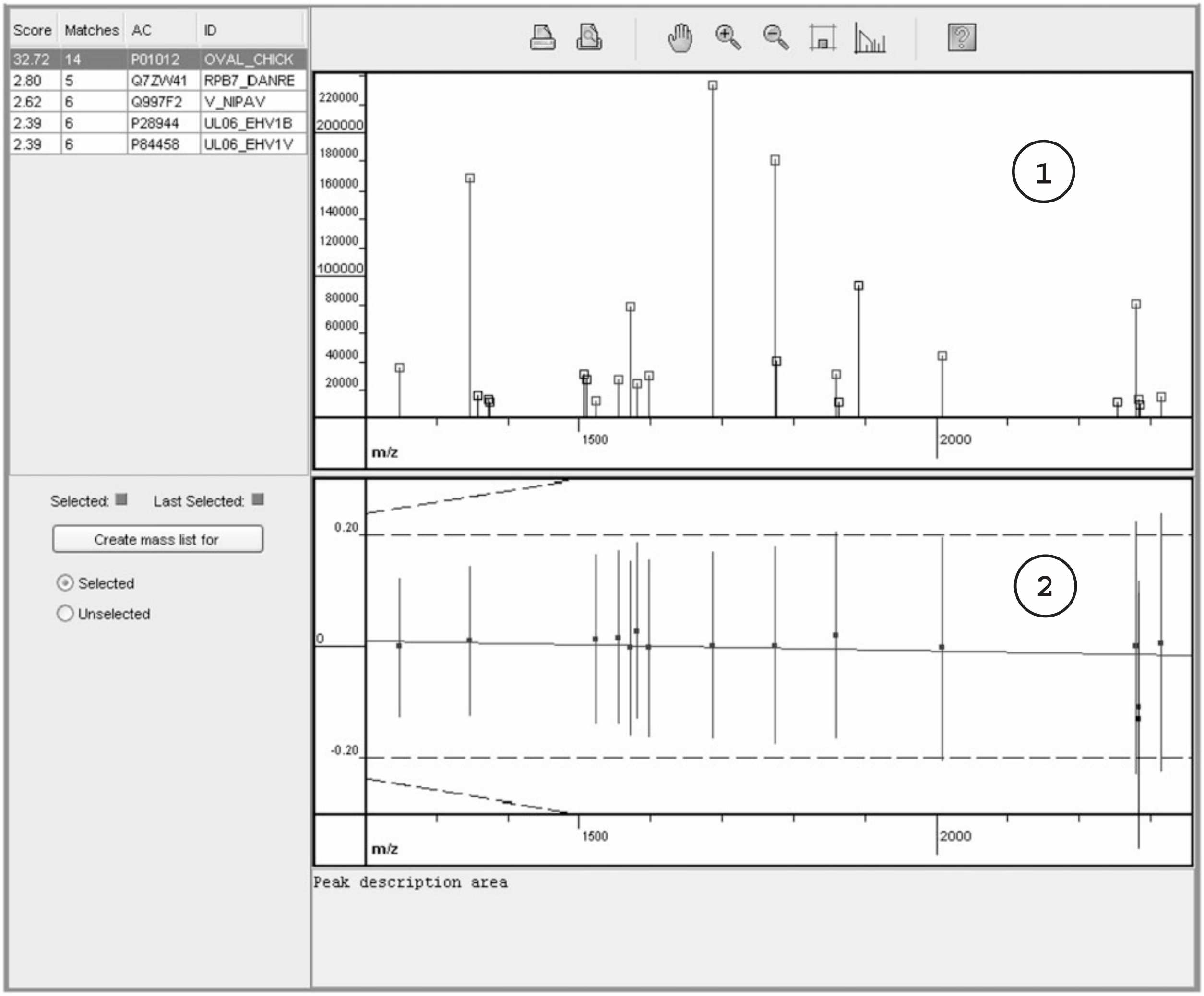
These suppression effects also make MALDI-TOF MS susceptible to loss of sample peptide signals caused by the presence of other abundant ions, whether these come from extraneous sources (e.g. keratin and polyethylene glycol) or from the sample itself (e.g. albumin) if procedures to remove them are inadequate. In practice it proves difficult to confidently identify more than two or three proteins in the same sample by peptide mass fingerprinting. This was clearly illustrated by a comparative analysis of 98 2DE gel spots from *Methanococcus jannaschii* [21]. By MALDI-TOF MS, 88% of spots contained a single protein, 11% contained two and 1% a mixture of three different proteins. In contrast, the same samples analysed by LC-MS/MS revealed that 41% were single-protein spots, with the majority containing multiple proteins—as many as six per spot.

Dry lab analysis of MALDI-TOF MS data

Despite the limitations mentioned in the previous section, peptide mass fingerprinting remains a popular protein identification technique and there are a number of bioinformatics tools available for the analysis of MALDI-TOF MS data [3].

The accuracy of protein identifications obtained by peptide mass fingerprinting depends on the successful matching of the experimental peptide masses by a search algorithm to theoretical masses derived from a protein sequence database. Various outputs from the search result may help to indicate whether a particular match is correct or not. Different search engines calculate their search scores in different ways, and interpreting this information can be difficult. Three simple metrics; hit ratio, mass coverage and excess of limit-digested peptides, have been proposed as universal measures of the quality of a protein identification by peptide mass fingerprinting that can be combined into a single score and used to validate such protein identifications, particularly in large data sets [22].

The *Molecular & Cellular Proteomics* journal guidelines suggest that, for peptide mass fingerprinting, the number of masses matched to the identified protein, the number of masses not matched in the spectrum and the sequence coverage should be reported along with the input parameters used in the database search [6]. These guidelines advise the use of probability-based scoring schemes or the reporting of the expected false-positive rate. Of the various peptide mass fingerprinting tools available, Aldente [23, 24], Mascot [25] and ProFound [26] provide sufficient information to fulfil these requirements. Aldente gives statistics on random sequences and colour-codes results depending on whether the score is greater or lower than the best random score. It also has a viewer (requires web browser Java plugin) that displays peaks matched in the spectrum and their peptide mass error distribution (Figure 3)—information that is useful for validating the result. However, the choice of database is limited to Swiss-Prot or TrEMBL. Mascot has a ‘decoy’ database search option that returns the score of the best random hit, reports a probability-based Mowse score and expectation value, and displays error distribution plots but not the peptide mass spectrum. The public web server allows searches to be run over the MSDB, NCBInr and Swiss-Prot databases, while the purchase of Mascot Server software allows the user to install any database if a suitable FASTA sequence (DNA or amino acid) file is available. ProFound can report either an expectation value or a probability value and *Z*-score, and displays the spectrum coverage and mass error distribution. An evaluation



Score	Matches	AC	ID
32.72	14	P01012	OVAL_CHICK
2.80	5	Q7ZW41	RPB7_DANRE
2.62	6	Q997F2	V_NIPAV
2.39	6	P28944	UL06_EHV1B
2.39	6	P84458	UL06_EHV1V

Figure 3: Output from the BioGraph viewer of the Aldente peptide mass fingerprinting tool. Ovalbumin standard was run on a ID gel, the band excised, and subjected to in-gel tryptic digestion. Peptide masses determined by MALDI-TOF MS were input as a peak table (PKT) file to Aldente. Panel (I) shows the spectrum coverage view and panel (2) the mass error distribution plot, for the best match in the Swiss-Prot database.

of search engines commonly used for peptide mass fingerprinting concluded that ProFound provided the best discrimination between random matches and correct identifications [27].

Dry lab analysis of tandem mass spectrometry (LC-MS/MS) data

Tandem mass spectrometry (LC-MS/MS) data contains sequence information that can be used to identify peptides (and from them, proteins) by *de novo* sequencing [28], peptide fragment fingerprinting [29], or peptide sequence tagging [30]. Of these, only peptide fragment fingerprinting is used on a large scale. Peptide fragment fingerprinting is an analogous protein identification technique to peptide mass fingerprinting in which peptide fragment, or MS/MS, spectra are matched to theoretical masses of peptide fragments generated *in silico* from a sequence database [29]. Knowledge of which bonds in the peptides break preferentially in the mass spectrometer is important in peptide fragment fingerprinting, in the same way that the specificity of the cleavage reagent (e.g. trypsin) is used to predict the peptides generated in a peptide mass fingerprinting experiment. Peptide fragmentation is dependent upon the type of tandem mass spectrometer and dissociation technique—these parameters should be specified when conducting the database search.

Of the various search engines available and reviewed in [3], Mascot [25] and SEQUEST [31] are probably the most widely used. Evaluations of different search engines, including Mascot and SEQUEST, have reached variable conclusions [27, 32, 33], suggesting that their relative performance may depend on the nature of the data set used for testing. Mascot MS/MS ions searches may be performed using the public web server (which may be quite slow and limits the size of the data file to a maximum of 300 masses) or on a local Mascot server. In order to achieve a significant gain in speed from a local server, a powerful cluster of processors is required. SEQUEST is exclusively marketed by Thermo Scientific as part of the BioWorks software designed for their instruments. Phenyx [34] provides a public web interface for low-throughput submissions [35], while the OMSSA web server [36] is designed for higher-throughput use, but does not consolidate the individual peptide identifications into protein 'hits'. X! Tandem [37] cannot be run via the Internet but must be downloaded and installed on a local web server—a process that may require specialist knowledge. In practice, the choice of search engine may depend on the type of instrument because of the limited compatibility of certain MS file formats. Only Mascot currently accepts proprietary data files from a wide range of different instruments. However, this situation will change as the generic XML-based file format specified by the mzData standard [38] becomes more widely implemented—a tangible benefit of the Proteome Standards Initiative (PSI) (see Standardisation of experiment description section).

Because the output files produced from LC-MS/MS experiments are large and complex, there is a danger that protein identification software systems are used as 'black boxes' by practitioners with little or no validation of the results. This reinforces the concerns over false-positive protein identifications mentioned previously (see Introduction section). Considerable attention has been focussed on the fact that many of the spectra generated by LC-MS/MS proteomics experiments cannot be confidently assigned to known peptide sequences. It has been argued that many of the MS/MS spectra are of low quality and should be filtered out of the analysis [39–42]. Another approach is to improve discrimination between correct and random matches, by using machine learning techniques to classify database search results [43], average peptide scores based on Mascot ion scores [44], or *S*-scores based on sequence tag information [45]. Mass deviance has been proposed as a suitable metric for assessing the quality of a peptide assignment from MS/MS data [46], and the excess of limit-digested peptides quality metric proposed for peptide mass fingerprinting [22] is also predicted to be of value for assessing protein identifications by peptide fragment fingerprinting. A third approach, taken by commercial software systems such as ProteinScape (Bruker Daltonics), PEAKS (Bioinformatics Solutions Inc.) and Spectrum Mill (Agilent Technologies) (see Laboratory information management systems section), is to send the input masses to more than one search engine and to cross-validate or consolidate the results in an attempt to increase confidence in the protein identification. Identifications of peptides from MS/MS data may be validated by comparing the results with experimental peptide fragment spectra from other labs, for which the probability of correct identification has been uniformly tested. This is the idea behind data repositories such as the Global Proteome Machine (GPM) [47] and PeptideAtlas [48, 49] (see Proteomics data repositories section).

'Decoy' databases, in which the protein sequences are reversed or randomised, are particularly useful for estimating false-positive identification rates in peptide fragment fingerprint searches [50–52]. False-negative identifications, caused by analytical incompleteness, may be a serious problem in qualitative differential display LC-MS/MS experiments [53]. Quantitative mass spectrometry-based approaches overcome this problem.

Quantification can be achieved in LC-MS/MS by using stable isotope mass-tagging techniques (reviewed in [54–56]). Introducing mass tags and pooling the different samples early in the workflow, as in the SILAC technique (stable isotope labelling by amino acids in cell culture) [57], has the advantage of removing the effect of analytical variance in subsequent processing steps (Figure 4).

An interesting recent advance is the virtual 2D mapping of LC-MS data [60, 61]. This technique can be used for comparative proteomics based on the display of two separate LC-MS/MS runs. In this case, the problem of spot matching in 2DE is replaced by one of peak retention time matching. While 2D gels display protein spots separated by charge and molecular size, a virtual 2D map displays LC-MS data as ion intensities distributed by mass-to-charge ratio (*m/z*) and retention time. LC-MS image analysis may also have potential applications

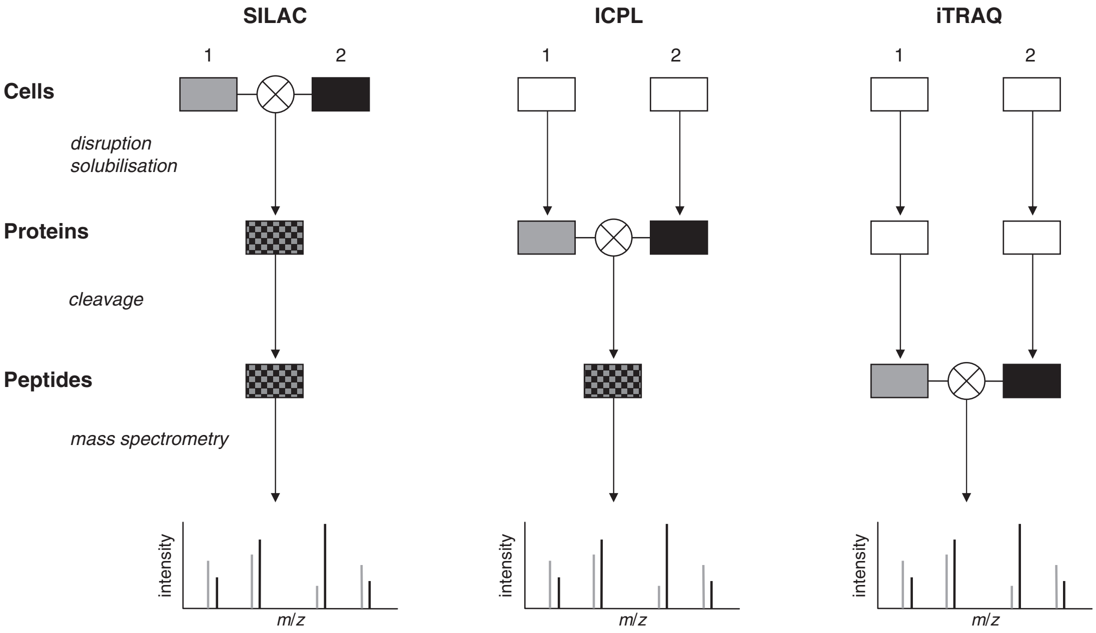


Figure 4: Workflows for quantitative MS-based differential proteomics. Generalised workflows for the comparison of two biological samples (1, 2) using stable isotope labelling by amino acids in cell culture (SILAC) [57], isotope coded protein labelling (ICPL) [58] or isobaric tag for relative and absolute quantification (iTRAQ) [59]. Labelling certain amino acids with light (grey) or heavy (black) mass tags then mixing (crossed circle) the samples is performed before harvesting in the case of SILAC, after the extraction of proteins in the case of ICPL, and after digestion of the proteins to peptides in the case of iTRAQ. For clarity, protein or peptide separation steps are not shown, but these are necessary to reduce the complexity of the sample prior to mass spectrometry.

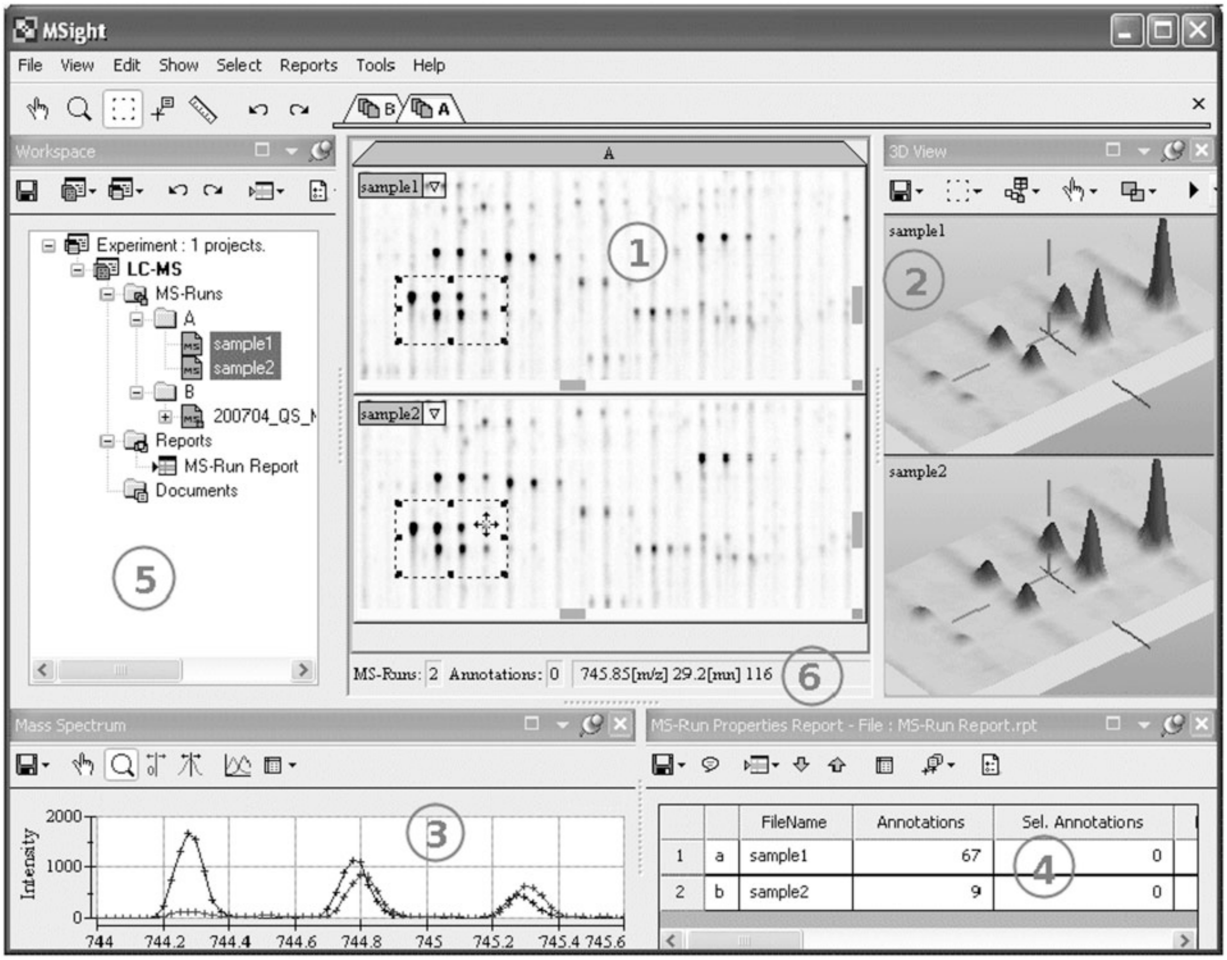
in quality assessment (artifacts created during sample processing may be recognised as characteristic patterns on the virtual map) and post-translational modification (PTM) discovery, and may prove to be a powerful tool that helps the researcher to cope with the size and complexity of the data produced by LC-MS/MS experiments. The MSight mass spectrometry imaging software (Figure 5) is freely available [62] and supports MS and MS/MS data in a variety of proprietary and generic data formats.

ARCHIVING AND SHARING PROTEOMICS DATA

This section describes the pragmatics of experimental data management and sharing, with a view to identifying the relationship between data quality and information management systems, data standards and public repositories.

Laboratory information management systems

Many software tools for proteomics have been developed to accomplish specific tasks, but there is a need to support the whole data-gathering and reporting process. Laboratory information management systems (LIMS) have been used for many years for sample tracking and reporting purposes, for example in the clinical biochemistry laboratory where maintaining an accurate sample audit trail is essential. Similar systems are being developed for proteomics laboratories that aim to combine sample tracking and automated data analysis (including protein identification and validation) with the functionality to generate output files that are compliant with proteomics standards and compatible with data repositories [63]. Not only does this help to increase throughput in the proteomics laboratory, but it should also improve the consistency of data processing and minimise failures in the sample audit trail. These systems have been described as ‘pipeline tools’ or ‘workflow systems’ [3, 64]. They include the trans-proteomic pipeline (TPP) [65, 66], ProteinScape (Bruker Daltonics) and Spectrum Mill (Agilent Technologies). However, such systems cannot be expected to work seamlessly with data repositories until the proteomics standards that shape them become mature and stable.



MS-Runs: 2 Annotations: 0 745.85[m/z] 29.2[mn] 116

		FileName	Annotations	Sel. Annotations
1	a	sample1	67	0
2	b	sample2	9	0

Figure 5: Screenshot from MSight mass spectrometry imaging software. Panel (1): Images of two related experiments aligned in the same sheet (A). Panel (2): The same region of both images marked in panel (1) seen in a 3D view. Panel (3): Mass spectrum view. Panel (4): Text report of both experiments. Panel (5): The Workspace window that is used to organise experiments. Panel (6): The Status Bar indicates the number of MS runs in the current sheet and the number of selected annotations. It also indicates coordinates in real value and intensity of the data located under mouse cursor.

Standardisation of experiment description

Where a sufficiently large and organised community exists, it is possible to prevent some of the more common and serious information quality problems by coming to agreement on how members of the community will record and structure their experimental results. A data set may be wholly accurate, but if it is stored in an obscure format, using undocumented conventions, then the information it contains may still be unusable. Standardisation efforts facilitate data archiving and sharing by defining data formats that allow experimental data produced across the community to be described consistently, and by characterising and promoting good practice in data collection and publishing.

The PSI [67] of the Human Proteome Organisation (HUPO) [68] is the principal organisation associated with the development of proteomics standards. Its goal is to facilitate the systematic capture, comparison, exchange and verification of proteomics data [69]. Three different kinds of proteomics standard can be identified, as illustrated in Figure 6:

Minimum information guidelines

In common with other standards bodies, the PSI is defining a collection of documents under the MIAPE heading (Minimum Information About a Proteomics Experiment) that state what should be recorded about a proteomics experiment [70]. The guidelines take the form of a checklist, in which a collection

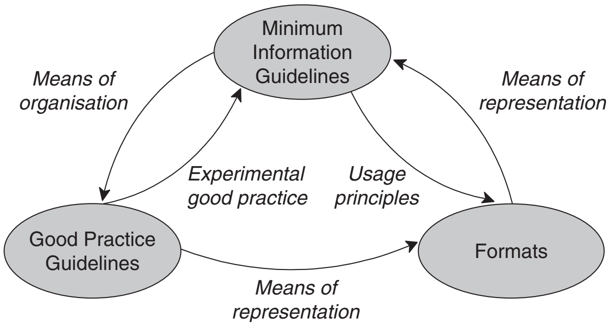


Figure 6: Relationships between types of standard document. An overview of the relationships between the different forms of standard, with the arrows indicating how each one benefits from the other. For example, good practice guidelines provide a means of arguing for the presence of a particular data item in minimum information guidelines, while this latter form of standard helps to provide the basic vocabulary and structuring principles for the definition of the former. Similarly, minimum information guidelines provide a baseline set of concepts for the designers of standard data formats, while the formats themselves help the designers of minimum information guidelines to state more precisely what information is to be considered mandatory and that which is not.

of data properties are defined that together constitute the minimum amount of information required to carry out the stated tasks. For example, in the case of protein identification by peptide mass fingerprinting (see Dry lab analysis for MALDI-TOF MS data section) the number of masses matched to the identified protein, the number of masses not matched in the spectrum (these two can be combined together as the 'hit ratio') and the sequence coverage are properties of the search result that may be used to validate a protein identification [22]. Minimum information guidelines say nothing about the accuracy of the experimental results reported; rather they seek to ensure that sufficient information is recorded about each experiment to allow informed observers to judge the effectiveness of the approach adopted for the problem at hand. In other words, they promote *completeness* of information.

Good practice guidelines

When applying sophisticated experimental techniques, considerable care is required in experiment design and result interpretation. For example, see 'Wet lab quality issues for 2DE' section for a discussion of how many replicate gels should be run in a 2DE experiment. The journals *Molecular and Cellular Proteomics* and *Proteomics* have developed guidelines for authors that specify not only what information should be provided, but also what characterises a well-designed experiment [6, 53]. Good practice guidelines seek to encourage the reporting of high quality results (e.g., results with a small and known false-positive rate) by indicating how experiments should be designed and carried out. In other words, this form of standard promotes *credibility* of information.

Data formats

Given minimum information guidelines, the question remains as to *how* data should be described. Both minimum information and good practice guidelines tend to result in textual descriptions suitable for manual access by scientists. Formats, in contrast, are designed to support computational searching and manipulation of the data. They typically include some form of structured file format, often represented using XML (eXtensible Markup Language), and some form of terminology to be used when populating data elements in the XML file. In the PSI, there is normally a one-to-one correspondence between MIAPE documents [70] and data formats [71]. Proteomics data formats, whether those produced by the PSI or by independent researchers [72, 73], should increase the *consistency* of the data, thereby allowing software to use data from different sites in ways that influence the quality of the data. For example, searches for identifications can be repeated using consistent software settings or underlying sequence databases, thereby allowing results to be compared more directly.

Proteomics data repositories

Public repositories of proteomics data seek to fulfil one of the early hopes of these information-rich experiments, i.e. that added value may be gained by combining data sets from different studies and enabling complex queries to be run over them [74]. Realisation of the difficulties involved in capturing the information that might be useful from such experiments was the driving force behind the proteomics standards movement (see 'Standardisation of experiment description' section). It is important that principles and formats emerging from standards-based approaches are used to shape the structure of proteomics data repositories in order to facilitate querying and re-use of the stored information.

The various proteomics data repositories currently available have been reviewed elsewhere from the viewpoint of data integration [64]. PRIDE [75] is unique in that it contains protein identifications generated from both peptide mass fingerprinting and peptide fragment fingerprinting, whereas the GPM database [47], Open Proteomics Database (OPD) [76] and PeptideAtlas [48] are limited to LC-MS/MS data. The Gene Expression Omnibus (GEO) [77] is mostly concerned with microarray data, but also supports non-array techniques, including MS-based proteomics.

The PRIDE database, as of January 2008, held over 3000 experiments on 26 different species, with 370 000 identified proteins and over 2 million identified peptides with rich experimental descriptions. The 'compare experiments' function in PRIDE produces a Venn diagram showing the protein identifications unique to each experiment and those common to both. PRIDE allows registered users to submit data files generated by the Proteome Harvest Spreadsheet [78]—a Microsoft Excel workbook with functionality that enforces the inclusion of certain required data (mandatory fields) and enables the user to select controlled vocabulary terms from the Ontology Lookup Service, thereby promoting data completeness and consistency.

The OPD consists entirely of a collection of MS/MS data files generated by the host laboratory (University of Texas) together with descriptions of the sample processing procedure and MS parameters used in their generation. The only functionality provided is downloading of the zipped data files.

The GPM contained over 40 million peptide identifications from 14 different proteomes (as of January 2008). It allows the user to search the database with MS/MS data (DTA, PKL or MGF file formats) and gives the option of adding the input data to the database either as a named or anonymous contribution. The GPM database does not store contextual or experimental protocol information.

Although sharing and re-use of proteome data sets is not yet widespread, projects have integrated data from coordinated experiments in multiple laboratories [79], and studies have been undertaken that seek to establish properties of experimental techniques through systematic studies of large data sets (e.g. [80, 81]).

Software tool support for managing data quality in proteomics

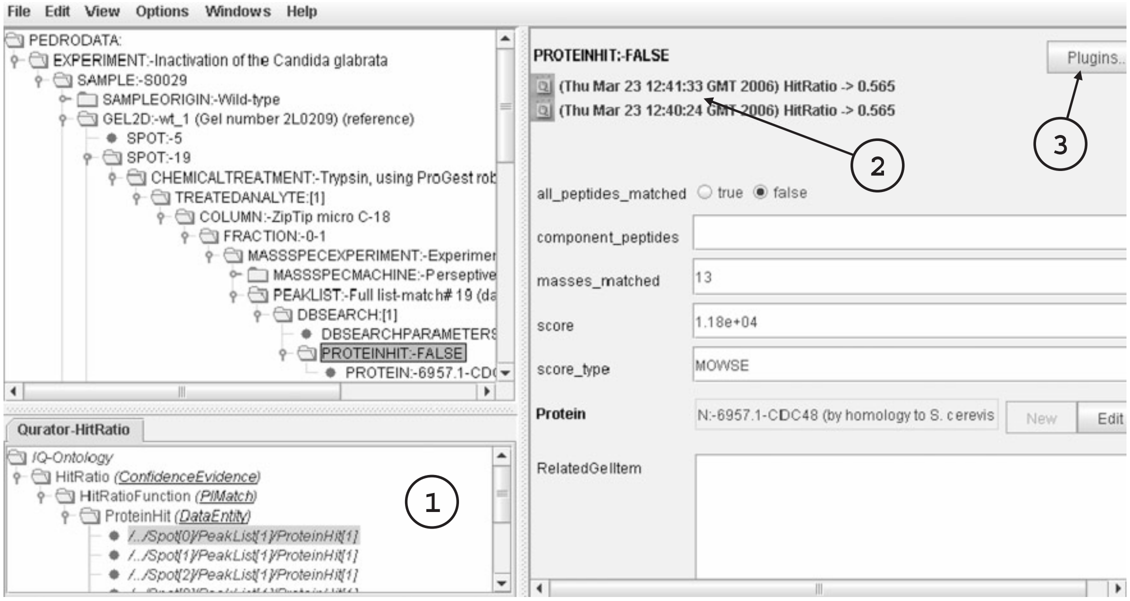
It was recognised early in the work on establishing proteomics data repositories that associated software tools should have functionality built-in to help scientists manage the quality of their data. The PEDRo [74] repository was provided with a software application (originally called the PEDRo Data Collator, later renamed to simply 'Pedro' [82]) which helped ensure that data captured for entry into the repository conformed to the constraints of the XML Schema that defined the PEDRo data model. Moreover, the Pedro tool was designed to allow users to check entered data against controlled vocabularies, to ensure that meaningful values were entered into particular fields.

Developing this idea further, recent work has exploited the 'plug-in' architecture of the Pedro tool in order to allow users to access quality-checking services appropriate for their data. This 'quality-aware' Pedro plugin [83] uses the identifier of the XML Schema (e.g. PEDRo) to look up available quality-checking services on the Web, and allows users to call those services (as Web services) to check the data they have loaded into the tool. Currently, in proteomics, services are available to apply the metrics described earlier in this article [22]. The augmented Pedro tool with quality-aware plugin is shown in Figure 7 and is downloadable from [84].

The approach of making quality-checking services available as Web services has the significant benefit that they can then be re-used within other software environments. As an example, such services can be invoked as part of bioinformatics workflows in the Taverna Workbench [85], to enable automated filtering or flagging of the data as it is processed, according to criteria set by the user. An example of this approach in the proteomics domain is presented in [86].

CONCLUSIONS

The major proteomics technologies (2DE and LC-MS/MS) and their associated data analysis systems are continually under development in order to improve the quality of information generated by proteomics experiments. Increased automation, the provision of well-designed experimental workflows and robust (multivariate) statistical methods, should increase confidence in the results generated by high-throughput experiments. Information standards offer a valuable mechanism by which certain forms


File Edit View Options Windows Help
PEDRODATA:
EXPERIMENT:-Inactivation of the Candida glabrata
SAMPLE:-S0029
SAMPLEORIGIN:-Wild-type
GEL2D:-wt_1 (Gel number 2L0209) (reference)
SPOT:-5
SPOT:-19
CHEMICALTREATMENT:-Trypsin, using ProGest rob
TREATEDANALYTE:[1]
COLUMN:-ZipTip micro C-18
FRACTION:-0-1
MASSSPECEXPERIMENT:-Experimer
MASSSPECMACHINE:-Perseptive
PEAKLIST:-Full list-match# 19 (da
DBSEARCH:[1]
DBSEARCHPARAMETERS
PROTEINHIT:-FALSE
PROTEIN:-6957.1-CD(
Qurator-HitRatio
IQ-Ontology
HitRatio (ConfidenceEvidence)
HitRatioFunction (PIMatch)
ProteinHit (DataEntity)
/../Spot[0]/PeakList[1]/ProteinHit[1]
/../Spot[1]/PeakList[1]/ProteinHit[1]
/../Spot[2]/PeakList[1]/ProteinHit[1]
1
2
3
Plugins..
New
Edit

PROTEINHIT:-FALSE

(Thu Mar 23 12:41:33 GMT 2006) HitRatio -> 0.565

(Thu Mar 23 12:40:24 GMT 2006) HitRatio -> 0.565

all_peptides_matched ○ true ● false

component_peptides

masses_matched 13

score 1.18e+04

score_type MOWSE

Protein N:-6957.1-CDC48 (by homology to S. cerevis

RelatedGelItem

Figure 7: 'Quality-aware' version of the Pedro tool. The 'quality aware' version of Pedro introduces lower-left panel (1), which contains a tree view of the various information quality attributes. An *IQ ontology* [83] relevant to the loaded data model is used to populate panel (1) with concepts including test functions, annotatable data elements, IQ indicators and metrics. For the PEDRo data model, these include attributes such as *HitRatio* and *MassCoverage*. This panel allows users to discover available indicators for the data model at hand, and follow hyperlinks to explore related quality attributes. Annotations on the selected data element are shown in the right-hand panel at (2), summarised along with basic provenance information (test function used, timestamp). The 'Plugins' button (3) invokes available annotation services, using bindings to the displayed data element (and to the loaded data model) to look up these services.

of quality (in particular, completeness, credibility and consistency) can be attained by the proteomics community. Repositories, such as PRIDE, that are closely linked to the standards initiatives, play an important role in sharing experimental data and are becoming another form of quality assurance, alongside peer-review, for the publication of large-scale functional genomic studies. Finally, the need for scientists to apply their own measures of information quality has been stated, and the importance of having 'quality-aware' software tools, which make quality measures more explicit has been highlighted.

Key Points

- The variable and hidden quality of proteomics data sets are major barriers to their maximal exploitation.
- Careful experimental design and choice of quality control procedures is important for maintaining the quality of information that emerges at the end of the proteomics experimental workflow.
- Several computational tools are available to validate the results at various stages of proteomics experiments.
- Community-wide adoption of standards for information capture and formatting is a key tool in the management of information quality in proteomics.
- Information management environments with the capability to measure, filter and flag proteomics data based on its quality are beginning to emerge.

Acknowledgements

We thank Dr Zhikang Yin, School of Medical Sciences, University of Aberdeen for providing the data on which Figure 2 is based. Funds were provided by Engineering and Physical Sciences Research Council, GR/S67593/01 (SME) and GR/S67609/01 (AP).

References

1. Rocke DM. Design and analysis of experiments with high throughput biological assay data. *Semin Cell Dev Biol* 2004; **15**:703–13.
2. Biron DG, Brun C, Lefevre T, *et al*. The pitfalls of proteomics experiments without the correct use of bioinformatics tools. *Proteomics* 2006;**6**:5577–96.
3. Palagi PM, Hernandez P, Walther D, *et al*. Proteome informatics I: bioinformatics tools for processing experimental data. *Proteomics* 2006;**6**:5435–44.
4. Gorg A, Weiss W, Dunn MJ. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* 2004; **4**:3665–85.
5. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;**422**:198–207.
6. Carr S, Aebersold R, Baldwin M, *et al*. The need for guidelines in publication of peptide and protein identification data. *Mol Cell Proteomics* 2004;**3**:531–3.
7. Souchelnytskyi S. Bridging proteomics and systems biology: what are the roads to be traveled? *Proteomics* 2005; **5**:4123–37.

8. Bell AW, Nilsson T, Kearney RE, *et al*. The protein microscope: incorporating mass spectrometry into cell biology. *Nat Methods* 2007;**4**:783–4.
9. Lexander H, Hellman U, Palmberg C, *et al*. Evaluation of two sample preparation methods for prostate proteome analysis. *Proteomics* 2006;**6**:3918–25.
10. Thongboonkerd V, Chutipongtanate S, Kanlaya R. Systematic evaluation of sample preparation methods for gel-based human urinary proteomics: quantity, quality, and variability. *J Proteome Res* 2006;**5**:183–91.
11. Schwarz K, Fiedler T, Fischer RJ, *et al*. A standard operating procedure (SOP) for the preparation of intra- and extracellular proteins of Clostridium acetobutylicum for proteome analysis. *J Microbiol Methods* 2007;**68**:396–402.
12. Hunt SM, Thomas MR, Sebastian LT, *et al*. Optimal replication and the importance of experimental design for gel-based quantitative proteomics. *J Proteome Res* 2005;**4**: 809–19.
13. Karp NA, Spencer M, Lindsay H, *et al*. Impact of replicate types on proteomics expression analysis. *J Proteome Res* 2005; **4**:1867–71.
14. Gorg A, Postel W, Gunther S. The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* 1988;**9**:531–46.
15. Yin Z, Stead D, Selway L, *et al*. Proteomic response to amino acid starvation in Candida albicans and Saccharomyces cerevisiae. *Proteomics* 2004;**4**:2425–36.
16. Zivy M. Quantitative analysis of 2D gels. *Methods Mol Biol* 2007;**355**:175–94.
17. Karp NA, Lilley KS. Maximising sensitivity for detecting changes in protein expression: experimental design using minimal CyDyes. *Proteomics* 2005;**5**:3105–15.
18. Marengo E, Robotti E, Bobba M, *et al*. Multivariate statistical tools applied to the characterization of the proteomic profiles of two human lymphoma cell lines by two-dimensional gel electrophoresis. *Electrophoresis* 2006;**27**: 484–94.
19. Engkilde K, Jacobsen S, Sondergaard I. Multivariate data analysis of proteome data. *Methods Mol Biol* 2007;**355**: 195–210.
20. Hu J, Coombes KR, Morris JS, *et al*. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic* 2005;**3**:322–31.
21. Lim H, Eng J, Yates JR 3rd, *et al*. Identification of 2D-gel proteins: a comparison of MALDI/TOF peptide mass mapping to μ LC-ESI tandem mass spectrometry. *J Am Soc Mass Spectrom* 2003;**14**:957–70.
22. Stead DA, Preece A, Brown AJP. Universal metrics for quality assessment of protein identifications by mass spectrometry. *Mol Cell Proteomics* 2006;**5**:1205–11.
23. Aldente. http://ca.expasy.org/tools/aldente/ (7 January 2008, date last accessed).
24. Tuloup M, Hernandez C, Coro I, *et al*. Aldente and BioGraph: an improved peptide mass fingerprinting protein identification environment. Swiss Proteomics Society 2003 Congress: Understanding Biological Systems through Proteomics, Basel, Switzerland, 2–4 December, 2003, Ed. FontisMedia (ISBN 2-88476-004-0), 174–176.
25. Mascot. http://www.matrixscience.com/search_form_select.html/ (7 January 2008, date last accessed).
26. ProFound. http://prowl.rockefeller.edu/prowl-cgi/profound.exe/ (7 January 2008, date last accessed).
27. Chamrad DC, Korting G, Stuhler K, *et al*. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics* 2004;**4**: 619–28.
28. Hunt DF, Yates JR 3rd, Shabanowitz J, *et al*. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A* 1986;**83**:6233–7.
29. Nesvizhskii AI. Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol Biol* 2007;**367**:87–119.
30. Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;**66**:4390–9.
31. SEQUEST. http://fields.scripps.edu/sequest/ (7 January 2008, date last accessed).
32. Boutilier K, Ross M, Podtelejnikov AV, *et al*. Comparison of different search engines using validated MS/MS test datasets. *Anal Chim Acta* 2005;**534**:11–20.
33. Kapp EA, Schutz F, Connolly LM, *et al*. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 2005;**5**:3475–90.
34. Phenyx main site. http://www.phenyx-ms.com/ (7 January 2008, date last accessed).
35. Phenyx public server. http://phenyx.vital-it.ch/pwi/ (7 January 2008, date last accessed).
36. OMSSA. http://pubchem.ncbi.nlm.nih.gov/omssa/ (7 January 2008, date last accessed).
37. X! Tandem. http://www.thegpm.org/TANDEM/index.html/ (7 January 2008, date last accessed).
38. Orchard S, Jones P, Taylor C, *et al*. Proteomic data exchange and storage: the need for common standards and public repositories. *Methods Mol Biol* 2007;**367**:261–70.
39. Bern M, Goldberg D, McDonald WH, *et al*. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics* 2004;**20**(Suppl 1):I49–54.
40. Flikka K, Martens L, Vandekerckhove J, *et al*. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics* 2006;**6**: 2086–94.
41. Nesvizhskii AI, Roos FF, Grossmann J, *et al*. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* 2006;**5**:652–70.
42. Wong JW, Sullivan MJ, Cartwright HM, *et al*. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics* 2007;**8**:51.
43. Ulintz PJ, Zhu J, Qin ZS, *et al*. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol Cell Proteomics* 2006;**5**:497–509.
44. Shadforth I, Dunkley T, Lilley K, *et al*. Confident protein identification using the average peptide score method

coupled with search-specific, ab initio thresholds. *Rapid Commun Mass Spectrom* 2005;**19**:3363–8.

45. Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol Cell Proteomics* 2005; **4**:1180–8.
46. Piening BD, Wang P, Bangur CS, *et al*. Quality control metrics for LC-MS feature detection tools demonstrated on Saccharomyces cerevisiae proteomic profiles. *J Proteome Res* 2006;**5**:1527–34.
47. Global Proteome Machine. http://gpmdb.thegpm.org/ (7 January 2008, date last accessed).
48. PeptideAtlas. http://www.peptideatlas.org/ (7 January 2008, date last accessed).
49. Desiere F, Deutsch EW, King NL, *et al*. The PeptideAtlas project. *Nucleic Acids Res* 2006;**34**:D655–8.
50. Feng J, Naiman DQ, Cooper B. Probability-based pattern recognition and statistical framework for randomization: modeling tandem mass spectrum/peptide sequence false match frequencies. *Bioinformatics* 2007;**23**:2210–7.
51. Higgs RE, Knierman MD, Freeman AB, *et al*. Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. *J Proteome Res* 2007;**6**: 1758–67.
52. Huttlin EL, Hegeman AD, Harms AC, *et al*. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J Proteome Res* 2007;**6**:392–8.
53. Wilkins MR, Appel RD, Van Eyk JE, *et al*. Guidelines for the next 10 years of proteomics. *Proteomics* 2006;**6**:4–8.
54. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* 2005;**1**:252–62.
55. Yan W, Chen SS. Mass spectrometry-based quantitative proteomic profiling. *Brief Funct Genomic Proteomic* 2005;**4**: 27–38.
56. Chen X, Sun L, Yu Y, *et al*. Amino acid-coded tagging approaches in quantitative proteomics. *Expert Rev Proteomics* 2007;**4**:25–37.
57. Ong SE, Blagoev B, Kratchmarova I, *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 2002;**1**:376–86.
58. Schmidt A, Kellermann J, Lottspeich F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* 2005;**5**:4–15.
59. Ross PL, Huang YN, Marchese JN, *et al*. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004;**3**:1154–69.
60. Linsen L, Locherbach J, Berth M, *et al*. Visual analysis of gel-free proteome data. *IEEE Trans Vis Comput Graph* 2006; **12**:497–508.
61. Palagi PM, Walther D, Quadroni M, *et al*. MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005;**5**:2381–4.
62. MSight. http://www.expasy.org/MSight/ (7 January 2008, date last accessed).
63. Levander F, Krogh M, Warell K, *et al*. Automated reporting from gel-based proteomics experiments using the open source Proteios database application. *Proteomics* 2007;**7**: 668–74.
64. Lisacek F, Cohen-Boulakia S, Appel RD. Proteome informatics II: bioinformatics for comparative proteomics. *Proteomics* 2006;**6**:5445–66.
65. Trans-Proteomic Pipeline. http://tools.proteomecenter.org/software.php/ (7 January 2008, date last accessed).
66. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 2005;**1**:2005.0017 [Epub 2005 Aug 2].
67. Proteome Standards Initiative. http://www.psidev.info/ (7 January 2008, date last accessed).
68. Human Proteome Organisation. http://www.hupo.org/ (7 January 2008, date last accessed).
69. Hermjakob H. The HUPO proteomics standards initiative – overcoming the fragmentation of proteomics data. *Proteomics* 2006;**6**:34–8.
70. Taylor CF, Paton NW, Lilley KS, *et al*. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* 2007;**25**:887–93.
71. Hermjakob H, Montecchi-Palazzi L, Bader G, *et al*. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol* 2004;**22**:177–83.
72. Taylor CF, Paton NW, Garwood KL, *et al*. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol* 2003;**21**: 247–54.
73. Pedrioli PG, Eng JK, Hubley R, *et al*. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004;**22**: 1459–66.
74. Garwood K, McLaughlin T, Garwood C, *et al*. PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* 2004;**5**: 68.
75. PRIDE. http://www.ebi.ac.uk/pride/ (7 January 2008, date last accessed).
76. Open Proteomics Database. http://bioinformatics.icmb.utexas.edu/OPD/ (7 January 2008, date last accessed).
77. Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/ (7 January 2008, date last accessed).
78. Proteome Harvest Spreadsheet. http://www.ebi.ac.uk/pride/proteomeharvest/index.html (7 January 2008, date last accessed).
79. Omenn GS, States DJ, Adamski M, *et al*. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005;**5**:3226–45.
80. Prakash A, Piening B, Whiteaker J, *et al*. Assessing bias in experiment design for large scale mass spectrometry-based quantitative proteomics. *Mol Cell Proteomics* 2007;**6**:1741–8.

81. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007;**4**:787–97.

82. PEDRo Data Collator. http://pedrodownload.man.ac.uk/ (7 January 2008, date last accessed).

83. Preece A, Jin B, Missier P, *et al*. Towards the management of information quality in proteomics. *Proceedings of the 19th IEEE International Symposium on Computer-Based Medical Systems* (CBMS 2006), 2006, 936–40.

84. Qurator. http://www.qurator.org/ (7 January 2008, date last accessed).

85. Taverna. http://taverna.sourceforge.net/ (7 January 2008, date last accessed).

86. Missier P, Embury S, Greenwood M, *et al*. Quality views: capturing and exploiting the user perspective on data quality. *Proceedings of the 32nd International Conference on Very Large Data Bases* (VLDB 2006), Seoul, Korea, 2006, 977–88.