Informatics solutions for high-throughput proteomics

Thodoros Topaloglou

Information Engineering Center, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, M5S 3G8, Canada

The success of mass-spectrometry-based proteomics as a method for analyzing proteins in biological samples is accompanied by challenges owning to demands for increased throughput. These challenges arise from the vast volume of data generated by proteomics experiments combined with the heterogeneity in data formats, processing methods, software tools and databases that are involved in the translation of spectral data into relevant and actionable information for scientists. Informatics aims to provide answers to these challenges by transferring existing solutions from information management to proteomics and/or by generating novel computational methods for automation of proteomics data processing.

The rapid advancement of mass spectrometry (MS), the development of new laboratory techniques, such as isotope-tagging methods, and the demand for protein biomarker discovery, have led to a rapid increase in the number, size and rate at which proteomics datasets are generated. Managing and extracting valuable information from such datasets is a new challenge for proteomics laboratories. There is currently a limited number of mature commercially available data management platforms that enable the acquisition, organization, analysis and frequent querying of high-volume MS-based proteomics data.

The goal of proteomics is to study all the expressed proteins in a biological system [1]. The realization of this goal depends on three interdependent technological factors: instrument sensitivity, protein separation technology and data analysis software. Although the thrust for higher sensitivity of analysis continues, we currently have the technological capabilities to identify most of the proteins in a sample, quantify large numbers of proteins in a sample [2] and interrogate complex sample types, such as whole cell lysates, tissues, blood and urine [3,4]. This allows the application of proteomics in the study of disease processes, the development of new diagnostics and the acceleration of the drug development process [5]. A common requirement of these applications is higher throughput.

Throughput refers to the rate of experiments performed per unit of time. In addition to the complexity of data management, a concern in high-throughput efforts relates to the speed and comprehensiveness of data analysis. Typically, there is a trade-off between the level of throughput and the comprehensiveness of the analysis. Informatics has an important role to play in balancing this trade-off. Automation of computational tools enables fast processing of large datasets based on common parameters but can admit some false answers, or miss correct ones. The alternative is to incorporate a manual review step, which slows down the throughput of data processing. High-throughput processing necessitates informatics strategies that combine automated tools with empirical models to filter out the false answers. The development of such models involves expert knowledge and methodologies based on statistical quality-control theory, or statistical baseline information from prior experiments to help to differentiate between technical and experimental outliers.

Until recently, the focus of informatics in proteomics, henceforth referred to as proteoinformatics, was the development of methods for interpreting spectral data and, particularly, tools for identification (Mascot [6], Sequest [7], Pepsea [8], X!Tandem [9]) and quantitation (Xpress [10], RelEx [11]) of MS–MS and MS spectra. As the volume of data produced by liquid chromatography (LC)-MS experiments and subsequent analyses, became an issue, stand-alone tools were extended with data management

Corresponding author: Topaloglou, T. (thodoros@mie.utoronto.ca)

1359-6446/06/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2006.04.011

functionality (Mascot Integra, Spectrum Mill) or integrated in laboratory information management systems (LIMS). Such systems are common in most proteomics laboratories; they are either developed in-house using relational database and/or web technology, or purchased from a vendor.

The particular needs of large-scale high-throughput operations are not easily supported by a single LIMS system. In a high-throughput proteomics (HTP) facility that supports multiple investigators, projects and proteomics analysis protocols, such as the proteomics laboratory at the Pacific Northwest National Laboratory, the data management requirements range from handling collaborators' data to the management of resources, samples, data files, production processes and analysis results [12]. It is uncommon for all these functions to be part of a single information system because they differ in subject matter and are performed by different units within an organization. Most often, different functions are implemented by different systems, which need to cooperate to support the entire proteomics laboratory data lifecycle. The shift from disparate systems to an integrated platform is now seen as essential in industrial laboratories and academic research facilities [13] conducting large-scale proteomics studies.

HTP requires more than just scaling-up informatics. It requires rapid production of high-quality samples with reduced protein complexity and a robust and reproducible method of profiling samples. The two main priorities of informatics in HTP are (i) acceleration of computational methods and (ii) information management throughout the proteomic-data lifecycle. The objective of this review is to discuss these priorities and to describe the components of the informatics architecture for HTP. The goals and challenges will be introduced for each component, and recent solutions that have emerged in the literature and in commercially available products will be discussed. An implemented platform that was used in a high-throughput industrial proteomics laboratory will be briefly described.

This review focuses on MS-based proteomics, which is the only proteomics discipline that currently works at high throughput. Emergent array-based proteomic platforms [14] are promising for high throughput but still in early phase.

Computational methods for high-throughput proteomics

LC-MS–MS experiments involve the collection of thousands of MS–MS spectra generated from a sample digested with a proteolytic enzyme (e.g. trypsin). One goal of these experiments is to identify the proteins that are present in the sample. This analysis includes assigning peptides to spectra, validating these peptide assignments to remove incorrect results and inferring protein identities from the assigned peptides. A second goal is to compare protein expression between two or more samples, which involves reconstructing the LC elution profiles of peptides, quantifying the abundance of peptides and enabling the comparison of peptides or proteins between samples. The core concepts and technical underpinnings in protein identification and quantitation, and strategies that apply them in high-throughput conditions are described in this section.

Protein identification

The idea of identifying proteins using peptide MS–MS spectra to search protein sequence databases [8,15,16] marks the start to the development of search engines and the automation of protein identification. The basic principle of a search engine is described in Box 1. Search engines, although significantly advancing the throughput of protein identification, have several limitations. Understanding their limitations is essential for their successful application in high-throughput environments.

One limitation of the search engine approach is that many of the experimentally acquired spectra do not match any database-derived spectra, or they are mismatched (Figure I). A lot of recent work has tackled this problem by trying to minimize unmatched or mismatched MS–MS spectra using improved scoring schemes [17], filtering of bad spectra [18] and exploring proteotypic peptide libraries [19]. A recent review by Johnson *et al.* [20] provides a detailed account of why not all MS–MS spectra match a sequence entry. Several studies [21,22] have compared results from different search engines. Boutilier *et al.* [23] recommend using two search engines for better coverage of the protein space. The choice and size [24,25] of the database also plays a role in peptide matching. A large, comprehensive, multi-species database increases the chances of a match and of false-positives, whereas a smaller species-specific database reduces false-positive matches. The protein reference sequence database from the US National Center of Biotechnology Information (NCBI) is an example of a popular non-redundant, multi-species database, whereas the International Protein Index (IPI) from the European Bioinformatics Institute (EBI) is a curated and species-specific protein sequence database. (For a review of the recent developments in the field of protein sequence databases, see Ref. [26].)

The inference of proteins from MS–MS peptide hits is probabilistic. A key issue in high-throughput studies is to establish acceptance criteria of protein identifications based on peptide evidence. Search engine score cut-offs are not a sufficient solution. A revealing analysis by Carlige *et al.* [25] shows that large databases, such as those representing eukaryotic systems, contain enough peptides so that a random match with a high score is likely to occur. This problem is addressed by methods that complement the search engine by applying probabilistic statistical models to compute the confidence of protein assignments [17,21,27]. These methods allow filtering high-throughput datasets at a predictable sensitivity and false-positive identification error rates. Another approach to inferring proteins identities from peptide hits that is suitable for high-throughput processing, is to combine peptide evidence from several searches (replicate experiments) and select proteins following the principle of parsimony [28]: peptides mapping to a single rather than a complicated set of proteins are most likely to account for the observed data. Finally, a different approach to protein inference emerged from the hypothesis that there is class of 'proteotypic' peptides that are more likely to be detected with confidence by current MS-based methods and mapped to proteins. If a database of such peptides and the protein sequences that contain them becomes available, then the problem of protein assignment is simplified to a simple database lookup. Characterizing the properties of proteotypic peptides is still a research topic [29]. There are currently efforts by various groups to construct repositories of such data (e.g. PeptideAtlas and GPMdb) [19,30].

There are different search engines and protein identification methods (reviewed in Refs [20,22,31]) several of which are

BOX 1

Protein identification

Identification of proteins by mass spectrometry (MS) is assisted by search engines. The basic principle of a search engine is matching the fragmentation pattern of MS–MS spectra representing peptides to the theoretical spectra that are derived computationally from a protein sequence database. In the database, each protein sequence is cleaved *in silico* with the same proteolytic enzyme as in the MS experiment, producing peptide sequences. The mass of each peptide is calculated, and for those peptides whose mass matches that of an experimentally generated peptide – plus or minus a tolerance – a theoretical spectrum is calculated. The similarity of the theoretical to the experimental spectrum is captured in a score, which is used to select peptide matches (hits). Although their basic principle is always the same, search engines differ in how they score hits. Some use partially interpreted spectra, the peptide sequence tag approach, to assist the matching of experimental to theoretical spectra (Pepsea), but the dominant trend is to search uninterpreted spectra against the database. This approach is used by several popular search engines, including Mascot, Sequest and X!Tandem.

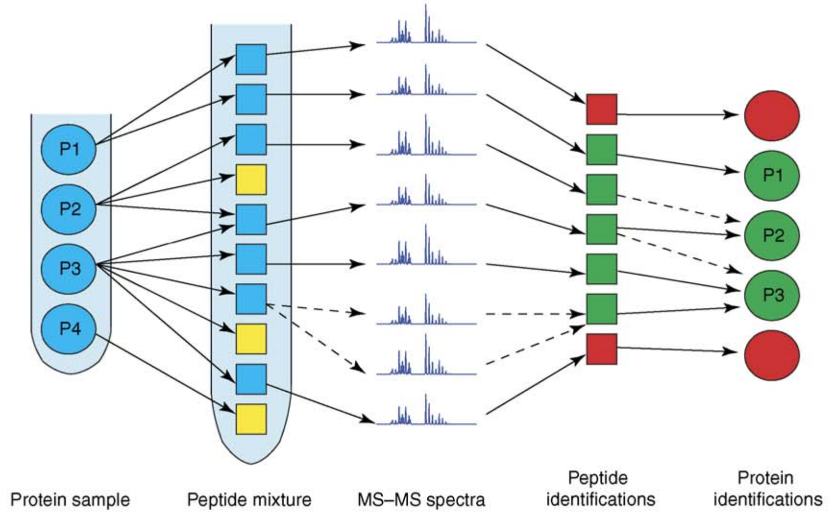

FIGURE I

Protein identification process. Circles represent proteins and boxes represent peptides. Each sample protein is cleaved into smaller peptides. Peptides are then ionized, and selected peptides are fragmented to produce MS–MS spectra. Some of these peptides (yellow) do not ionize. Other peptides are selected for fragmentation multiple times (broken arrows) and therefore yield multiple MS–MS spectra. Acquired MS–MS spectra are searched against a sequence database and assigned a best-matching peptide. Correct peptide assignments are shown in green, incorrect assignments in red. The list of peptides is then used to infer the proteins that are present in the sample. Identified proteins that are present in the original sample are shown as green circles; false-positives are shown as red circles. Protein P4 is not detected and it is a false negative. Adapted, with permission, from [17].

amenable to high throughput. Not many comparative studies are available to educate practitioners on how to choose one. What makes evaluating and comparing protein identification methods so difficult is the availability of reliable benchmark datasets where the correct and incorrect spectra are characterized. It is encouraging that independently generated datasets of this type are now emerging [32] (www.abrf.org/index.cfm/group.show/Proteomics.34.htm). Another important question is how to apply identification methods. Recently, a working group of proteomics experts published a set of guidelines [33] on how to report peptide and protein identification data. Their two principal suggestions are, first, to report the supporting information so that calls can be reviewed and confirmed and, second, to increase the stringency of protein assignments based on a single peptide hit.

Quantitation

Quantitative measurement of peptide and protein expression largely depends on data analysis tools to process and compare spectral data [34–36], and is essential for applying proteomics in disease and drug research. A description of the principle of MS quantitation appears in Box 2.

Data analysis of quantitative MS measurements uses advanced algorithms for de-isotoping, charge-state and peak detection [37], peptide recognition [38–40], label recognition, peptide matching [41], peak alignment between samples [42–44] and data normalization [45]. The description of these algorithms is beyond the scope of this review. For a detailed account of the computational and statistical methods in quantitative proteomics, see a recent review by Listgarten and Emili [46]. Effectively, the quantitative analysis workflow converts mass peaks from MS scans to data points in the mass–time space. The data points correspond to peptides, which enables the detection and measurement of differences in peptide expression. Figure 1 provides a visual illustration of this process.

High-throughput quantitative proteomics studies need tens or hundreds of comparisons of two or more samples. There are two strategies that support these analyses. The first strategy uses the computational step that is outlined in the previous paragraph to construct and compare peptides, representing data points in the

BOX 2

Mass spectrometry quantitation

Current mass spectrometry (MS)-based approaches to quantitative proteomics are either based on isotope-tagging methods (e.g., ICAT) or use the absolute ion intensity to quantify peptides. In ICAT, two protein mixtures are labelled with heavy and light reagents. The labelled mixtures are combined and digested using trypsin. The peptide mixture is analyzed with liquid chromatography (LC)-MS. The produced spectra contain masses of the same peptide from both samples that are separated by the mass difference between the heavy and light labels. As peptides elute in LC over a period of time, charged ions are detected in successive MS scans at a precise mass over charge value (m/z). Peptide elution profiles are reconstructed by tracking ion masses over time. The area below a peptide profile curve is a measure of its abundance. In the case of sequential analysis (unlabelled samples profiled sequentially), the same data analysis principle applies except that elution profiles have to be aligned and normalized before they are compared. An essential part of peptide quantitation techniques (labelled and unlabelled) is availability of software tools to reconstruct peptide elution profiles and calculate their abundance. Figure II shows the basic principle of ICAT.

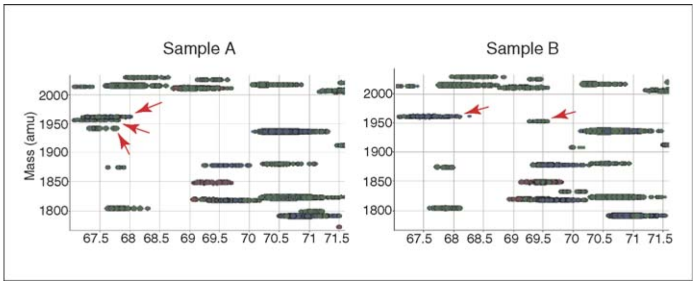

FIGURE II

ICAT quantitation strategy. Red and the green ion elution profiles come from samples A and B, respectively, and appear to correspond to the same peptide because the profiles are apart by the heavy–light mass difference and peak approximately at the same time. The peptide from sample A appears to be twice as abundant as its matching peptide of sample B. The peptide shown in blue is not matched and it is unclear which sample it is measured from.

mass–time space across multiple runs [38,45], and then transfers these changes to a data matrix similar to those used in gene-expression analysis [47] where questions of normalization and statistical significance of differences can be addressed using existing methodology. The differential peptides are selected for identification by MS–MS. This strategy is better known as the ratio-driven approach. The second strategy involves a comprehensive MS–MS analysis to first build a map of peptide identifications in mass–time space and then perform MS-only experiments for quantitation [48]. These two strategies enhance throughput by reducing or replacing the expensive and time-consuming MS–MS-based identification experiments with faster computational inference.

Information management requirements of high-throughput proteomics

The advances in proteoinformatics and the proliferation of methods for protein identification and quantitation, address a particular bottleneck – the MS data interpretation – that is only one phase of the proteomics data lifecycle. They also introduce a new challenge: the interoperability and integration of applications and data in the proteomics laboratory. Let us take as an example the proteomics facility at the Institute of Systems Biology (ISB, www.proteomecenter.org). They use a range of instruments that produce MS data in proprietary binary formats. Extracted peak lists from the binary data are transformed into the appropriate data format accepted by the search engine (e.g. .dta for Sequest). The search results are directed to Peptide Prophet for identification, and to XPRESS for relative quantitation. This demonstrates that even routine analyses require the use of several applications and data formats. To simplify this complexity, laboratories need to store, organize and validate the results, as well as information management that improves the coordination of analytical data-processing steps. The solution used by ISB is the introduction of the Trans-Proteomic Pipeline [13] and the adoption of open XML formats for storage and data exchange [49].

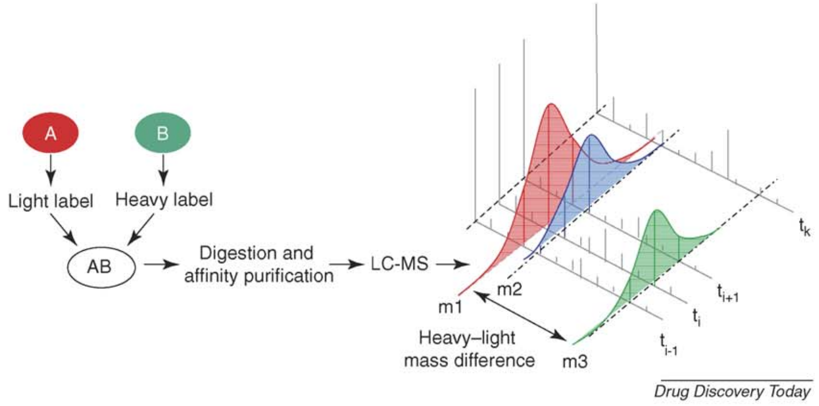

FIGURE 1

Time–mass space. Peptide profiles from samples A and B are projected in the time-mass space. Peptide profiles appear as a connected horizontal sequence of dots of the same mass (*y* axis) over a time span (*x* axis) – each dot corresponds to a measured ion. The colour of the dot corresponds to the charge state at which it was detected and the size corresponds to its intensity. A matching peptide in both samples appears at the same mass and time coordinates plus or minus tolerance. Differential peptides (red arrows) can be present in one and not in the other sample, or appear in both samples but their profiles differ in length and dot size. Plots produced by M. Dharsee.

Proteomics data lifecycle

The lifecycle of proteomics data involves the following discrete phases: sample management, MS data acquisition, spectral data processing, biological interpretation, and reporting and dissemination.

Sample management. The success of proteomics experiments depends on careful experimental design and high-quality samples [50]. Equally important is the documentation of sample parameters, quality factors and study structure in databases that guide the analysis and interpretation of proteomics studies. Without careful sample ascertainment and/or the availability of detailed sample annotations, the interpretation of the study results can be difficult. Biological samples go through a complex preparation process before MS [51]. Part of the preparation involves fractionation or gel-separation experiments, which alter the sample by producing other derivative samples. Documentation of the fractionation, pooling and replication of samples is important not only for interpretation, but also for tracking experiments in the laboratory pipeline. Management of samples and sample-processing steps are onerous tasks that must be entrusted to systems with the appropriate functionality. Such functionality is offered by several special-purpose commercial LIMS products for proteomics (ProteusLIMS, Nautilus and Mascot Integra).

Data acquisition. The commercially available software for instrument control and data acquisition in MS contributes to, but is not sufficient to handle, the information flow needs of the high-throughput laboratory. Because a laboratory often uses multiple instruments from multiple vendors, a broader data-management solution is necessary. The functionality of such a solution includes file management, linking to the sample tracking and documentation of the protocol parameters, such as operator, settings and so forth. This functionality is provided by a LIMS system, occasionally the same but often different to the sample management one.

Data processing. As mentioned earlier, there is an array of methods for interpreting spectral data to identify or quantify proteins. The problem is how to integrate different and heterogeneous applications into high-throughput scientific workflows. Heterogeneity is encountered at different levels: in the data formats that application generate or receive as input, in different platforms that applications are running on; and in differences in definitions of common concepts that are shared by multiple applications. Several ‘out of the box’ systems support generic data processing workflows (EPICenter [52], Scaffold) but perhaps not the exact workflow of a specific laboratory. The fields of scientific workflow management research [53] and process modelling are making progress on these questions but their results are not yet visible in the field of proteomics. As the integration requirements of proteomics increase, we believe that collaboration between the two disciplines is inevitable. Ideally, scientists need components like Mascot, data formats like mzXML, ‘connectors’ like mzXML2<other> [where <other> is one of the proprietary formats of search engines (e.g. dta for Sequest, mgf for Mascot etc.)], and a high-level notation like business process execution language (BPEL) [54] to build virtual data analysis pipelines that run on a distributed infrastructure

TABLE 1

End-to-end informatics platform for high-throughput proteomics[a,b]

Stage	Sample processing	MS acquisition	Spectral analysis	Interpretation	Data reporting
Goals	Maintain sample data records. Document sample processing steps.	Collect MS data. Document MS analysis process. Manage data files.	Identify and quantify proteins from mass spectra. Document analysis steps.	Evaluate biological relevance. Summarize lists of proteins from experiments.	Evaluate statistical significance. Disseminate comprehensive result reports to collaborators.
Data	Sample record. Quality control Assays. Study info.	Acquisition parameters. Quality control protocol tracking.	Mass spectra. Peptides and proteins. Peptide–protein differentials.	Protein sequences. Annotations. Protein lists.	Data tables. Plots. *p* Values.
Processes		Workflow management. Data file management.	Quantitation. Peptide ID. Integration. Differential analysis. Validation.	Protein annotation. Sequence analysis. Literature mining. Pathway analysis.	Statistical confidence. Biological relevance. Reporting. Data mining.
Systems	Sample tracking LIMS.	Sample tracking LIMS. Custom scripts to transfer files to centralized storage. MS vendor software.	MS database. MS data extractor. Mascot cluster. Differential analysis software. MS data viewer. Spectra browser. Analytical workflows.	AIDA bioinformatics database. Discovery portal bioinformatics analysis environment. Workspace db. Variety of external data sources and tools.	Spotfire visualization. *R* statistical routines. Results database. Custom reports.
Support	DB2 DBMS.	DB2 DBMS. Network of mass spectrometers. Network storage.	Turboworx workflow manager. DB2 DBMS. DB2 Info Integrator (Discovery Link). Computing farm.	DB2 DBMS. DB2 Info Integrator. Jetspeed. Hybernate. Lucene. Computing farm.	DB2 DBMS. *R*. Spotfire decision site.

[a] The platform supports the goals, data and processes of all stages of data lifecycle in a proteomics laboratory. The implemented component systems and the supporting technology are displayed by stage. In addition, a significant aspect of the platform is a data- and workflow-integration strategy that is implicit in the support layer.

[b] Abbreviations: DBMS, database management system; LIMS, laboratory information management system; MS, mass spectrometry.

(such as the grid [55]). The area of data standards for proteomics is extremely active and guided by consortia such as the Proteomics Standards Initiative [56] of the Human Proteome Organization and the Functional Genomics Ontology group (http://fugo.sourceforge.net).

Biological interpretation. There is no shortage of data results in HTP experiments. But do the results make sense? Are they related to the biological inputs in a meaningful way? For example, a study identifies differentially expressed proteins in disease (compared with control samples). Do the results corroborate or contradict what is known from other sources? To answer this type of question, elaborate bioinformatics environments have been built that integrate protein annotations from all major public domain sources [57]. Much can be learned about the biology of an experiment by examining the available annotation for the proteins resulting from it – such as sequence-based annotations (e.g. functional domains and signal peptides) or function-based annotations (e.g. biological process, pathways, etc.). Such systems are available in the public domain [57,58] or from commercial vendors, and can also be built to meet the needs of a specific organization.

Data reporting. Two important factors related to data reporting are confidence and provenance. Much like gene expression analysis, protein expression analysis uses statistical reasoning to identify proteins that are changing in a significant way. Statistical analysis of high-throughput LC-MS experiments generates sizable datasets consisting of average intensities, differential effects, p values and summary statistics on peptides and proteins. It is often necessary to preserve the results of statistical analysis in a database. Such databases enable the

TABLE 2

List of proteomics software systems and resources

Software	Focus	URL
Mascot	MS–MS search engine	www.matrixscience.com
Mascot Integra	Data management for proteomics	www.matrixscience.com
Sequest	MS–MS search engine	www.thermo.com
EPICenter	Data management and validation for peptide ID data	www.proxeon.com
Spectrum Mill	MS–MS search engine environment	www.agilent.com
Proteus LIMS	Proteomics data management LIMS	www.genologics.com
Peptide Prophet	High-throughput validation of peptide identifications	www.proteomecenter.org
Protein Prophet	Protein identification (statistical)	www.proteomecenter.org
X!Tandem	Open source search engine for MS–MS	www.thegpm.org
GPM	Public database of identified peptides	www.thegpm.org
XPRESS	Quantitative differential analysis for ICAT	www.proteomecenter.org
SBEAMS	Systems biology experiments analysis and management	www.proteomecenter.org
PRISM	High-throughput proteomics information management system	http://ncrr.pnl.gov/prism
Scaffold	Protein identification automation software	www.proteomesoftware.com
Phenyx	Protein identification and validation platform	www.phenyx-ms.com
DBParser	Protein identification and validation platform	http://proteome.nih.gov
MZmine	Differential LC-MS analysis of metabolomics data	http://mzmine.sourcefourge.net
ProDB	Storage and analysis of identification proteomics experiments	http://www.cebitec.uni-bielefeld.de
PROTEIOS	Storage, analysis and annotation of proteomics experiments	www.proteios.org
ProteomIQ	Integrated proteomics data management platform	www.proteomesystems.com
Proteome Browser	Protein sequence annotation	http://genome.ucsc.edu
PepLine	Software pipeline for protein identification	http://www-helix.inrialpes.fr
Protein Expression System	Quantitative and qualitative proteomics analysis	www.waters.com
Xome	Quantitative and qualitative proteomics analysis	http://bio.mki.co.jp/product/xome
CellCarta	Integrated suite for quantitative proteomics analysis	www.caprion.com
mzXML	File format (standard) for mass spectra data	www.proteomecenter.org
Trans-Pproteomic Pipeline	XML-based analysis pipeline for proteomics data	www.proteomecenter.org
ProICAT	Protein quantitation and identification for ICAT	www.appliedbiosystems.com
ProQuant	Protein quantitation and for iTRAQ	www.appliedbiosystems.com
DeCyder MS	Identification and quantitation analysis platform	www.amershambiosciences.com
MS peaks	*De novo* protein identification	www.bioinformaticssolutions.com
Expasy proteomics server	Protein sequence analysis tools and databases	http://ca.expasy.org
Ion Source	On line resource of mass spectrometry methods	www.ionsource.com

storage and provenance of statistical results by storing the computed results of statistical analyses and the information necessary to document and recompute them, including the input data, information about the software used in the analysis, and various procedural choices and settings. Data provenance provides information about where a piece of information comes from and the process by which it arrived [59]. Given the complexity of the laboratory life cycle and the different systems involved, a solution to the data provenance problem is not easy, but it is absolutely essential if the data is to be trusted.

A common theme across the various stages of the proteomics life-cycle is that data need to be captured, organized and stored together with supporting data in databases that can be explored, aggregated and shared. Component applications interact with each other and/or with databases by importing and exporting data in formats that are not well aligned. Notable approaches to address these problems include the definition of common open representation of MS data [49,56], the development of laboratory-specific systems where the interactions between components are explicitly defined [12] and the deployment of a workflow-management solution that enables researchers to create, modify and run workflows that involve multiple applications.

An end-to-end informatics platform for high-throughput proteomics

Industrial and research laboratories invest resources and time in responding to the issues raised in this review. An informatics platform has been developed at MDS Proteomics to enable HTP analysis. The technological foundations of this platform include robust data management, rigorous application and data integration, workflow management and high-performance implementation of proteomics analysis algorithms. Table 1 outlines the components and the overall architecture of the platform. Each component corresponds to a stage of the data lifecycle. For each component, the supporting infrastructure, goals, data, processes and systems are summarized. It is important to point out the heterogeneity of processes and systems involved. The big picture view also gives a sense of the importance of issues beyond the pure computational aspects of proteomics, such as information tracking, workflow management and data integration.

An uncompromised goal in the development of the platform was that no data should be managed outside a database. To achieve this, multiple databases were combined to form a federated database system that could answer end-to-end process and data-specific queries. The database-driven architecture enhanced the comprehensiveness and performance of spectral data analysis. For each sample, the results of identification and absolute quantitation were stored in the database and linked using the mass–time coordinates of precursor ions. Incorporation of multiple identifications and additional scoring schemes was easy to integrate (Table 2).

Another goal of the platform was to achieve a high degree of integration between applications. One strategy was to use the database as a broker. Application A stores in the database, application B retrieves from the database; the database schema is the ‘contract’ of the interoperation. Because many applications operate on lists of proteins, a list of protein identifiers was the minimum contract required for interoperation. A second strategy was to connect different data-processing tasks together to form workflows. This strategy was pursued in two phases. In phase one, workflows were programmed directly, (i.e. without employing a formal workflow language and execution model). Phase two involved a commercial workflow-management environment. Selected proteomics analysis workflows were successfully migrated to the new environment, but only after they were prototyped in custom code and validated. The gain from this exercise was a more flexible easy-to-maintain system. The fact that a workflow modelling solution was not used in the prototyping might sound counter intuitive. The reason is that notations and tools for modelling workflows are not mature yet, especially in bioinformatics [60]. Recent work in bioinformatics workflows [61], service-oriented architectures for bioinformatics [62], process management [63] and tools support (www.alphaworks.ibm.com/tech/ii4bpel) will change this in the future.

Concluding remarks

A few important points have been highlighted in this review: (i) the importance of data and process management and integration; (ii) the need to incorporate validation steps that quantify the error rate of automated high-throughput analyses; and (iii) the breadth and heterogeneity of the proteomics data lifecycle. Reaching high-throughput is a multivariate problem, and successful informatics platforms for HTP must bring together high-performance algorithms, analysis automation, data quality factors and robust information management.

Acknowledgements

The author gratefully acknowledges his former colleagues at MDS Proteomics/Protana for sharing a unique experience of building a high-throughput informatics platform for proteomics.

References

1 Tyers, M. and Mann, M. (2003) From genomics to proteomics. *Nature* 422, 193–197

2 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207

3 Hanash, S. (2003) Disease proteomics. *Nature* 422, 226–232

4 Liotta, L.A. and Petricoin, E.F., 3rd (2003) The promise of proteomics. *Clin. Adv. Hematol. Oncol.* 1, 460–462

5 FDA, (2004) *Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*. U.S. Food and Drug Administration

6 Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567

7 Yates, J.R., 3rd *et al.* (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* 67, 1426–1436

8 Mann, M. *et al.* (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345

9 Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467

10 Han, D.K. *et al.* (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope- coded affinity tags and mass spectrometry. *Nat. Biotechnol.* 19, 946–951

11 MacCoss, M.J. *et al.* (2003) A Correlation Algorithm for the Automated Quantitative Analysis of Shothun Proteomics Data. *Anal. Chem.* 75, 6912–6921

12 Kiebel, G.R. *et al.* (2004) Proteomics Research Information Storage and Management (PRISM) System. *Pacific Northwest National Laboratory*

13 Keller, A. *et al.* (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Molecular Systems Biology* 10.1038/msb4100024 (www.nature.com/msb)

14 MacBeath, G. (2002) Protein microarrays and proteomics. *Nat. Genet.* 32 (Suppl.), 526–532

15 Pappin, D.J. *et al.* (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3, 327–332

16 Yates, J.R., 3rd *et al.* (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 214, 397–408

17 Nesvizhskii, A.I. *et al.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 75, 4646–4658

18 Bern, M. *et al.* (2004) Automatic quality assessment of Peptide tandem mass spectra. *Bioinformatics* 20 (Suppl. 1), I49–I54

19 Craig, R. *et al.* (2005) The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* 19, 1844–1850

20 Johnson, R.S. *et al.* (2005) Informatics for protein identification by mass spectrometry. *Methods* 35, 223–236

21 Fenyo, D. and Beavis, R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* 75, 768–774

22 Baldwin, M.A. (2004) Protein identification by mass spectrometry: issues to be considered. *Mol. Cell. Proteomics* 3, 1–9

23 Boutilier, K. *et al.* (2004) Comparison of different search engines using validated MS/MS test datasets. *Analytica. Chemica. Acta* 534, 11–20

24 Eriksson, J. and Fenyo, D. (2004) The statistical significance of protein identification results as a function of the number of protein sequences searched. *J. Proteome Res.* 3, 979–982

25 Cargile, B.J. *et al.* (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* 3, 1082–1085

26 Apweiler, R. *et al.* (2004) Protein sequence databases. *Curr. Opin. Chem. Biol.* 8, 76–80

27 Colinge, J. *et al.* (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* 4, 1977–1984

28 Yang, X. *et al.* (2004) DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* 3, 1002–1008

29 Le Bihan, T. *et al.* (2004) Definition and characterization of a ''trypsinosome'' from specific peptide characteristics by nano-HPLC-MS/MS and in silico analysis of complex protein mixtures. *J. Proteome Res.* 3, 1138–1148

30 Desiere, F. *et al.* (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 6, R9

31 Sadygov, R.G. *et al.* (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* 1, 195–202

32 Keller, A. *et al.* (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* 6, 207–212

33 Carr, S. *et al.* (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell Proteomics* 3531–3533

34 Gygi, S. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–998

35 Li, X.-J. *et al.* (2003) Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry. *Anal. Chem.* 75, 6648–6657

36 Silva, J.C. *et al.* (2005) Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* 77, 2187–2200

37 Horn, D.M. *et al.* (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* 11, 320–322

38 Kearney, P. and Thibault, P. (2003) Bioinformatics meets proteomics–bridging the gap between mass spectrometry data analysis and cell biology. *J. Bioinform. Comput. Biol.* 1, 183–200

39 Dharsee, M. *et al.* (2005) Automated quantitation and comparative proteomic analysis of complex biological samples. ASMS Abstract#2701

40 Katajamaa, M. and Oresic, M. (2005) Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* 6, 179

41 Anderle, M. *et al.* (2004) Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* 20, 3575–3582

42 Torgrip, R. *et al.* (2003) Peak alignment using reduced set mapping. *J. Chemometrics* 17, 573–582

43 Wong, J.W. *et al.* (2005) SpecAlign–processing and alignment of mass spectra datasets. *Bioinformatics* 21, 2088–2090

44 Zhang, X. *et al.* (2005) Data pre-processing in liquid chromatography-mass spectrometry based proteomics. *Bioinformatics* 21, 4054–4059

45 Wang, W. *et al.* (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* 75, 4818–4826

46 Listgarten, J. and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics* 4, 419–434

47 NatureGenetics, (2002) Chipping Forecast II. *Nat. Genet.* 32, 4s

48 Smith, R.D. *et al.* (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* 2, 513–523

49 Pedrioli, P.G. *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* 22, 1459–1466

50 Boguski, M.S. and McIntosh, M.W. (2003) Biomedical informatics for proteomics. *Nature* 422, 233–237

51 Figeys, D. (2005) *Industrial Proteomics: Applications for Biotechnology and Pharmaceuticals*. Wiley

52 Kristensen, D.B. *et al.* (2004) Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* 3, 1023–1038

53 Ludaescher, B. and Goble, C. (2005) Guest editors' introduction to the special section on scientific workflows. *SIGMOD Rec.* 34, 3–4

54 Juric, M. *et al.* (2004) *Business Process Execution Language for Web Services: BPEL and BPEL4WS*. Packt Publishing

55 Stevens, R. *et al.* (2004) myGrid and the drug discovery process. *Drug Discov. Today: BIOSILICO* 2, 140–148

56 Orchard, S. *et al.* (2004) Common interchange standards for proteomics data: Public availability of tools and schema. *Proteomics* 4, 490–491

57 Shah, S.P. *et al.* (2005) Atlas – a data warehouse for integrative bioinformatics. *BMC Bioinformatics* 6, 34

58 Hsu, F. *et al.* (2005) The UCSC Proteome Browser. *Nucleic Acids Res.* 33, D454–D458 (Database issue)

59 Buneman, P. *et al.* (2001) Why and where: a characterization of data provenance. In *International Conference on Database Theory (ICDT). Lecture Notes in Computer Science*. 1973. 316–330

60 Addis, M. *et al.* (2003) Experiences with eScience workflow specification and enactment in bioinformatics. In *Proc. UK e-Science All Hands Meeting*. pp. 459–466 (www.nesc.ac.uk/events/ahm2003/AHMCD)

61 Oinn, T. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045–3054

62 Lord, P. *et al.* (2004) Applying Semantic Web Services to Bioinformatics: Experiences Gained, Lessons Learnt. In *Proc. of the 3rd Internation Semantic Web Conference. Lecture Notes in Computer Science*. 2004. 3298, 350–364

63 Leymann, F. *et al.* (2002) Web services and business process management. *IBM Systems Journal* 41, 198–211