Publishing large proteome datasets: scientific policy meets emerging technologies

William S. Hancock, Shiaw L. Wu, Robert R. Stanley and Erich A. Gombocz

Currently, there are various approaches to proteomic analyses based on either 2D gel or HPLC separation platforms, generating data of different formats, structures and types. Identification of these separated proteins or peptide fragments is typically achieved by mass spectrometry (MS) measurements that use either accurate mass measurements or fragmentation (MS–MS) information. Integrating the information generated from these different platforms is essential if proteomics is to succeed. A further challenge lies in generating standards that can accept the hundreds-of-thousands of mass spectra produced per analysis based on threshold or probability measurements. Finally, peer review and electronic publication processes will be crucial to the dissemination and use of proteomic information. Merging the policy requirements of data-intensive research with information technology will enable scientists to gain real value from global proteomics information.

A dataset is of limited use unless it can be integrated dynamically with the diverse knowledge surrounding disease, genomics and proteins. Both the scientific and technological communities must respond to the anticipated flood of complex, related data, presented in different structures and formats and defined by different (but overlapping) data ontologies. This review discusses such challenges in terms of documenting achievements in the field in a permanently archived and reviewable manner, and the emerging computational challenges faced by proteomics researchers using 2D gel–matrix-assisted laser desorption–ionization (MALDI) time-of-flight (TOF), and liquid chromatography (LC)–mass spectrometry (MS)–MS approaches.

2D gel–MALDI-TOF

Two-dimensional gels separate proteins based on the charge [i.e. isoelectric focusing (IEF)] and molecular weight (SDS-PAGE) of the protein, and are well-established for the visualization of up to 2000 components in a given sample [1]. The identity of a protein(s) in a given spot is usually determined by an approach such as MALDI-TOF meaning that a database of such a study would need to contain, at the very least, information such as spot location, intensity parameters and peptide and/or protein identifiers. Increasing the resolving power of this approach can be achieved by using sample pre-fractionation steps, for example, narrow-range IEF strips and by subcellular fractionation and/or abundant protein removal of the sample. However, the number and complexity of sample aliquots generated in a typical study remains a challenge for proteomics researchers (e.g. one sample might require three subcellular fractionations, which, in turn, might generate 2000 spots per fraction, multiple MALDI-TOF measurements, and ESI ionization coupled to various MS platforms for each spot). Informatics systems designed specifically for such applications are now available, such as the BioinformatIQ™ software (Proteome Systems Ltd; http://www.proteomesystems.com/product/profile.asp?Category=Bioinformatics) combined with a DB2 database from IBM (http://www-3.ibm.com/solutions/lifesciences/).

LC–MS–MS

In shotgun sequencing, the protein mixture is digested with a protease(s) (e.g. trypsin), and the resulting peptides are separated by HPLC (ion exchange and/or reversed phase) [2]. The peptide is then further characterized by comparison with additional online MS measurements where the peptide mass spectra are related to either genomic or proteomic databases. It is therefore possible to generate a complex set of related information, either in the form of a report that references related files, or as metadata (structured 'data about data') connected to each sample file. Thus, shotgun sequencing enables the number and complexity of aliquots to be minimized in comparison to 2D gels, but this is balanced by the lower degree of sequence coverage that is typically achieved with the approach. An additional informatics challenge is to determine whether the peptide data provide a unique protein identifier because typical analyses might only identify a single peptide or very few peptides.

William S. Hancock*
Shiaw L. Wu
ThermoFinnigan
355 River Oaks Parkway,
San Jose, CA 95145, USA.
*e-mail: WHancock@ThermoFinnigan.com

Robert R. Stanley
Erich A. Gombocz
Biosentients,
1325 61st Street,
Emeryville, CA 94608-2117, USA.

0167-7799/02/$ – see front matter PII: S0167-7799(02)02093-0

Two common approaches to identifying a given peptide by mass spectrometric measurements use either accurate mass measurement, or fragmentation (MS–MS) information. As the quality of the protein identification relies on this measurement, one issue in publishing such experiments is the perceived need to store this information in a database that permits reviewer scrutiny. For example, computerized systems used in clinical trials must meet FDA requirements and the data files and metadata must be attributable by persistent supporting information, that is, who acquired the data and the conditions under which data was acquired. Ideally, metadata and annotation should refer to data that are linked, validated, original and relevant.

These conditions require re-evaluation of both the structure of the data and the methods for storing it because the requirements for information-linking across data structures, formats, types and locations can be complex, and individual data files can be extensive for some types of MS spectra (a single MS analysis might produce thousands of spectra). Technology also needs to be developed that can enable peer-review of the criteria used to validate data and metadata, such as thresholds used to delineate good spectra from bad, and how results have been related to approved probability measurements.

Quantitation

Quantitation is important in the comparison of different samples, and 2D gels typically use estimates of the amount of protein per gel spot, which requires sufficient resolution or pre-fractionation to avoid co-migrations of proteins. The shotgun sequencing approach typically uses an isotopically labeled tag that is quantitated in the mass spectrometer [3]. In both cases, the method, performance parameters and recording consistency of measurements (either within or between samples) should be recorded in the database, in addition to the quantitative information.

There is also an emerging focus on post-translational modifications, such as phosphorylation and glycosylation, which generate another layer of data complexity. Much of this supporting information could be stored as abstracted metadata, but it is preferable to keep the supporting attribute information sources directly linked to the (entity) data for validation purposes and for future research. Innovative technologies need to be developed that can enable dynamic re-evaluation of data inter-relationships based on direct linking of non-interpreted data, rather than on abstracted metadata alone. For example, data entities in one research scenario might become supporting data attributes in another.

In summary, although there is the need for disciplined, reviewable structuring of data and their attributes and relationships, data ontologies must also be flexible and, ideally, should keep the data proprietary until publication.

Other parameters

Information generated for each parameter of a given proteomic measurement should be integrated into a master database, for example, the biological source, including clinical records and genomic information, sample preparation protocols, protein structure, results of homology searches, protein interaction information, cellular locations, and known protein function, including role in diseases. As the volume of information in a laboratory increases, so does the need for intelligent and automated searching and integrating of all relevant information into the experimental dataset.

Such information should be updated on a regular, if not daily, basis and raises the challenge of extracting information from a variety of file formats, different database structures and different instrument platforms. An immediate need is the integration of information from a 2D gel with that of an LC–MS–MS study, and to establish a correlation between the two studies. The two approaches are generally complementary in the lists of proteins identified [4]. For publication purposes, the reviewers must have access to all significant data, but typical space limitations of today's journals means that many of the results are published as supplementary information. A further concern is the functionality and preservation of archived information together with maintenance of the informatics tools used in the original study. Furthermore, it is not clear who will pay for the cost of archiving such information. It would be of great concern if the large public expenditures on proteomic studies were lost in the future owing to the inability to protect these enormous datasets. Finally, flexibility and proprietary requirements for advanced and evolving research must be established alongside the new technologies that are being developed to enable well-organized, integratable data structures.

Informatics challenges

With the advent of both genomics, proteomics and, more recently, 'metabolomics', the isolated nature of the laboratory is being replaced by a high degree of global interconnection. In this context, the selection of an appropriate bioinformatics and hardware strategy is becoming a significant part of discovery science, and issues such as cost, training, functionality, interconnectivity and speed will have a bearing on such decisions.

One of the major difficulties with this higher degree of global interconnection is that software approaches and data definitions are often proprietary. This is typically for valid reasons, for example, when a new idea or unproven data inter-relationships are being explored. Even openly published information technology approaches and data definitions will probably conflict for a variety of persistent reasons. It is therefore difficult to promote a broadly based discussion of the needs and possible solutions.

The next section of this review focuses on the strategies that are being developed to meet these needs, and gives examples of the technical challenges and types of solutions that are emerging.

Barriers to integrating proteomics data

Structuring data accurately for integration and the publication of proteomics and related biological databases pose a significant challenge to information technology. These include: (1) the disparate nature of related data; (2) programming and processing overheads associated with analyzing massive amounts of heterogeneous data; (3) the difficulty in analyzing relationships between heterogeneous datapoints without common annotation; (4) conflicting standards for integration; and (5) the importance of maintaining flexible and proprietary data definitions.

Historically, software solutions have been expensive, time consuming and unable to provide effective integration [5]. Current methods for integrating proteomics databases include: software 'wrapping' (IBM Life Sciences); 'open' or 'federated' data standards (Rosetta Inpharmatics; http://www.rosettabio.com/home.html; LabBook; http://www.labbook.com; The Open Bioinformatics Foundation; http://www.open-bio.org); and newer 'intelligent data' methods (Biosentients; http://www.biosentients.com). The following section focuses on conventional and emerging technologies in proteomics data integration, outlining the associated problems and highlighting examples of recent improvements.

Middleware and software wrappers

Lion Bioscience (http://www.lionbioscience.com) and IBM have each utilized Common Object Request Broker (CORBA), Java [e.g. Enterprise Java Beans (EJB)] and object-oriented middleware as core technology to develop their respective data integration solutions (Lion, SRS Platform; IBM, DiscoveryLink solutions) based on software 'wrapping'. In simple terms, 'wrappers' are software layers that translate between middleware and backend (e.g. databases and application servers) interfaces. Pioneered in the early 1990s, this interface integration method enables a high level of interoperability [6]. However, with the emergence of very large datasets, one can anticipate crucial shortcomings, such as high programming and processing overheads, interface protocol conflicts, limited integration with the diverse world of biology, and problems with proprietary application programming interfaces (APIs).

Existing challenges to wrapper-based integration include problems with programming, compiling, updating and, perhaps more importantly, considerable processing overheads. Solutions to such problems often demand clustered servers or mainframes to address the high volumes of data [7]. However, even with such a massive amount of hardware, time delays in the order of hours to days can be expected when comparing inter-laboratory global studies [8].

Conflicting protocols are also a problem; middleware must comply with diverse interface conventions or exchange APIs for translation. Interoperability requires adherence to or wrapping between changing interface and invocation protocols, including Interface Definition Language (IDL), Internet Inter-ORB Protocol (IIOP), Java and/or CORBA Remote Methods Invocation linked to JAVA and/or CORBA (JAVA/RMI, CORBA/RMI-IIOP), Open Database Connectivity (ODBC) and Distributed Component Object Model (DCOM) [9,10]. These barriers result in wrapper definitions of limited value, which are useful only for limited time periods and thereby result in decreased performance. Importantly, the dynamic programming-free definition of application and data inter-relationships required to meet changing analytical needs is, at present, restricted.

Software standards for proteomics data integration

Groups such as Rosetta Inpharmatics and LabBook have developed 'open' data annotation for integration. Annotation standards (e.g. ontologies and taxonomies) are commonly written in markup languages such as XML. In the same way as wrapping, these methods provide integration but also have shortcomings that could become crucial factors as the proteomics challenge develops [11]. For example, open annotation standards can conflict, federated standards can restrict innovation, annotation markup increases processing overheads, data interdependencies are missed, and data annotation is exposed, which limits security in sensitive applications such as medical research.

Two of the many open bioinformatics annotation standards are Gene Expression Markup Language (GEML) from Rosetta Inpharmatics and Bioinformatic Sequence Markup Language (BSML; http://www.bsml.org) used by LabBook. Other open or federated data standards include Array Extensible Markup Language (AXML; http://www.mged.org), Biopolymer Markup Language (BioML; http://www.bioml.com), Biological Distribution Annotation Sequence (BioDAS; http://www.biodas.org), Gene Ontology Markup Language (GO; http://www.geneontology.org), MicroArray Markup Language (MAML; http://www.omg.org) and Systems Biology Markup Language (SBML; http://www.cds.caltech.edu/erato). Conflicting annotation standards are proposed for several valid reasons [12]. Unfortunately, data described according to different standards are usually incomparable for a researcher attempting to enrich a given analysis. This leads to problems for those scientists interested in analyzing data rather than arbitrating between standards.

Furthermore, dependence on open data standards limits security and propriety. Exposure of annotation limits the researcher's ability to develop unique data views, and adherence to rigid federated standards limits analytical flexibility. If an annotation defining data inter-relationships has not been previously inferred by a standard that defines that group, the analysis will not be supported by the software. In short, exposure of data annotation for compliance to rigid standards restricts a major intellectual property arena: the definition of proteomics data within its relationships.

Software overhead

XML adds overhead to each data file and, in many cases, the XML 'header' is larger than the data content. High-volume analysis results in poor performance even with massive computing hardware. This is a widely recognized problem, leading Lincoln Stein from Cold Spring Harbor (http://www.cshl.org/) to state that 'XML is the worst possible solution [to integration of distributed data], except for all the other solutions' [13].

Generating highly functional large datasets

To a large extent, proteomics is a child of uncertain parentage in that it has inherited informatics platforms from fields such as biology, genomics and protein chemistry without considering the needs of the new field. This leaves middleware 'enterprise solutions' with the unenviable task of patching or 'wrapping' together the varying platforms. Coupled with new computing hardware expenses, many conventional integration solutions are too expensive for many of the smaller laboratories. However, 'intelligent data' approaches are now emerging based on object-oriented programming that are designed to address the complex proteomics challenge more efficiently. The following section outlines examples of this new approach.

Emerging requirements and intelligent software objects

Emerging requirements of proteomics software include: (1) improving performance for high volume processing of heterogeneous data; (2) integrating protocols and standards without exposure of proprietary APIs or annotation; and (3) providing dynamic definition and linking of multidimensional data attributes to enable a flexible and meaningful analysis. An 'intelligent data' approach has appeared recently that attempts to meet these emerging requirements and advance the multidimensional analysis of heterogeneous proteomics data.

A solution to one of the requirements of data-enabled processing is to develop lightweight (~5K) data structures that refer to distributed data uniformly from either a given laboratory or over the Internet. Such methods need to include: (1) vectorized data access and direct weighted linking of normalized content subsets; (2) persistent data level agency ('active listening') and state management; and (3) flexible content-attribute ontologies enabled by multidimensional 'property panes'.

Unified interfaces to intelligent global datasets

Another requirement is a unified interface to distributed applications and data to enable integrated analysis across heterogeneous resources. This should also be capable of simultaneous deployment across different network connections, such as client-server, server-server, peer-to-peer, grid, and other 'any-to-any' combinations.

However, distributed data access when taken alone is of little value to the individual proteomics researcher because of the time required for intelligent integration of data in a given proteomic study. Access to global data, therefore, must be linked according to user-definable multidimensional relationships, with learning algorithms provided by an appropriate neural network functionality [14]. Such algorithms must support clustering, self-organizing maps, learning algorithms, and additional data-mining functions.

Speed and security

The two main challenges in proteomics research are the need for security and increased speed. Increased processing would enable real-time processing of LC–MS–MS data to be developed. For example, MS–MS measurements could be targeted to new peptide spectra or concentration changes without re-sequencing known peptides. Security is another concern in proteomics research; clinical studies or pharmaceutical discovery must preserve patient confidentiality yet still enable the sharing of data subsets with different collaborators.

Vectorized addressing makes it possible to achieve fine-grained data access. Vector subset workspaces within data can be defined to address heterogeneous data content to as fine as single-byte granularity. In addition, weighted linking of vectorized workspace subsets enable data-enabled parallel processing. These methods minimize processing and network transport and enable multiple-user analysis without data locking. Such an approach can result in impressive performance benchmarks, even within a small laboratory format [15].

Examples of data integration

As described previously, conflicting annotation and interface standards can hinder the extraction and integration of data from different analytical platforms. Object state and translation engines can overcome this problem by enabling meaningful data extraction from disparate sources (Fig. 1). Redundant or conflicting annotations can

be ranked according to source validation, and annotation content can be given an alias by unifying ontological and semantic information provided in content-attribute lookup tables [12]. Such an approach has the potential of providing the 'home run' for proteomics researchers: training intelligent software based on the integration of trusted data annotations, without programming or intensive input from the user.

Another goal of such methods would be to create optimized global standards within local content-attribute tables. Integration of global data annotation for research should not require exposure of local experimental data annotation to external parties until the formal review process is initiated.

Intelligent data annotation should also be capable of dynamically (and locally) describing all functional relationships and analytical linkages in a given experiment using distributed data. Curated local annotation must be exportable in XML or in other open and regulatory compliant formats thereby enabling regulatory or peer review, and facilitating publication.

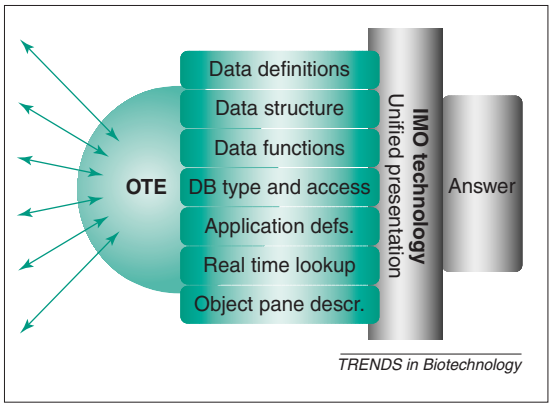

Figure 1. Integration of object state and translation for extracting information from disparate sources

This software was produced by Biosentients (http://www.biosentients.com) and contains an Object Translation Engine (OTE) with nested vector lookup tables (e.g. data definitions, structure, function, type and access definitions). This setup is designed to provide fast data integration to the unified data structure (provided by Intelligent Molecular Objects™ in the architecture shown here).

Figure 2 gives an example of where property panes from such an analysis can provide unified presentation layers for flexible, validated analysis. Property panes define multidimensional data properties, providing access to unique electronic identification, data state history, metadata indices, processing pipelines, graphical analysis, applications assembly, tagged annotation, and raw data matrices. Such an analysis can bypass distributed database interfaces

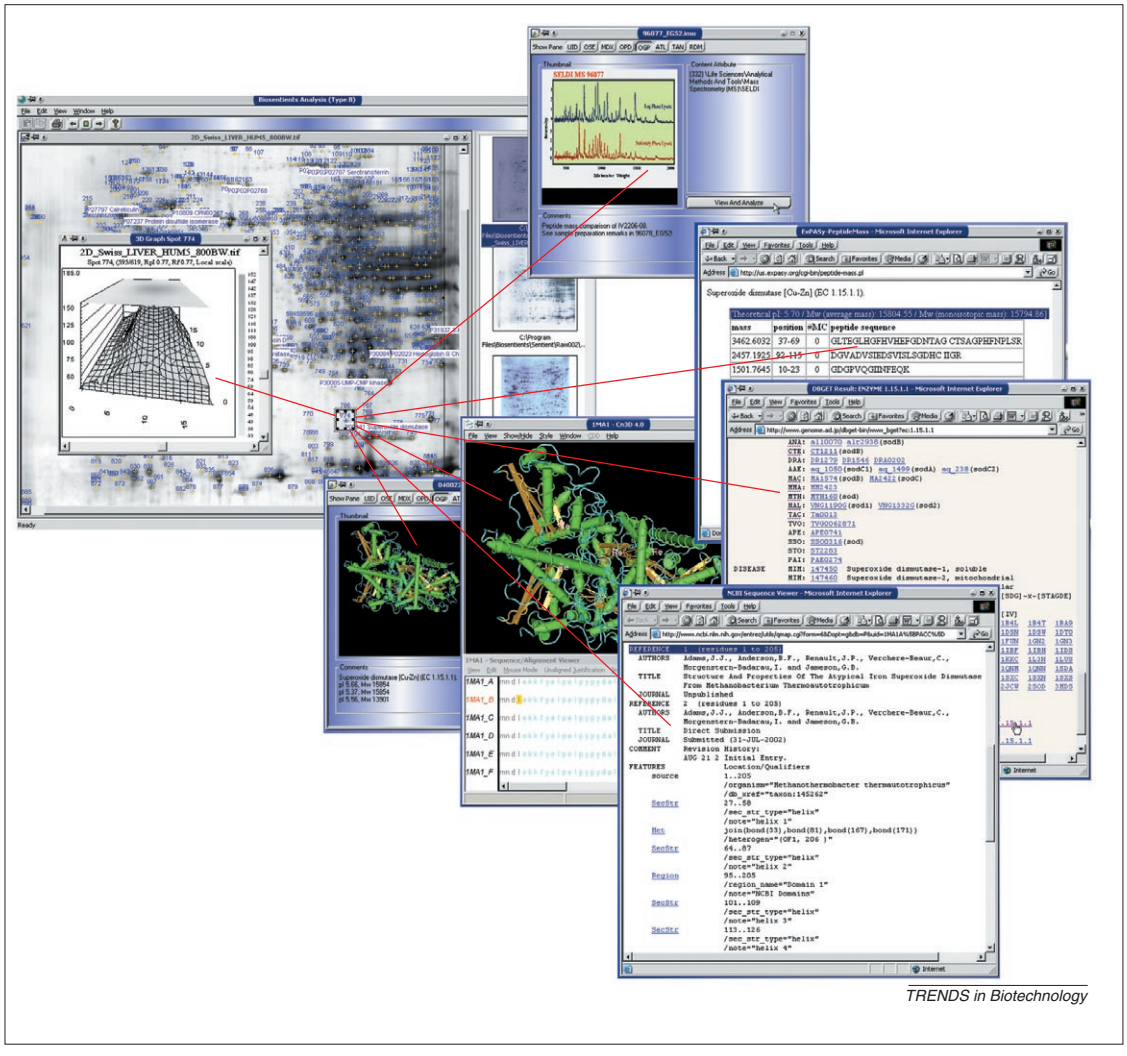

Figure 2. Direct integration of multidimensional structure and functional dependencies across data types and standards

2D gel electrophoresis (2DE) data for protein separation are shown on the left with (clockwise from top) directly related mass spectrometry fragmentation, sequence, metabolic function, literature and structure data. On the left, an analysis window and workspace opened from within Intelligent Molecular Objects™ (IMO™) (Biosentients; http://www.biosentients.com) (see Fig. 1) enables analysis of the 2DE protein data. Mass spectrometry data, see top, are shown within the graphic preview property pane, Object Graphics Preview Pane (OGP) of IMO™. Other IMO™ property pane toggles are shown within the IMO, at the top, for unique identification of the data (UID), persistent data state history (OSE), metadata indices (MDX), processing history and description (OPD), applications linking (ATL), interactive text annotation (TAN) and raw data matrix description (RDM).

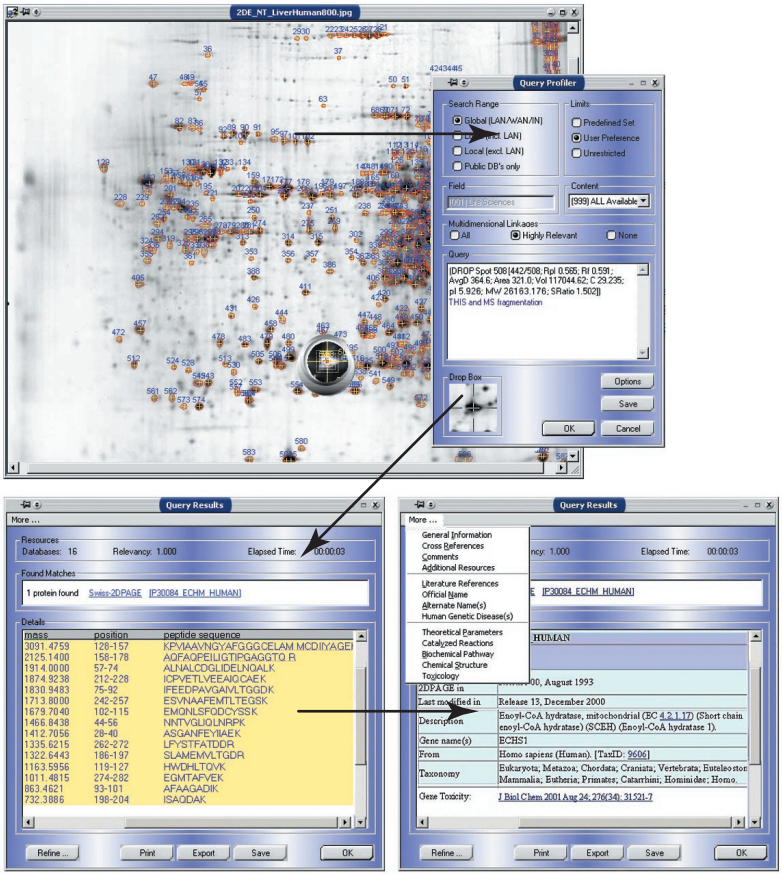

Figure 3. The Sentient IT Platform™

Shown here is the Sentient IT Platform™ (Biosentients; http://www.biosentients.com) query profiler, which integrates several distributed proteomics databases within one unified 'drag and drop plus plain English' query. The researcher clicks on the spot indicating a particular protein represented within a 2D-gel electrophoresis image. This automatically enters all known physical properties of the selected peptide to characterize the peptide for the search. Next, the researcher enters 'THIS and MS fragmentation' to investigate the mass spectrometry (MS) information relevant to that particular peptide. A unified, validated report is generated in seconds, with information ranked according to validation criteria. In this case, the MS fragmentation information is shown. Additional supporting information can be displayed based on user-defined selection criteria, which can be automatically linked to the original IMO data object.

and data headers, enabling locally defined multidimensional analysis without missing interdependencies.

Proteomics applications and benchmark results

Researchers and decision-makers in biotechnology traditionally are not experts in computer science and software. Informatics tools should therefore be intuitively useful with minimal steps, and results should be immediate and easily visualized.

A useful informatics platform must enable intuitive real-time decision-making based on valid information integrated from multiple sources. Ideally, multiple data sources and data types from multiple databases should be integrated within one 'plain English' query interface (Fig. 3). This interface should support formally structured as well as unstructured queries, and should be capable of analyzing multiple data qualities and relationships in real-time and across databases and data types.

Unified 'natural language' query profilers capable of supporting structured queries (Fig. 3) will improve the ability of proteomics researchers and decision-makers to extract knowledge from global data efficiently and in near real-time. The application of vector workspace technology for data input–output has been shown to improve the performance of data search, access and analysis by factors benchmarked up to 200×, compared with leading bioinformatics software [15]. Importantly, vector subset methods for accessing large high-functionality datasets continue to provide relative performance improvement as the volume and complexity of data for a given analytical task increases. These types of emerging technology promise more useful and intelligent access and analysis of heterogeneous, distributed proteomics data in the context of related genomics and metabolomics data.

Concluding remarks

One often hears the complaint at the major proteomics meetings that there is little value in publishing large datasets. However, it is not that the datasets have no value, but rather the data is of limited use without intelligent annotation that can exist in a functional form once published. In addition, the field needs improved visualization tools to enable the individual researcher to extract information from such a study. Such needs are magnified when one considers the challenges of publication, where the harried reviewer is confronted with, for example, a dataset of 100 000 proteins in a dynamic study involving 100 samples, and the journal is attempting to published well-reviewed and timely research on a limited budget.

References

1 Patterson, S.D. and Aebersold, R.A. (1995) Mass spectrometric approaches for the identification of gel separated proteins. *Electrophoresis* 16, 1791–1814
2 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682
3 Gygi, S.P. *et al.* (2001) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
4 Koller, A. *et al.* (2002) Comparative analysis of protein identification results using two different techniques, 2D gel Electrophoresis and MUDPIT. *Proceedings of the 5th Siena Meeting From Genome to Proteome: Functional Proteomics* (2–5 September, Siena, Italy) Abstract 374
5 Watkins, K.J. (2001) Bioinformatics. *Chem. Eng. News* 79, 29–45
6 Goble, C.A. *et al.* (2001) Transparent access to multiple bioinformatics information sources. *IBM Systems Journal* 40, 532–551
7 Head-Gordon, T. *et al.* (2001) Computational challenges in structural and functional genomics. *IBM Systems Journal* 40, 265–296
8 Stanley, R.A. *et al.* (2002) Intelligent molecular objects: a new paradigm in bioinformatics. *Am. Genomic/Proteomic Technol.* 2, 2–27
9 Object Management Group (2001) OMG IDL syntax and semantics. Common Object Request Broker Architecture (CORBA). 2.6, 1–3.60, Object Management Group (Needham, MA, USA)
10 Wang, L. *et al.* (2000) Accessing and distributing EMBL data using CORBA (Common Object Request Broker Architecture). *Genome Biol.*, 1, 0010.1-0010.10
11 Lehr, W. (1995) Compatibility standards and interoperability: lessons from the Internet. In *Standards Policy for Information Infrastructure* (Kahin, B. *et al.*, eds), pp. 121–147, Harvard University Press
12 Gardner, S. (2002) Ontologies, taxonomies, and other buzz words. *Am. Genomic/Proteomic Technol.* 2, 10–12
13 Stewart, B. (2002) *Bioinformatics conference – 4 Day Wrap Up*, O'Reilly & Associates Online Publications (http://www.oreillynet.com/pub/wlg/1109)
14 Gallant, S. *et al.* (1993) *Neural Network Learning*, MIT Press
15 Gombocz, E. A. *et al.* (2001) Intelligent molecular objects: integrated, data-enabled, global, multidimensional real-time analysis in BI and CI. 2001 *Annual Meeting AES/AIChE*, Nov 4–9 2001 (San Reno, NV, USA). Abstract 108e