REGULAR ARTICLE

PRIDE: The proteomics identifications database

Lennart Martens[1], Henning Hermjakob[2], Philip Jones[2], Marcin Adamski[3], Chris Taylor[2], David States[3], Kris Gevaert[1], Joël Vandekerckhove[1] and Rolf Apweiler[2]

[1] Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
[2] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3] Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

The advent of high-throughput proteomics has enabled the identification of ever increasing numbers of proteins. Correspondingly, the number of publications centered on these protein identifications has increased dramatically. With the first results of the HUPO Plasma Proteome Project being analyzed and many other large-scale proteomics projects about to disseminate their data, this trend is not likely to flatten out any time soon. However, the publication mechanism of these identified proteins has lagged behind in technical terms. Often very long lists of identifications are either published directly with the article, resulting in both a voluminous and rather tedious read, or are included on the publisher's website as supplementary information. In either case, these lists are typically only provided as portable document format documents with a custom-made layout, making it practically impossible for computer programs to interpret them, let alone efficiently query them. Here we propose the proteomics identifications (PRIDE) database (http://www.ebi.ac.uk/pride) as a means to finally turn publicly available data into publicly accessible data. PRIDE offers a web-based query interface, a user-friendly data upload facility, and a documented application programming interface for direct computational access. The complete PRIDE database, source code, data, and support tools are freely available for web access or download and local installation.

Received: October 11, 2004
Revised: January 25, 2005
Accepted: March 1, 2005

Keywords:
Bioinformatics / Databases / Protein identification

Correspondence: Dr. Lennart Martens, Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium
E-mail: lennart.martens@UGent.be
Fax: +32-9264-9484

Abbreviations: GPS, general proteomics standards; **PDF**, portable document format; **PPP**, Plasma Proteome Project; **PRIDE**, proteomics identifications database; **PSI**, proteomics standards initiative; **RDBMS**, relational database management system; **SQL**, structured query language; **UM**, University of Michigan; **W3C**, WWW Consortium; **XML**, extensible markup language; **XSL**, XML stylesheet language

1 Introduction

The field of proteomics has rapidly grown into one of the most active research areas in life sciences today. This growth is largely attributable to the availability of ever increasing amounts of gene and protein sequence information and the many technical improvements in the elaborate machinery used to identify proteins in complex mixtures, along with many novel techniques that reduce the complexities of analyte mixtures, allowing protein identification and characterization at an ever increasing

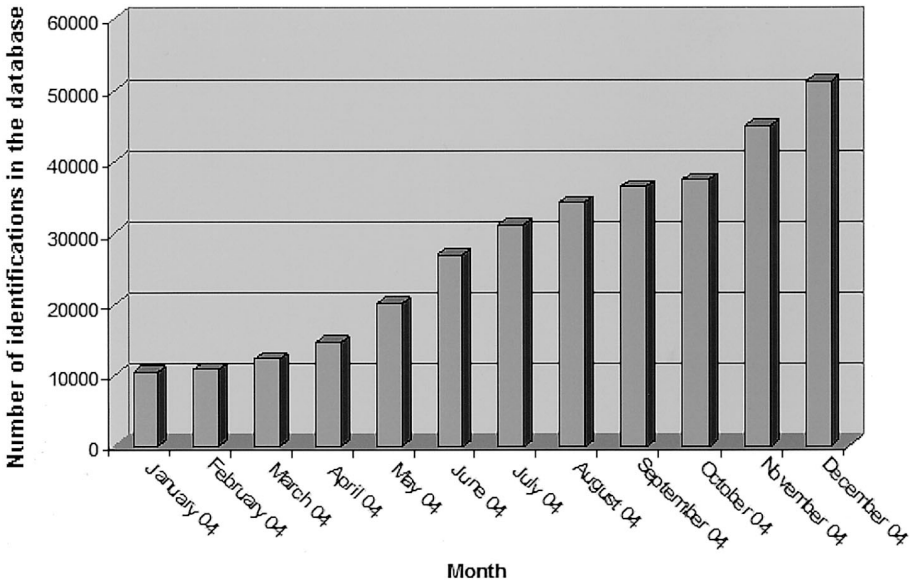

Figure 1. Illustration of the increase in the number of protein identifications over time. Number of identifications in the local database of the Department of Biochemistry at the University of Ghent throughout 2004 is shown. Source data originate from fragmentation spectra obtained from three different mass spectrometers: ESI-Q-TOF, ESI-IT, and MALDI-TOF/TOF. Note that the average increase since January 2004 amounts to 4000 identifications *per* month, often punctuated by particularly sharp increases when all three machines are fully operational in parallel. The vast majority of the identifications stems from experiments using the COFRADIC gel-free technology [2–4].

pace, as reviewed in [1]. To illustrate this further, the actual identification rate for a typical proteomics laboratory is given in Fig. 1.

Correspondingly, publication of protein identification data has been steadily on the rise over the past few years. The number of hits returned *per* year since 2000 for a PubMed query illustrates this in a simplistic, yet straightforward manner (Fig. 2). Together with the growing number of publications, the lists of identifications have grown considerably in size as well. Since these listings can easily contain thousands of peptides or hundreds of proteins, they are often published as supplementary information. In almost all cases this supplementary information consists of one or more portable document format (PDF) files detailing the identifications in a tabular format.

Yet even though PDF does an admirable job as a truly portable format and is therefore a natural choice for publishers, it is definitely not designed to convey structured information. Tables in PDF are notoriously difficult to extract and this problem is further exacerbated by the fact that nearly every author uses a different formatting for these tables.

Since this relative inaccessibility of proteomics data presents a considerable stumbling block on the way to making all these identifications really count for life sciences, the construction of a centralized, freely accessible repository for proteomics data is one of the primary requirements in proteomics today [5, 6]. In a single sentence: publicly *available* data needs to become publicly *accessible* data.

The pilot phase of the Plasma Proteome Project (PPP) [7], the first of the HUPO proteomics projects [8] to reach an important milestone, has been invaluable in achieving the ambitious goal of designing and implementing such a repository [9].

Indeed, the need for a centralized data repository was quickly realized during the initial planning phase for the PPP and this resulted in both a short-term and a long-term

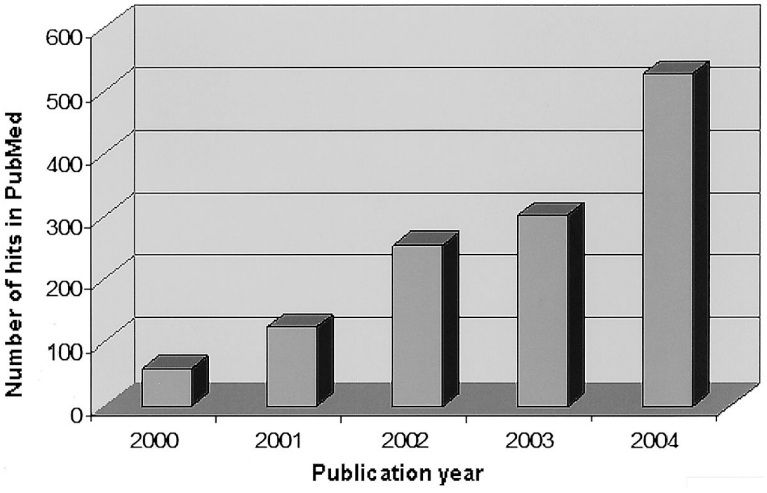

Figure 2. Illustration of the increase in protein identification papers over the past few years. Number of PubMed hits for each year since 2000 for the query "*proteom* AND proteins AND identified AND mass spectrometry*" are shown. Although this query is by no means exhaustive, it provides a meaningful sampling of the available literature.

approach to solving the data management problems. The short-term solution dealt with the immediate need for data storage and consisted of a relational database implemented using Microsoft Structured Query Language Server (MS-SQL). This database was constructed and continuously updated by Marcin Adamski from the core bioinformatics unit of David States at the University of Michigan (UM), Ann Arbor. This database actually had a two-fold objective: first of all, it served the vital purpose of centralizing the data produced in the PPP collaboration as it started to trickle (and later pour) in from the different labs, and second, it served as a test-bed for the construction of a centralized, project-independent database for protein identifications at the European Bioinformatics Institute (EBI). The aims of proteomics identifications database project are three-fold: developing an open source, publicly available set of tools to aid developers in implementing ms are three-fold: (1) providing a central repository for protein identification data, (2) building an efficient web-based interface for queries and data submission, and (3) developing an open source, publicity available set of tools to aid developers in inplementing custom analysis tools.

The MS-SQL database constructed at UM proved to be an excellent source of inspiration for the PRIDE data model as it had been refined throughout the PPP in order to contain detailed proteomics data from many different collaborating laboratories across the globe. The design of PRIDE and the functionality of its web interface will be discussed next, along with future prospects for the data model as proteomics standards evolve.

2 Materials and methods

The PRIDE project was completely developed in the Java 2 programming language (Sun Microsystems) using the Java™ Development Kit (JDK) 1.4 from Sun Microsystems (http://www.java.com/en/download/manual.jsp) as well as the Java™ 2 Enterprise Edition (J2EE) extensions from Sun Microsystems (http://java.sun.com/j2ee/index.jsp) for the web development.

PRIDE makes use of many open source software tools, components, and libraries. Object-relational bridge (OJB) (http://db.apache.org/ojb) takes care of the declarative object-relational mapping, Tomcat (http://jakarta.apache.org/tomcat) functions as web server and servlet engine, Log4J (http://logging.apache.org/log4j) as the logging framework, and Maven (http://maven.apache.org) as the project management tool. All of the above were obtained from the Apache Software Foundation (http://www.apache.org). Extensible markup language (XML) parsing and writing relies on the XML pull parser (XPP) libraries (http://www.extreme.indiana.edu/xgws/xsoap/xpp). Unit testing was performed using the JUnit framework (http://www.junit.org). During development, the relational database management system (RDBMS) employed was MySQL (http://www.mysql.com) and for the final prototyping and production version PRIDE was ported to Oracle (http://www.oracle.com).

3 Results and discussion

3.1 PRIDE as a set of components

The PRIDE project consists of a number of distinct parts, which are summarized in Fig. 3. The XML format represents the basic data structure, whereas the relational database implementation is just one of the possible renderings of the hierarchical XML format in a relational schema. The PRIDE core libraries contain an object model of the PRIDE data structure and allow the programmer to interact seamlessly and effortlessly with the PRIDE XML format and reference database implementation. The PRIDE web libraries provide a web-based view on an underlying reference database and use the PRIDE core libraries for data access. Query results from the web can be sent in PRIDE XML format or in HTML after XML stylesheet language (XSL) transformation of the XML.

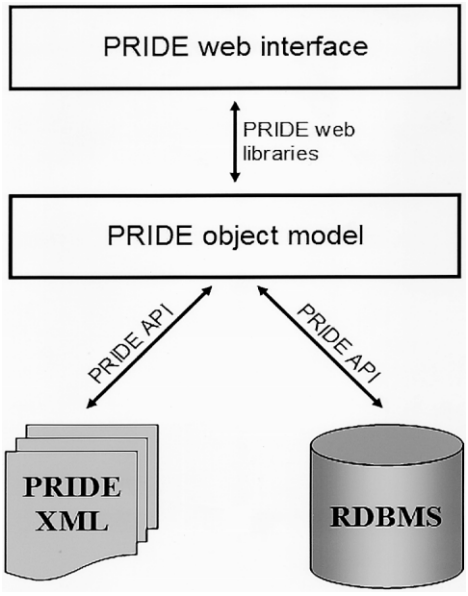

Figure 3. PRIDE components. PRIDE project consists of a number of separate components which are outlined here. PRIDE XML format is the basic data structure. RDBMS implementation is a possible rendering of the hierarchical XML format into a relational schema. PRIDE core libraries constitute the basic object model representations of the PRIDE data structure, as well as I/O objects that allow easy interaction with both the reference database implementation and XML format. PRIDE web libraries have been built to allow web-based submission and access to PRIDE. Web libraries use the PRIDE core libraries for data access and processing and can report query results in PRIDE XML as well as in XSL-transformed HTML.

3.2 XML data structure

XML is a standard text format developed by the WWW Consortium (W3C, http://www.w3c.org) that has quickly become a very popular and widely applied means of storing data as well as exchanging them. XML is a hierarchical (tree-like) structure, which fits well with typical proteomics experiments and can easily be validated. XML documents can also readily be extended, which allows them to retain a rather large degree of flexibility. For these reasons, XML was chosen to form the basic data structure for the PRIDE project rather than a relational database structure.

3.3 Relational database implementation

Relational databases provide highly efficient storage of structured data and can readily be optimized for extremely fast retrieval of data based on queries. These queries can be fed to the database through the use of a standardized interface: SQL.

Contrary to the traditional approach, which relies on a relational schema as the basic data structure, PRIDE instead builds upon the XML schema for reasons discussed above. Since XML is hierarchical in nature, and a database is relational, the mapping of an XML schema to a database schema is not straightforward. In fact, many different relational approaches can model exactly the same hierarchical schema. Therefore, the reference implementation provided by PRIDE is just one of the possible forms this database might take. The strength of the XML schema-based structure is that, depending on specific needs, other relational implementations can be created by third parties that emphasize or optimize different aspects of the data stored. As such, PRIDE can be molded to take many different queryable forms, each with distinct strengths and weaknesses.

3.4 The PRIDE data format

In PRIDE, one or more experiments are contained in the root tag "ExperimentCollection". The ExperimentCollection simply groups together one or more "Experiment" tags, which are the top-level tags for individual results. As such, a submitter is likely to submit a collection containing a single experiment, unless the data are extensive or varied enough to warrant the creation of multiple, distinct Experiment elements within PRIDE. The downloadable flat file on the other hand, will hold an ExperimentCollection root consisting of all experiments that constitute the PRIDE database at its release time. The ExperimentCollection structure will enable easy splitting of the download into multiple files when the PRIDE download file eventually becomes very large.

The top-level structure of an experiment, schematically represented and exemplified in Fig. 4, consists of seven conceptually distinct parts, which will be summarized next.

The first of these is the experiment accession number. This number is assigned after successful submission of an experiment and provides a unique pointer to all the associated data. The experiment accession number would be the data element of choice for inclusion in papers as the PRIDE reference because of its conciseness and since interested readers can easily use it to quickly retrieve all relevant data from the PRIDE web interface.

The second part contains meta-data about the experiment: a descriptive title, contact person and/or address, a short label, a description, and finally location information. The contact person and location information are meant to be complementary, *i.e.*, to have geographic and laboratory information in the location field, and contact person information in the contact information. Typically, one would expect the "contact person" field to contain the e-mail address of the corresponding author of a publication.

The third element concerns the sample studied. It consists of a description field and an attribute list. The structure and usage of the latter is discussed in more detail below.

Protocol information constitutes the fourth part of an experiment. Apart from a description and attribute list, it also holds one or more sections about the mass spectrometer(s) used. This latter section contains manufacturer, model, source, and analyzer information which can be further supplemented through an attribute list.

The fifth part details the information derived from the mass spectrometer. This section holds the MS coefficient (*e.g.*, MS^2, MS^3), peak lists, optional raw data references, comments, and an attribute list.

The most intricate subsection of an experiment is the sixth part and deals with the identifications obtained from the data specified in part five. Identifications have been split in two different types: 2-D PAGE-based identifications and nongel-based identifications. A schematic representation of the shared and specific elements for both of these subtypes is shown in Fig. 5. The shared elements are wrapped up in an abstract ancestor element called "IdentificationType". Note that the additional information for 2-D PAGE-based identifications centers on protein-related data gathered during the gel-separation phase, whereas the gel-free identifications typically require more information about the effective identification score and threshold (if available). This has been done to accommodate more stringent standards for identifications, as discussed in recent publications [10–12].

Finally, the seventh part is not restricted to the experiment level but can be found in many of the smaller branches as well. This is the "AttributeList" which represents a list of attributes, to be keyed from controlled vocabularies, allowing an extremely flexible way of integrating additional information into the core schema without sacrificing the structure of the whole. In fact, the PRIDE schema presents a *minimum minimorum* of information about protein identifications in present day proteomics. Many additional pieces of information (*e.g.*, from cone voltages and temperatures on ESI-type ion sources to the specific search parameters used for

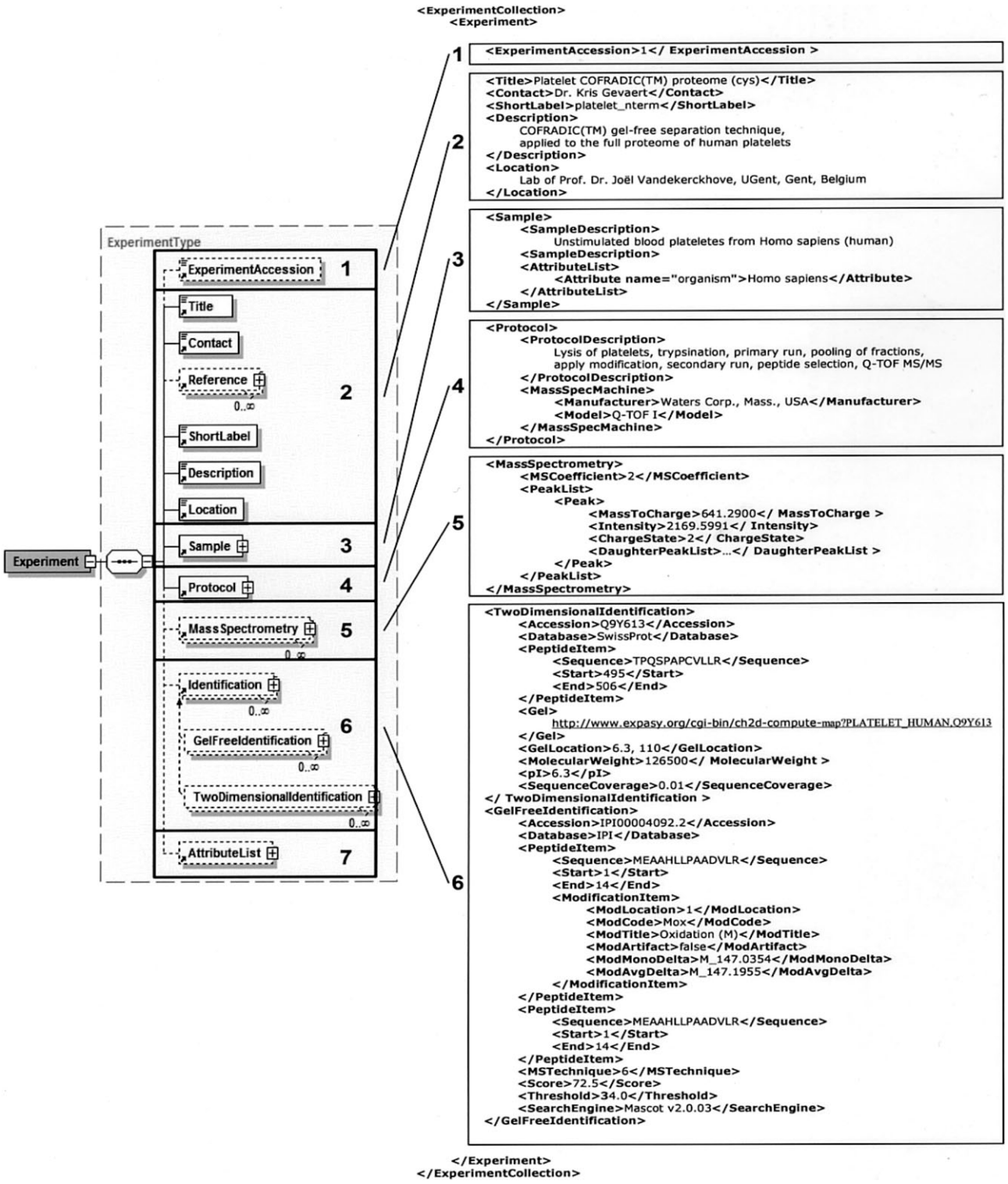

Figure 4. Top-level view of the PRIDE XML data structure and example document. Boxed and numbered areas represent conceptually distinct parts of the Experiment node top-level structure (see text for details). Optional elements are indicated by dashed boxes and multiplicity (if applicable) is shown below the box on the right. Note that an abbreviation has been introduced in the "PeakList" element due to space constraints. Recursive "DaughterPeakList" element is symbolically filled out by "...". An example of a simple AttributeList element (number 7) is represented in the "Sample" element (number 3). Please also note that the presented document is meant to provide an example only. As such, the occurrence of a "TwoDimensionalIdentification" in what is described as a "gel-free separation technique" has only demonstrative purposes.

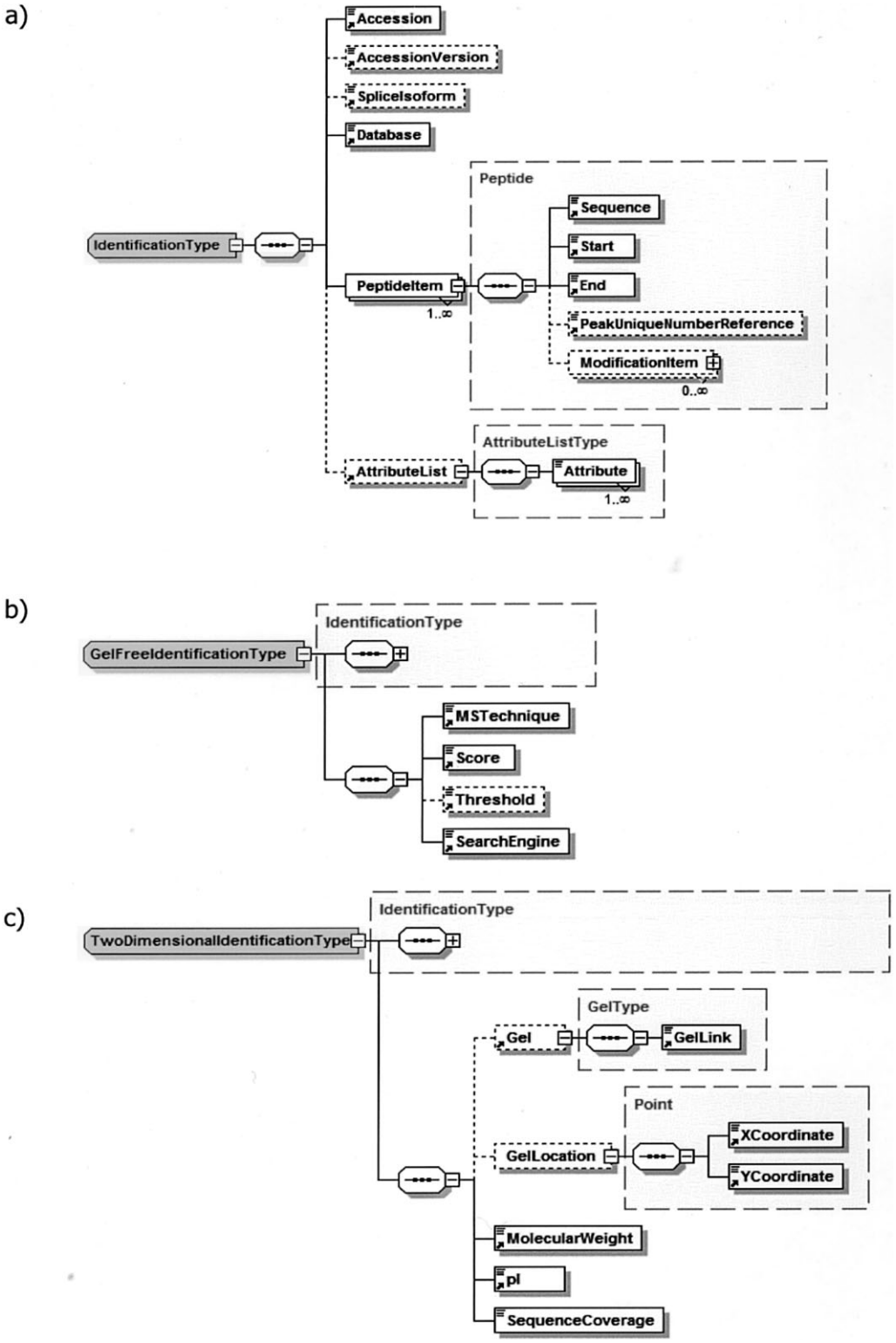

Figure 5. Detailed view of the identification data structures. Abstract ancestor element IdentificationType, shown in (a) contains the shared data elements for the two implementing forms, "GelFreeIdentificationType" (b) and "TwoDimensionalIdentificationType" (c). Both have specific properties which are displayed for each. Note that the properties that were inherited from IdentificationType have been collapsed for clarity in (b) and (c). Optional elements are indicated by dashed boxes and multiplicity (if applicable) is shown below the box on the right.

querying a sequence database with a fragmentation spectrum) that are gathered during the identification process are extremely useful to other researchers, yet few people actively gather and store this information in a structured way. Therefore, the inclusion of this information cannot now be mandatory, nor is it possible to mold it to a defined structure. Indeed, the field of proteomics is still evolving quite rapidly, and allowing for this kind of semistructured data makes PRIDE flexible enough to accommodate future requirements without a major overhaul. In the long term, it is conceivable that some of these attributes become standard elements, whereas others will become obsolete.

PRIDE will in fact be gradually extended to embrace the proteomics standards initiative-general proteomics standards (PSI-GPS) [13] as they become available, thus shaping the PRIDE format into an implementation of this broader format. Most notably, the mzData format for storage of mass-spectrometer derived information is quickly reaching maturity, and as soon as the controlled vocabularies for this format are released (expected by spring 2005), PRIDE will incorporate this format. The mzIdent format, meant to capture the identifications that result from searches based on MS data, is in a more primitive stage at this point, but PRIDE will adopt this standard upon availability as well. The success of previous PSI standard formats [14] has led us to committing ourselves to their GPS standards, yet other proposed standards for data interchange formats which have been published in peer-reviewed literature [15, 16] will also be accommodated by automated conversion in the near future.

3.5 Comparison between the PRIDE data format and the PPP database at UM

Even though the PRIDE database draws in part upon the PPP structure devised at UM, both models do not overlap in full. This is mainly due to the slightly different focus of the respective databases. PRIDE, being developed as a generic repository for proteomics data, necessarily lacks some of the very detailed structures present in a single-purpose database such as the PPP database at UM. Specifically, the PPP database provides more structured detail both at the level of the protocol description as well as the identification process (including full details about the searches performed). This additional level of detail in the PPP database enables the collaboration to compare the findings across similar yet slightly different plasma samples, across technologies and platforms used, and across identification algorithms applied to the data.

Although this information is not present in PRIDE as structured data (*i.e.*, there are no database columns or XML tags with corresponding names), the ability to cope with this data is inherently present through the use of attribute lists. Coupled to controlled vocabularies, which can be both particular to as well as shared across experiments, these attribute lists enable storage of any desirable additional level of detail at several crucial points in the PRIDE data structure.

3.6 The PRIDE web interface

PRIDE presents a default web interface that provides three areas of functionality: the ability to search and query the PRIDE database, the facility to register as a data submitter or collaborator, and the ability to submit data to PRIDE.

Five types of queries are supported in the current release (Fig. 6). Queries on experiment title and accession number are particularly useful when the user wants to view the full list of identifications for an (published) experiment. Additionally, queries can also be performed on a text fragment from a reference, enabling users to obtain the identifications associated with a certain publication or author. Querying the database by protein accession number lists all known identifications of the specified protein across experiments, together with detailed identification information such as the peptides identified and their modifications. Finally, PRIDE can be searched by sample name. This allows the user to see all proteins identified in a certain tissue, cell type, or organism, again across all experiments and with full details.

The PRIDE web interface can be configured to return HTML formatted results as well as XML formatted results. As such, it caters for both human readers (HTML format) and machine readers (XML format). The latter allows users to write scripts that perform an off-line meta-analysis on the results of PRIDE queries in an efficient way.

The ability to register as a data submitter or collaborator has been implemented with several goals in mind. First, restricting data submission to registered users coupled with a simple level of data curation will help to avoid spurious data being uploaded into PRIDE. The system allows the creation of collaborations, such that submitted data can be kept private and shared only amongst collaborators until such a time as they wish to make their data public. This functionality allows PRIDE to be used as a tool for collaboration and data sharing within a consortium as well as serving as a final repository for published data. The same applies to PRIDE as a system for peer-reviewing data that has been submitted for publication, since data can be privately shared between author(s), the journal, and the reviewers. Obviously it is also possible for a data submitter to declare their data publicly at the point of submission, excluding it from any of the restrictions described above.

4 Concluding remarks

The PRIDE project has resulted in the construction of a unique combination of tools, standards, and infrastructure that for the first time enable the construction of a truly global, centralized proteomics data repository. The highly modular design makes PRIDE flexible enough to be adapted by third parties to create more or less differing mirrors with new or specialized views on the same data.

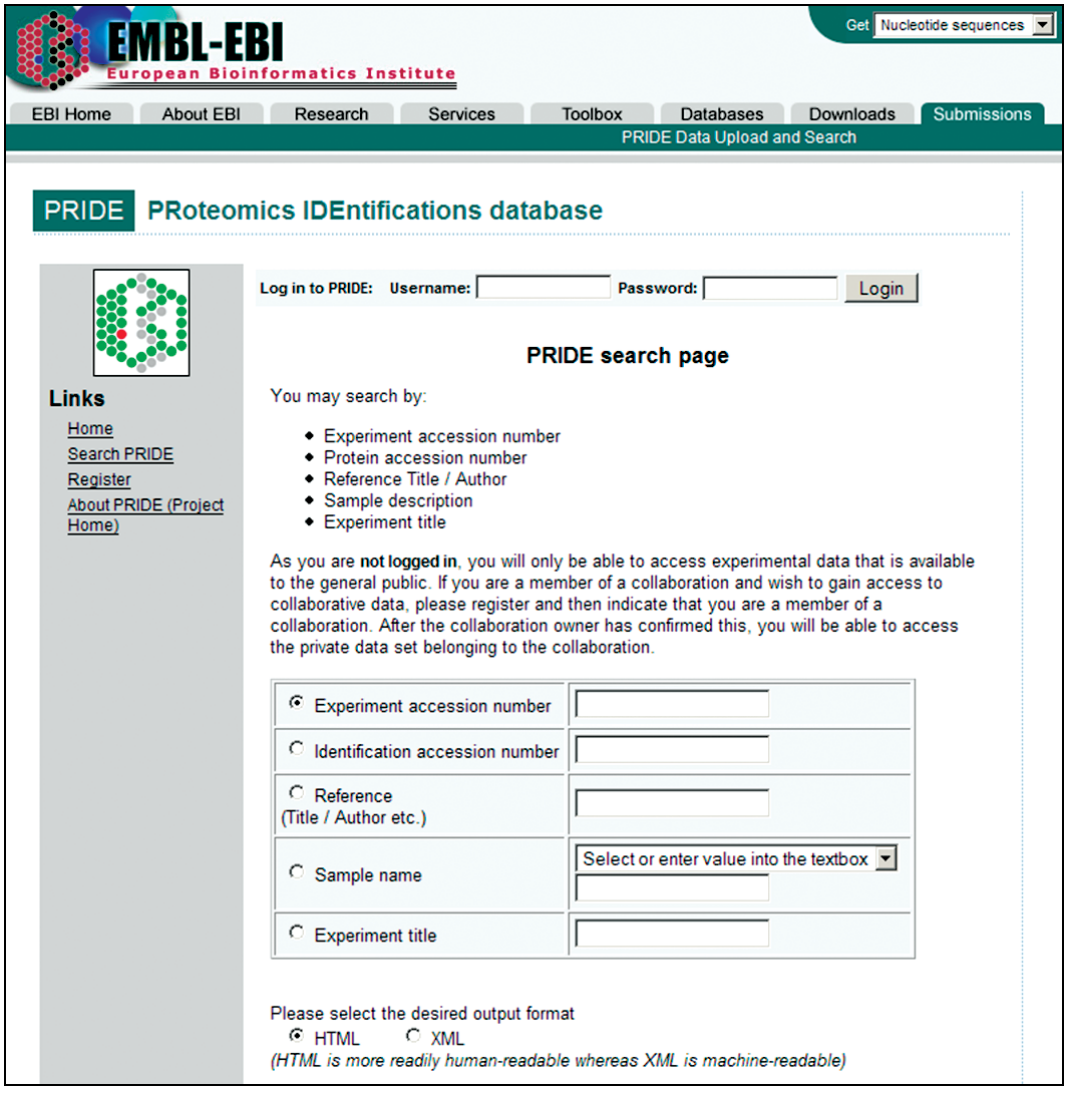

Figure 6. PRIDE web interface for queries. Three types of query are supported (1) by experiment accession number, (2) by protein accession number, (3) by reference text fragment, (4) by sample, or (5) by experiment title.

Indeed, as PRIDE starts to gather data, we expect new and unexpected uses of the publicly available data to come up. Statisticians might seize the opportunity to constructively contribute to the way database search software functions or could enhance the procedures used to distinguish between true identifications and false positives. Biologists can mine the data in search for new research targets and software developers can come up with new ways to store, visualize, and query the large amounts of data that will accumulate over time.

It is our conviction that PRIDE will be an important milestone in the evolution of the field of proteomics and it is our hope that it will become the highly active hub of proteomics data that we designed it to be.

The PRIDE database can be accessed on-line at http://www.ebi.ac.uk/pride.

The authors would like to extend their gratitude to Gilbert S. Omenn for his support of the project and for the many informative discussions which have contributed to this paper. L.M. would like to thank Samuel Kerrien, Mark Rijnbeek, and Kai Runte for their invaluable comments, suggestions, and contributions to the development of the PRIDE code base, reference relational database implementation, and XML schema. The PRIDE project was funded in part through the European Commission Programme "Quality of Life", Marie Curie Training Site Fellowship, Contract number: QLRI-1999-50595. Parts of the data used in this paper were generated in the context of the IWT-GBOU-research initiative (Project number 20204) of the Flanders Institute of Science and Technology (IWT). K.G. is a Postdoctoral Fellow and L.M. a Research Assistant of the Fund for Scientific Research-Flanders (Belgium) (F.W.O. Vlaanderen).

5 References

[1] Zhang, H., Yan, W., Aebersold, R., *Curr. Opin. Chem. Biol.* 2004, *8*, 66–75.

[2] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R., Hoorelbeke, B., Demal, H., Martens, L. , *Mol. Cell. Proteomics* 2002, *1*, 896–903.

[3] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., Vandekerckhove, J., *Nat. Biotechnol.* 2003, *21*, 566–569.

[4] Gevaert, K., Ghesquière, B., Staes, A., Martens, L., Van Damme, J., Thomas, G. R., Vandekerckhove, J., *Proteomics* 2004, *4*, 897–908.

[5] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., *Nat. Biotechnol.* 2004, *22*, 471–472.

[6] Rohlff, C., *Expert Rev. Proteomics* 2004, *1*, 267–274.

[7] Omenn, G. S., *Proteomics* 2004, *4*, 1235–1240.

[8] Hanash, S., Celis, J. E., *Mol. Cell. Proteomics* 2002, *1*, 413–414.

[9] Adamski, M., Blackwell, T. W., Menon, R., Martens, L., Hermjakob, H., Taylor, C. F., Omenn, G., States, D., *Proteomics* 2005, *5*, this issue.

[10] Carr, S. A., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., Nesvizhskii, A., *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[11] Veenstra, T. D., Conrads, T. P., Issaq, H. J., *Electrophoresis* 2004, *25*, 1278–1279.

[12] Nesvizhskii, A. I., Aebersold, R., *Drug Discov. Today* 2004, *9*, 173–181.

[13] Orchard, S., Taylor, C. F., Hermjakob, H., Zhu, W., Julian, R. K. Jr., Apweiler, R., *Proteomics* 2004, *4*, 2363–2365.

[14] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Saliwinski, L., Ceol, A., Moore, S. *et al.*, *Nat. Biotechnol.* 2004, *22*, 177–183.

[15] McDonald, W. H., Tabb, D. L., Sadygov, R. G., MacCoss, M. J., Venable, J., Graumann, J., Johnson, J. R. *et al.*, *Rapid Commun. Mass Spectrom.* 2004, *18*, 2162–2168.

[16] Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

