REGULAR ARTICLE

Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project

Marcin Adamski[1], *Thomas Blackwell*[1], *Rajasree Menon*[1], *Lennart Martens*[2], *Henning Hermjakob*[3], *Chris Taylor*[3], *Gilbert S. Omenn*[1] *and David J. States*[1]

[1] University of Michigan, Ann Arbor, MI, USA
[2] University of Ghent, Ghent, Belgium
[3] European Bioinformatics Institute, Hinxton, UK

The pilot phase of the HUPO Plasma Proteome Project (PPP) is an international collaboration to catalog the protein composition of human blood plasma and serum by analyzing standardized aliquots of reference serum and plasma specimens using a variety of experimental techniques. Data management for this project included collection, integration, analysis, and dissemination of findings from participating organizations world-wide. Accomplishing this task required a communication and coordination infrastructure specific enough to support meaningful integration of results from all participants, but flexible enough to react to changing requirements and new insights gained during the course of the project and to allow participants with varying informatics capabilities to contribute. Challenges included integrating heterogeneous data, reducing redundant information to minimal identification sets, and data annotation. Our data integration workflow assembles a minimal and representative set of protein identifications, which account for the contributed data. It accommodates incomplete concordance of results from different laboratories, ambiguity and redundancy in contributed identifications, and redundancy in the protein sequence databases. Recommendations of the PPP for future large-scale proteomics endeavors are described.

Received: March 28, 2005
Accepted: April 29, 2005

Keywords:
HUPO / Plasma Proteome Project / Protein identification / Proteomics database / Proteomics data integration

1 Introduction

Data management was one of the key elements in the pilot phase of the HUPO Plasma Proteome Project (PPP). Data submission and collection approaches were defined collaboratively by the Bioinformatics and Technologies Committees, and were extensively discussed at the PPP Workshop in Bethesda, USA in July 2003 [1].

Ideally, experimental methods and the data generated by their execution would be fully described in a thoroughly decomposed manner, facilitating sophisticated searches and analyses. However, when dealing with the results from real experiments multiple compromises must be made. The first concerns the level of detail that can be requested: while it is, in principle, desirable to have all methodological steps, parameters, data, and analyses described in full detail, many laboratories lack automated laboratory information management systems and manual record keeping is laborious, limiting the granularity of information that can be captured. The second compromise concerns the degree to which experimental reports will be decomposed and structured by the submitter: from a long run of free text as in a journal paper to a fully annotated list of all the relevant items of information, arranged in an elaborate and well-specified hierarchy that captures the

Correspondence: Dr. David J. States, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA
E-mail: dstates@bioinformatics.med.umich.edu
Fax: +1-734-615-6553

Abbreviations: PPP, Plasma Proteome Project; **PRIDE**, proteomics identifications interrelationships of those items. It is notoriously difficult to automatically extract even the simplest information from free text [2, 3]. However, thoroughly classifying information for submission is burdensome. Indeed, developing standards, data definitions, forms or submission tools, and the associated documentation and training material is a substantial task. Third, the pilot phase of the PPP was designed to encourage individual laboratories to push the limits of their technologies to detect and identify low-abundance proteins; the Technology Committee was not able to define in advance all the parameters that emerged as desirable inputs for analysis in this broad, largely voluntary collaboration. The fourth compromise concerns the design and implementation of the data systems used for storage of the data at the central repository. It is desirable to retain as close a link as possible to the original submissions from the participating laboratories in the central repository, but this implies that the details of which data sets superseded earlier submissions, exceptions encountered in the data loading, and other detailed information on submission processing need to be encoded in subsequent queries, complicating the task of writing and debugging software to analyze the data.

Finally, a compromise at the level of the overall project relates to the choice of sequence database used for analysis and whether to "freeze" on a particular release of the sequence database. The results of protein identification by search of mass spectra against a database are necessarily dependent on the database being searched. Freezing on a particular protein sequence database release not only facilitates comparison of identification data sets but also prevents corrections and revisions to the protein sequence collection from being incorporated into the identification process. Further, freezing on a particular protein sequence database release complicates the task of linking the findings of the current study to evolving knowledge of the human genome and its annotation, because many of the entries in the protein sequence database available at the initiation of the project have been revised, replaced, or withdrawn over the course of this project, and continue to be revised.

The major aim of the pilot phase of the HUPO PPP was the comparison of protein identifications made from multiple reference specimens by all participating laboratories. An additional important aim was the development of an efficient method of data acquisition, storage, and analysis in such a big collaborative proteomics experiment [3]. Here we describe the data management system developed during the pilot phase of the HUPO PPP.

2 Materials and methods

2.1 Development of the data model

To encourage participation by laboratories, the data model focused on identifications of whole proteins as a high-level, concise description of experimental results, requiring a minimum of data input, transmission, and potential reformatting. The guidance specified the collection of the protein accession numbers and names, binary descriptions of the confidence of the protein identifications (high or lower), lists of identified peptides, and free text descriptions of experimental protocols. Analysis of the preliminary results brought to the fore a major problem with a data integration and validation process based exclusively on protein accession numbers. Participating laboratories used not only different search databases but also different algorithms to assemble protein identifications from their database search results. Additionally, the estimation of confidence of the identification, based on search scores and laboratory binary judgment, was inconsistent. To address these problems, the original data model was enhanced to include the peak lists used to obtain protein identifications, and raw spectra in the instrument native format.

The expanded data model is generally in concert with recently proposed guidelines for publication of protein and peptide identification data [4]. Since our studies were started before publication of these guidelines, our data collecting decisions do not reflect all of the requirements proposed by Carr *et al*. [4] Table 1 compares the guidance proposed in [4] with the information collected in the present study. The HUPO PPP data model consists of the following main objects:

2.1.1 Laboratory

Information about the participating laboratories, such as principal investigator, contact person, postal and email addresses, identifiers, descriptions, *etc*.

2.1.2 Experimental protocol

Free text descriptions sufficiently detailed to allow the work to be reproduced. The level of experimental detail was specified to be sufficient for the protocol to be considered for publication in Proteomics or the Journal of Biological Chemistry.

2.1.3 Protein identification data set

The identified protein accession numbers, names, search database and version, sequences of the identified peptides, and an estimate of confidence for each protein identification, plus any supporting information about PTMs (from experimental measurements, or other sources), and estimates of relative protein abundance in the specimen. Identification data sets were stored as peptide lists, reflecting the fact that some laboratories applied significant protein fractionation prior to tryptic digest and mass spectral analysis. In a pure "bottom up" strategy, any protein can contribute any peptide and no information is gained by retaining group structure for peptides. However, when protein fractionation is used, knowledge that a group of peptides were all derived from the same protein fraction can enhance the power of identification.

Table 1. Comparison of the HUPO PPP data model with guidance for publishing peptide and protein identification data by Carr *et al.* [4]

Guideline proposed by Carr *et al.*[4]	HUPO PPP data model
1. Supporting information	
The method and/or program used to create the "peak list" from raw data and the parameters used in the creation of this peak list.	Data were collected as a part of free text description of performed experiments. Recommendation to use PEDRO tool was moot, since tool was not ready for use.
The name and version of the program(s) used for database searching and specific parameters used for its (their) operation.	Name of the search program collected, but not version or operation parameters.
Scores used to interpret MS/MS data and thresholds and values specific to judging certainty of identification, whether any statistical analysis was applied to validate the results, and a description of how it was applied.	Scores and thresholds were collected.
The name and version of sequence database used; the count of number of protein entries in it at the time searched.	Both name and version of the sequence database were collected. The sequence database itself was also recorded.
2. Information regarding the observed sequence coverage	
Table that lists for each protein the sequences of all identified peptides.	Peptides (sequences) identified for each protein were collected.
To calculate the sequence coverage different forms of the same peptide are to be counted as only a single peptide.	All forms of identified peptides were collected, but as long as they have the same amino acid sequence they were counted only once.
The total number of MS/MS-interpreted spectra assigned to peptides corresponding to each protein.	Raw spectra were collected.
3. Protein assignments based on single-peptide assignments	
The sequence of the peptide used to make each such assignment, together with the amino acids *N*- and *C*-terminals to that peptide's sequence.	Sequence of the peptide was collected but not the terminal information.
The precursor mass and charge.	The precursor charge state was collected as a part of the peptide data. The mass was requested as part of the peak list information.
The scores for this peptide.	Scores were collected.
4. Biological conclusions based on observation of a single peptide matching to a protein	
Such conclusions must be supported by inclusion of the corresponding MS/MS spectrum.	Raw spectra were requested for all the MS/MS identifications (including single peptide).
5. Peptide mass fingerprint data	
In addition to listing the number of masses matched to the identified protein, authors should also state the number of masses not matched in the spectrum and the sequence coverage observed.	Only peptides matched to the identified protein were collected. Sequence coverage was calculated.
Parameters and thresholds used to analyze the data.	Data collected only as a part of free text description of performed experiments. No particular information was requested.
6. Ambiguous protein identifications	
The same protein appears in many cases under different names and accession numbers in the database. When matching peptides to members of such a family, it is the authors' responsibility to demonstrate that they are aware of the problem and have taken reasonable measures to eliminate redundancy. In cases where a single-protein member of a multiprotein family has been singled out, the authors should explain how the other members of the group were ruled out.	A data integration workflow was specially designed to address this problem. It is described in the following sections.
7. Submission of MS/MS spectra	
Submission of all MS/MS spectra mentioned in the paper as supplemental material. The dta, pkl, and mgf files are accepted.	Raw spectra in the instrument native format were collected and are available on request. They may be converted to the other formats with use of special software.

2.1.4 Peak list

Lists of mass over charge peaks used by search engines for protein identifications. The peak lists were accompanied by amino acid modifications catalogs, lists of all modified residues, including the symbol for and mass of the modified residue, and the type of modification.

2.1.5 Summary of technologies and resources

This included estimates of the time, capital, and operating costs of the analyses.

2.1.6 MS/MS spectra

The unprocessed data from spectrometers.

2.1.7 SELDI peak list

Peak lists from direct MS/SELDI experiments (registered for a separate analysis; see Rai *et al.*, this issue).

2.2 Data submission process

The data submission strategy was designed to make the submission process simple for the participants and at the same time error-proof and relatively easy to process for the data collection and integration center. As stated above, the consensus data model of the PPP pilot phase included only a limited representation of methods and results, to minimize the time commitment for participating experimentalists. Two methods for submitting were offered: (a) a combination of Microsoft Excel™, Microsoft Word™, and text forms, or (b) an XML (http://www.w3.org/XML) schema-based file format (PEDRO [5, 6]). Those who chose the form-based submission were asked to fill out a set of preformatted Excel/Word/text document templates, and submit them online using a web-based submission server at the University of Michigan. Those who chose the XML format were asked to email their submissions to the European Bioinformatics Institute, after generating one or more XML documents using the provided XML schema. The schema of the XML document allowed for the collection of all the information in one, hierarchically organized file. To generate the XML documents the participants were encouraged to use the PEDRO data entry tool [6], or to export XML directly from their existing LIMS system. The XML documents were checked for compliance with the schema and forwarded to the University of Michigan for further processing.

During the course of the project, we decided to request the raw MS/MS spectra in the form of instrument files in spectrometer native format. The size of these files, sometimes in excess of several gigabytes, did not allow for their collection by the standard data submission route; instead, CD or DVD disks were submitted to the University of Michigan Core and distributed to three groups for special cross data set analyses (see Omenn *et al.*, Kapp *et al.*, and Beer *et al.*, this issue).

At the beginning of the project each participating laboratory received two distinct identifiers: the first, a numeric public identifier used for interactions with the submission centers and other laboratories, and the second, a three-character private code known only to the laboratory and the central data analysis group. These private identifiers were used to create data surveys without disclosing the identity of submitters.

2.3 Design of the data repository

The project data repository was built with a Structured Query Language (SQL) relational database server. The data structure was divided into two main parts: (1) an intermediate structure presenting an exact copy of the data from documents submitted by the project participants, to make the data available for further processing, and for checking correctness of the submitted documents; and (2) the main data structure designed to hold the integrated project data.

The structure can be divided into four main sections: (1) experiment description, (2) protein identifications made by data producers from peptide sequences, (3) MS/MS peak lists, and (4) protein identifications from database searches made by groups other than the data producers.

In the database design (Fig. 1), experiments performed by the project participants are stored in the entity Experiment. This entity is referenced directly by the entity Laboratory and by a set of look-up entities: Specimen, Depletion, SeparationProtein, ReductionAlkylation, SeparationPeptide, and MassSpec. Experiment also has a many-to-many relationship with a free text protocol description (entities Protocol and ExperimentProtocol). At the experiment level the database structure branches into two sections. The first section started by the entity IdentificationSet stores protein identifications submitted by the participants. The second section started by the entity MsRun stores MS peak lists and the results from their analysis. The two-branched database structure reflects the changes in the project data collection model, from identification-oriented at the beginning to a more fine-grained description utilized later.

The database can capture three sets of protein identifiers from the same experiment. The first set stores protein identifications made by data producers in the entity Identification. The second set stores the results of peptide list searches done by the data integration center, in the entity ProteinByPeptides. This set captures peptide group information. The third set of identifiers (multiple subsets of these identifiers are possible) is derived from the same experimental results, but this time by an analytical group other than the data producer, through the MsRun branch of the database (entities MsRun, MzPeak, and ProteinByMsSearch).

The main project database does not store SELDI peak lists or MS/MS raw spectra. These data are available as downloadable files.

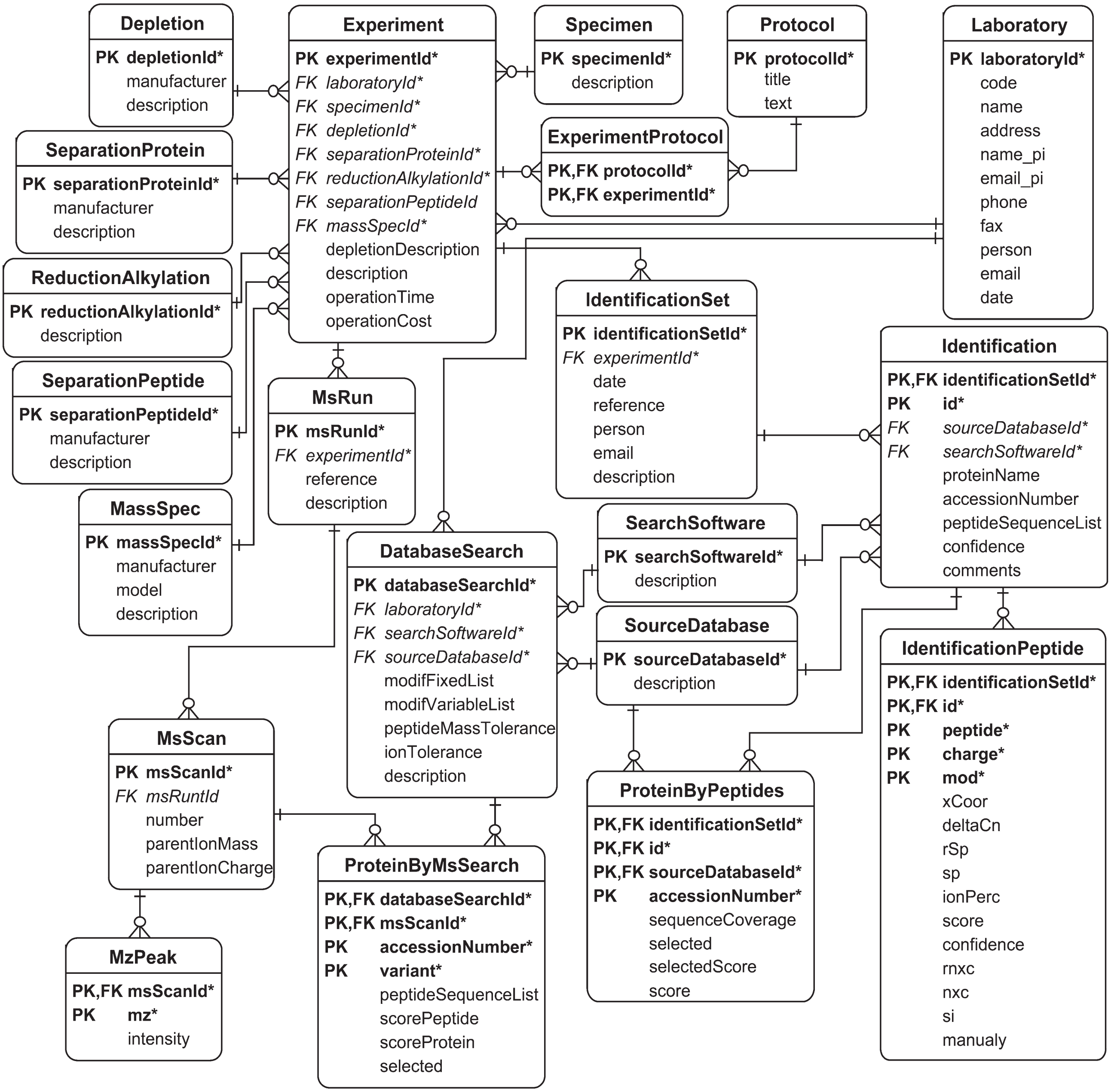

Figure 1. Entity-relationship diagram of the HUPO PPP data repository. Boxes symbolize entities or tables; connecting lines represent relations between the entities.

2.4 Receipt of the data

The data documents were uploaded using a web-based submission site established at the University of Michigan. During submission each document received a unique ID number used subsequently by the document tracking and transforming mechanism. The XML documents submitted by email were processed separately. Data from the received documents were transferred to an intermediate database. The transfer was done automatically for each web-submitted document and separately for the emailed XML submissions. The data in the intermediate structure represent an exact copy of the data from the original documents, without any transformation or integration. The intermediate database allows checking the correctness of the structure of the submitted documents and makes the data available for the integration procedures. Verified data were then rewritten using a consistent format for protein accession numbers, database names, peptide sequences, peak lists, and experimental categories.

3 Inference from peptide level to protein level

In the pilot phase of the HUPO PPP, proteins were identified by MS experiments, followed by searches of protein databases to find peptide sequences matching observed spectra. Often, such a search returns a cluster of proteins, all of which contain the same set of matching peptides. Problems with ambiguity of protein identifications obtained from searches of tandem mass spectra and methods for managing them have been widely discussed, *e.g.*, by Nesvizhskii *et al.* [7] and Sadygov *et al.* [8]. In these earlier works, protein identifications were inferred from lists of assigned peptides accompanied by probabilities that those assignments are correct. In the present report, however, we integrated lists of peptides obtained using several different search algorithms and different search databases, which frequently lacked identification probabilities. Although during the course of the project, participants were asked to additionally submit peptide and protein identification probabilities or scores, as well as peak lists and raw MS spectra, the main part of integrating the results was based solely on the sequences of the submitted peptides. The raw spectra and peak lists were subject to separate analysis and will be described elsewhere.

The integration workflow we describe here benefits from the collaborative character of the studies and is based on a heuristic approach that assumes that the proteins most likely to be truly present in the sample are those supported by the largest number of maximally independent experiments. The workflow additionally takes into account the "level of annotation" of the protein, thus preferentially selecting the proteins with the most extensive description available.

The workflow algorithm includes several consecutive steps:

(1) Assemble peptide sequence lists: Protein identifications submitted by the participating laboratories were accompanied by lists of sequences of matched peptides. All the lists were collected to form a set of distinct peptide sequence lists. Each list in that set preserves all references to its origin, *e.g.*, if a particular list is reported from more than one experiment, it has more than one reference.

(2) Search the peptide lists: Each peptide sequence list obtained in the previous step was subsequently searched against the IPI version 2.21 (July 2003) database [9]. This was selected as the standard database of the project. Each match requires 100% identity between sequences and disregards flanking residues.

(3) Select one representative protein from each cluster of equivalent protein hits: Often, more than one entry in the reference protein database matches all of the components of a peptide sequence list. We call this set of matching entries a "cluster of equivalent protein hits" for that peptide sequence list. The clusters for different lists may overlap. When they do, we wish to choose one protein entry from the intersection of several clusters to represent all proteins in each of the overlapping clusters,that is, the proteins identified by each of the associated peptide sequence lists. The selection is done as follows.

Each protein entry in the reference database receives three integer scores:

(a) The number of different laboratories reporting a peptide sequence list whose cluster includes this protein.

(b) The number of distinct experiments (laboratories × specimens × protocols) reporting a peptide sequence list whose cluster includes this protein.

(c) The number of identifications (laboratories × specimens × protocols × clusters) for clusters including this protein. For each peptide sequence list, the cluster member with the largest value of score (a) is chosen as the representative protein entry. Scores (b) and (c), followed by criteria (d–g) listed below, are applied in succession to break numeric ties at higher levels.

(d) Well-described protein – product of a well-described gene. The EnsEMBL gene model was used for the annotation. The "well-described" proteins and genes are those with a nonempty description line, and without words like "fragment", "similar to", "hypothetical", "putative", *etc.* in their description.

(e) Well-described protein-product of any gene.

(f) Well-described protein not assigned to any gene.

(g) Protein not assigned to any gene and described as a fragment, by its similarity to another protein, or with no IPI description line at all. Any remaining ties are broken by selecting the protein having the lower IPI number.

As a result, one protein will generally be chosen as the representative entry from several overlapping clusters of equivalent protein identifications. This simplifies later comparisons between laboratories and experiments. This particular choice for a representative protein is motivated by the idea that the protein whose identification is supported by the largest number of independent experiments is the protein most likely to be actually present in the specimen. Score (a) counts each laboratory only once, no matter from how many specimens or with how many different peptide sequence lists the laboratory identified this protein. Next in importance, score (b) counts the number of independent experiments in which the protein was identified. Score (c) counts all reported peptide sequence lists, even if several results are from the same experiment. Criteria (d–g) indicate the level of annotation for each database entry. They facilitate selection of the best-described proteins.

4 Summary of contributed data

Laboratories participating in the project submitted a total of 12 667 distinct protein accession numbers. This number includes 11 253 accession numbers from MS/MS – both MALDI and LC-ESI, and an additional 1414 IDs from FT-ICR-MS. FT-ICR-MS identified 2230 proteins, but 816 were also identified by the MS/MS technologies. In addition, participating laboratories contributed 653 identifications from MALDI-MS peptide mass fingerprints. These data were analyzed separately and will be reported elsewhere.

Table 2. Usage of the search databases

Category	Search database			
	IPI	Swiss-Prot	NCBInr	All three
Submitted protein identifications	11 960	199	508	12 667
Submitted identifications with peptide sequence lists found in IPI database	11 741 98%	196 98%	451 89%	12 388 98%
Entries in IPI database matching submitted peptide sequence lists	15 463	488	552	15 710
Average number of IPI entries *per* submitted protein identification	1.3	2.5	1.2	1.3

The majority of reported protein identifications from the MS/MS and FT-ICR-MS experiments (11 960 of 12 667 – 94%) were obtained by searching the tandem mass spectra against the IPI database. The remaining 6% were generated using either the Swiss-Prot or NCBInr databases (Table 2). Almost all of the submitted peptide sequence lists (12 388 of 12 667 – 98%) were matched in the standard database for the project, *i.e.*, IPI version 2.21. The 2% of peptide sequence lists for which no exact match was found in this database most likely represent up to 5% mismatch between database entries, which is permitted when constructing the IPI database (see [9]). We believe that the submitting laboratory searched one of the source databases for IPI, rather than IPI itself, and matched the spectrum to a source entry which is included in IPI as a secondary rather than a master entry.

The 12 388 reported identifications with peptides matching the IPI 2.21 database correspond to 18 098 distinct peptide sequence lists. Searching these lists against IPI 2.21 results in 15 710 matching entries. For each of 12 303 of these lists (68%), exactly one of 6601 IPI entries was matched. These were reported with 7000 different protein accession numbers, including Swiss-Prot and NCBI identifiers. The 6% reduction from 7000 to 6601 distinct identifiers comes from converting Swiss-Prot and NCBI identifiers to IPI identifiers. As these identifications are already unique, the integration workflow did not additionally reduce these 6601 accession numbers.

In the remaining 5795 (32%) cases, each peptide sequence list matches more than one IPI protein sequence, resulting in an ambiguous identification or a cluster of equivalent hits (Table 3). In this group of ambiguous identifications, searches of the 5795 peptide sequence lists return 9668 distinct IPI protein accession numbers. The integration workflow reduces this group to a set of 3273 distinct proteins, which explain the presence of all reported peptides. In the next step, the 6601 accession numbers from the group of uniquely identified proteins are combined with the 3273 accession numbers from the group of ambiguous identifications. Of the resulting 9874 identifications, 9506 represent distinct accession numbers.

Details of the integration process for the 5795 clusters of ambiguous hits are presented in Table 4. Scores (a–c) evaluate the level of confirmation of each protein identification by the number of completely independent experiments.

In 2044 (35%) of the cases, the decision of protein selection was done on the basis of the score (a): selecting a protein detected by the largest number of laboratories. In 1680 (82%) of those cases it was a single protein, and no additional selection step was required. In the remaining 18% of the cases, selection by score (a) returned more than one protein. The tie was then broken using additional scoring categories (d–g). In 2966 (51%) of the cases, all proteins in the cluster were indistinguishable using scores (a–c) and the decisions were made exclusive using categories (d–g).

Table 3. Effectiveness of the integration process

Category	Number of IPI entries matching single-peptide sequence list		
	One (distinct IDs)	More than one (indistinct IDs)	One or more (all IDs)
Submitted peptide sequence lists	12 303	5795	18 098
Submitted protein accession numbers	7000	5388	12 388
Matching entries in IPI database	6601	9668	15 710
Matching entries in IPI database after the integration	6601	3273	9506
Reduction level of submitted accession numbers to IPI entries	6%	39%	23%

Table 4. Number of clusters qualified on different levels of the integration

Integration level		Number of clusters	
A	Number of laboratories	1680	2288
B	Number of experiments	419	
C	Number of reports	189	
D	Well-described EnsEMBL gene	2429	3507
E	Any EnsEMBL gene reference	99	
F	No EnEMBL reference	286	
G	Poorly described protein	693	
	Total number of potentially ambiguous peptide sequence lists processed		5795

The categories (d–g) classify IPI database entries by the amount of detail in their description. It is then reasonable to compare such a classification of proteins in the project database with the same classification of proteins in the complete IPI database. Details of this comparison are given in Table 5. This shows that 41% of entries from the HUPO PPP database and 24% of the entries from the IPI database belong to the highest category (d) – the best-described proteins. The intermediate categories (e) and (f) include relatively few proteins while category (g) – the least described proteins – contains the majority of the entries, 49 and 63% for the HUPO PPP and IPI databases, respectively. For the HUPO PPP database, the ratio between the percentage of entries from categories (d) and (g) is 41/49% = 0.84. This ratio for the IPI database is 24/63% = 0.67. Thus, the laboratories were more likely to identify better-described proteins. This result can be interpreted as confirming the presence of proteins that were previously studied in detail, possibly because of their relative abundance or ease of identification. Alternatively, the integration workflow itself preferred the best-described proteins wherever possible, pushing the ratio toward category (d).

To further compare results from the HUPO PPP with all the proteins from IPI, we compared the distributions of peptide sequence length (number of amino acid residues *per* peptide) in both data sets (Fig. 2). The distribution of peptide length from the HUPO PPP database is noticeably shifted toward longer peptides – median equal to 12.9 residues – in comparison to the distribution of the lengths of tryptic peptides in IPI-median equal to 10.5 residues. We hypothesize that the under-representation of short peptides may be explained by the nature of the tandem mass spectrum search algorithms which require the spectra from short peptides to be of much better quality than spectra from longer peptides, to result in a significant match. Many laboratories did not report any peptides shorter than five residues. The fraction of nontryptic peptides in each peptide length bin is very small. These peptides were identified in a few nonenzyme-specific database searches and, as they passed quality control in the participating laboratories, they were included in our analysis. The origin of these peptides is not analyzed in this paper, but we speculate that they may be products of other endogenous proteases present in the tissue of origin or in human plasma [10].

Table 5. Distribution of numbers of entries from the HUPO PPP and complete IPI databases in the integration categories

Integration category	Complete IPI database		HUPO PPP database		
	No. of entries	Fraction of all entries	No. of proteins	Fraction of all identifications	Fraction of IPI entries
D	13 588	24%	3900	41%	29%
E	855	2%	220	2%	26%
F	6633	12%	716	8%	11%
G	35 454	63%	4670	49%	13%
All	56 530		9506		17%

Based on the nonuniform reporting of short peptides from participating laboratories, the limited spectral data available for short peptides, and the limited power for protein identification using a peptide present in multiple protein sequences, we decided to eliminate peptides shorter than six residues from further analysis. In doing so, we disregarded two protein identifications, each based on a single peptide of five amino acids. This reduces the number of accepted protein identifications from 9506 to 9504 accession numbers.

4.1 Cross-laboratory comparison, confidence of the identifications

The distribution of the number of protein identifications among participating laboratories is shown in Fig. 3. Individual laboratories are encoded using their numeric identifiers. The 18 laboratories identified a total of 9504 distinct IPI proteins. The number identified by individual laboratories varied from 52 to 4569. The laboratories were asked to mark as "high confidence" those identifications that passed more stringent criteria, chosen by each laboratory individually, although the PPP did issue guidance after the June 2004 Jamboree Workshop for SEQUEST searches to use $X_{\mathrm{corr}} \geq 1.9, 2.2, 3.75$ for $1+$, $2+$, and $3+$ ions, respectively, plus $\Delta C_{\mathrm{n}} \geq 0.1$ and $R_{\mathrm{Sp}} \leq 4$ for tryptic peptides. The number of these lab-reported high-confidence identifications ranged from 21 to 789. To further assess the confidence of protein identifications from individual laboratories, we counted the number of proteins, which were also reported by a second laboratory. We considered such identifications to be confirmed. The fraction of confirmed identifications is higher for laboratories, which submitted lower numbers of proteins. This may be caused by several factors including the followings. (1) Different stringencies for acceptance of the identifications – smaller sets may mean that more stringent criteria have been used and the resulting proteins are more likely to be true identifications. (2) Differences in experimental techniques – smaller sets of proteins may be obtained by shallower sampling, picking up only the more abundant, *i.e.*, more frequently identified proteins. (3) The

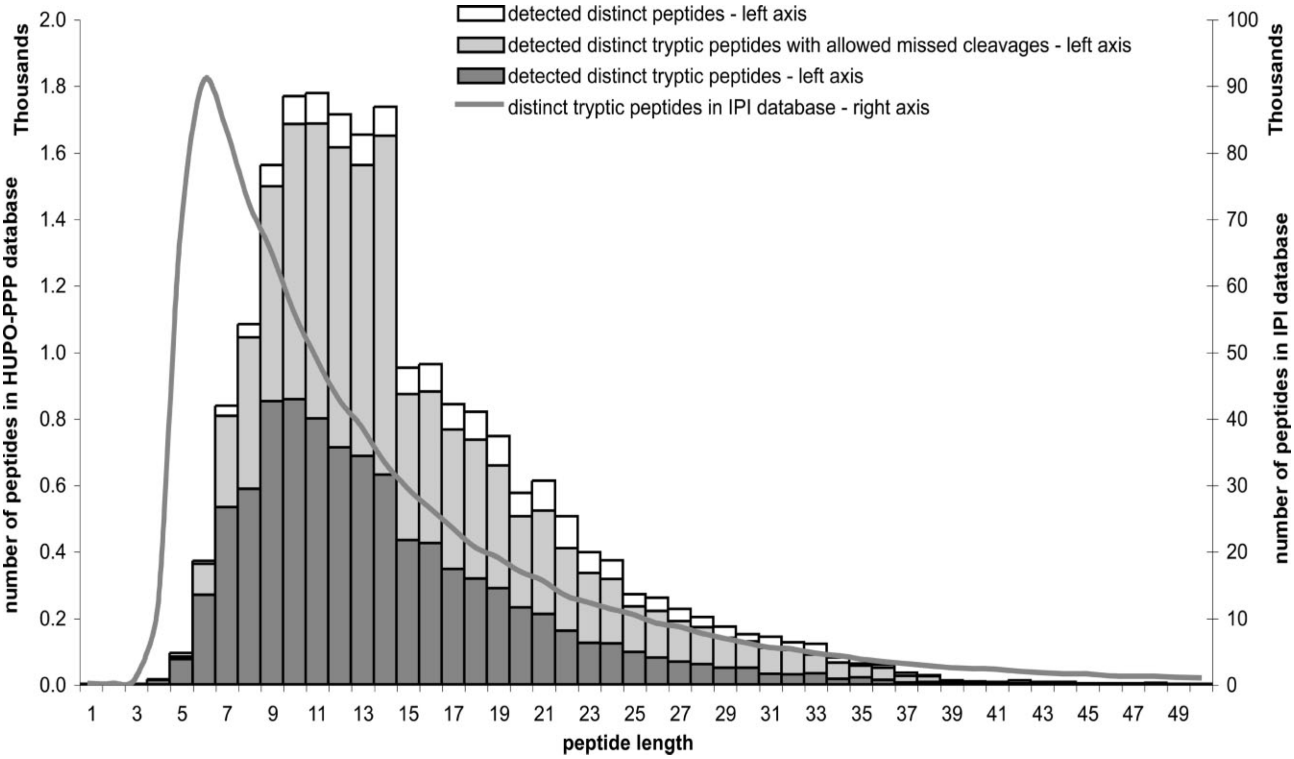

Figure 2. Comparison of distributions of length of tryptic peptides (dark gray bars), tryptic peptides with missed cleavages allowed (light gray bars), and all peptides, including nontryptic peptides (white bars) detected in the course of the project using MS/MS (both MALDI and LC) and FT-ICR-MS methods, to the distribution of the length of tryptic peptides from the complete IPI database (gray line).

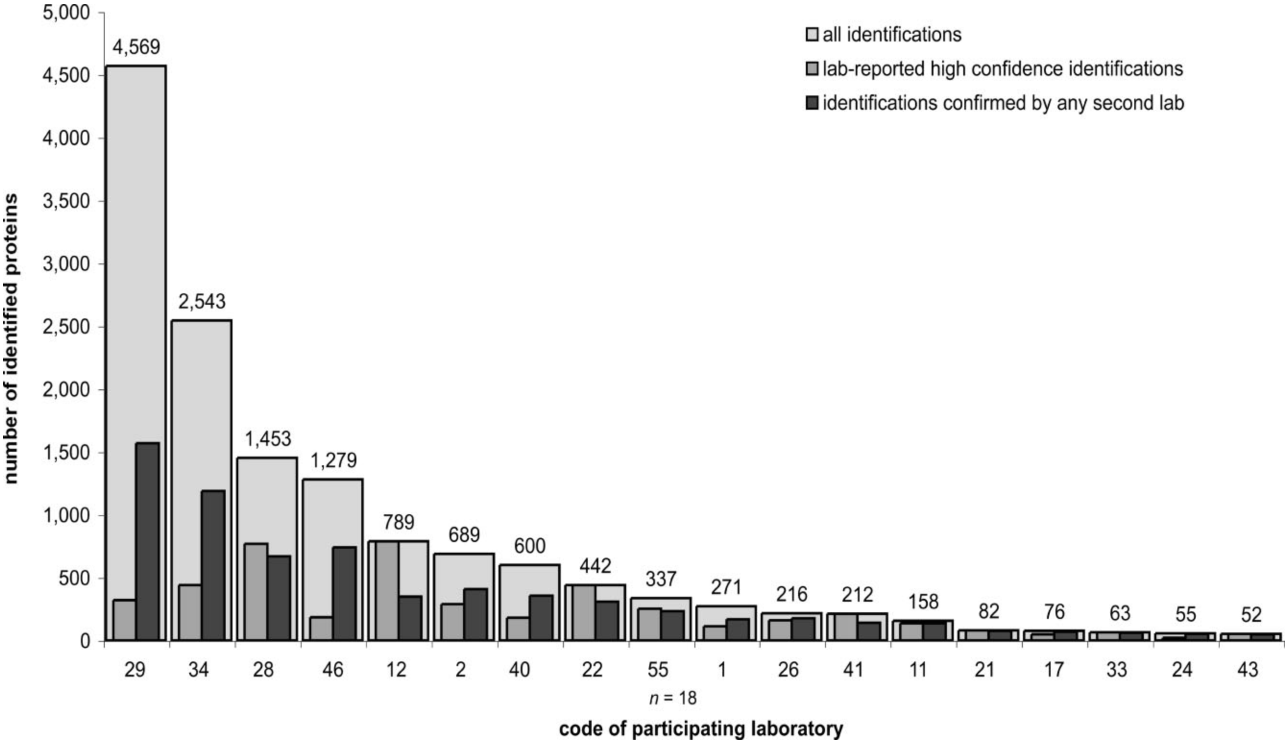

Figure 3. Distribution of MS/MS and FT-ICR-MS protein identifications among 18 participating laboratories, encoded using their numeric identifiers.

intrinsic nature of the confirmation process – the more sensitive the procedures used by a particular laboratory are, the more likely it is that it will be the only laboratory reporting a particular identification. Thus, the requirement for confirmation penalizes the laboratories that submitted the largest data sets.

The level of cross-laboratory confirmation of the identifications, as a function of the number of peptides detected across experiments and laboratories, is shown in Fig. 4. The first category – all identifications – has a confirmation level equal to 25%. The second category, resulting from elimination of single-peptide identifications, dramatically reduces the number of proteins from the original 9504 to 3020, and at the same time raises the confirmation level to 75%. The absolute number of confirmed identifications in these two categories is virtually the same, meaning that of 6484 single-peptide protein identifications almost none was confirmed. Limiting the identifications to those which are supported by an even larger number of peptides causes a further reduction in the number of proteins and a rise in the confirmation level.

The analysis described above led us to categorize protein identifications into four classes, based on the level of the identification confidence. The four categories are organized in a diamond-shaped parallelogram (Fig. 5). Identifications from the least stringent category – "all identifications" (9504 proteins) – are divided into two more stringent, parallel categories: "high-confidence identifications" (2857 proteins), including proteins reported at least once as high-confidence, and "multipeptide identifications" (3020 proteins), including proteins for which two or more distinct peptides were reported project wide, following data integration. The most stringent category "high-confidence multipeptide identifications" (1555 proteins) includes proteins from the intersection of the preceding categories. Proteins in this category are identified with two or more distinct peptides, requiring at least one to have been reported as part of a high-confidence protein identification.

5 False-positive identifications

False-positive peptide identifications exist and are widely acknowledged to be a problem [7, 8, 11–15]. One arises whenever the top-scoring database match for a particular spectrum has a score which passes all reporting thresholds, yet the matched database sequence is not the same as that of the biological specimen in the instrument. This will occur for a variety of reasons. The spectrum may represent a mixture of different peptides with almost equal parent masses and elution times. The biological specimen may be a contaminant or an allelic variant not recorded in the database being searched. Even if the database contains the correct amino acid sequence, this sequence may fall outside the scope of the search, due to PTMs or requirements for proteolytic cleavage. In each of these cases, the top-scoring match within scope and within the database is returned by the search software. If its score passes reporting thresholds, the (mis)match will be accepted and reported as a peptide identification.

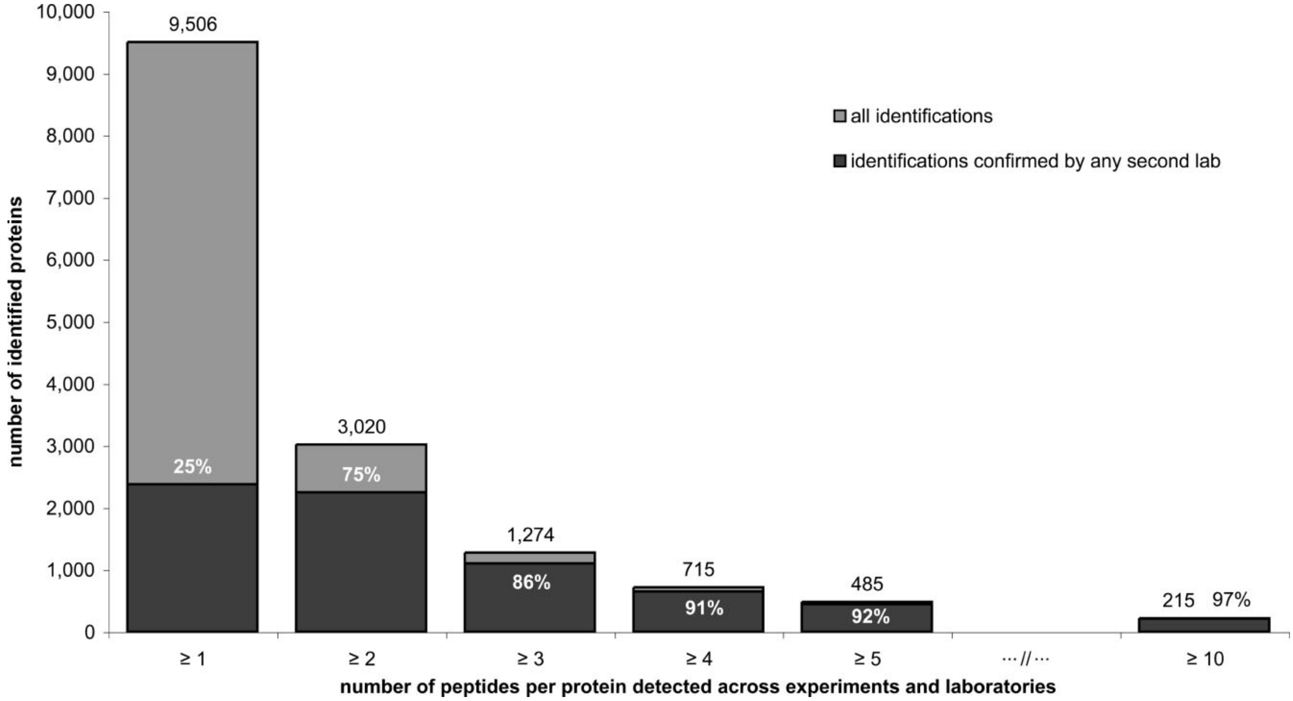

Figure 4. Distribution of MS/MS and FT-ICR-MS protein identifications as a function of the number of peptides detected *per* protein.

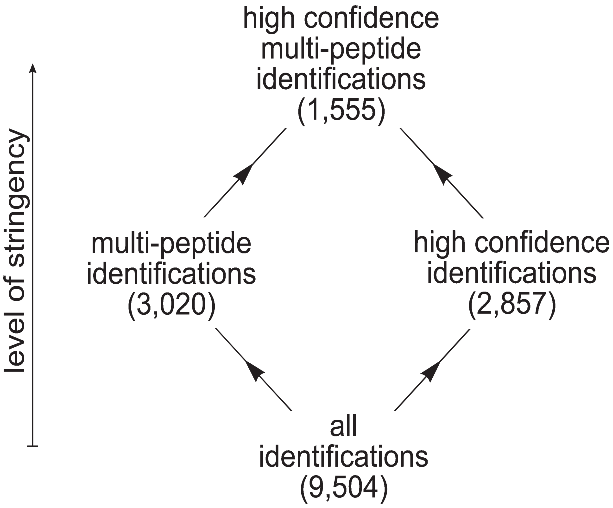

Figure 5. Proposed classification of the identification stringency levels; the number of protein identifications at each level is shown in parentheses.

False-positive and true-positive peptide identifications show opposite behavior when we accumulate large numbers of peptide identifications, as in this project [7, 11]. One expects false-positive peptide identifications to accumulate roughly proportional to total peptide identifications. However, the chance that two or more false-positive peptide identifications coincide on the same database entry should be no better than random. On the contrary, a protein which is present at a detectable concentration in the specimen will produce many tryptic peptides in nearly stoichiometric quantities. Increased sampling should increase the number of distinct peptides, which are reported, and all of these will map to the same (correct) database entry. This means that, as we accumulate more and more peptide identifications, the class of protein identifications based on a single peptide reported project wide is simultaneously depleted of correct peptide identifications (as these are promoted to multiple-peptide protein identifications) and refilled with false-positive protein identifications. Below, we consider a range of values for the fraction of such false-positive identifications. One major participating HUPO laboratory, after manually reviewing several hundred of their protein identifications, concluded that a single peptide constituted sufficient evidence in perhaps 20% of the cases where only one peptide from a protein had been seen. The acceptance rate after manual review was much larger for proteins identified using two or three peptides, precisely because of the selection described above. Manual review of all the spectra was not feasible, and all of their identifications were submitted to the database.

To assess the confidence of protein identifications, we use a Poisson model for the distribution of false-positive peptide matches. Two parameters are needed to specify the model: the total number of database proteins and the number of peptide level matches that are incorrect.

The IPI version 2.21 database contains 56 530 sequences, with some redundancy and overlap between entries. To model the database integration procedure, the two largest tryptic peptides from each database entry were calculated, and all entries containing exact matches to these two peptides were collapsed into a sequence group. This process resulted in 49 924 sequence groups. This is used as the number of bins in the random model.

Lower and upper bounds for the number of false peptide level matches are estimated by assuming either that all of the lower confidence single-peptide identifications are erroneous or that all single-peptide identifications, regardless of confidence, are erroneous. Of the 6484 identifications based on a single peptide project wide, 1956 were assigned with high confidence by at least one participating laboratory and 4528 are lower confidence identifications. The Poisson distribution parameter λ is chosen so that the random model predicts the assumed number of false single-peptide identifications. The range for λ lies between 0.146 and 0.211. The estimate of 80% false-positive rate cited above gives $\lambda = 0.168$, within this range. Values for λ larger than 0.211 would predict more protein-level identifications due to false positives alone than the 9504 total identifications reported, and are inconsistent with the random model.

For each $k = 0, 1, 2, 3, \ldots$ the expected number of database entries (out of 49 924) supported by exactly k false-positive peptide matches is calculated from a Poisson distribution. These are allocated in proportion among the reported protein identifications with $s \geq k$ supporting peptides. Only the predictions for which $s = k$ result in false-positive identifications at the protein level. The principle here is that a protein identification is considered correct if at least one of its supporting peptide identifications is correct. The allocation is illustrated in Table 6, and protein-level confidence is summarized in Fig. 6 and Table 7.

At the lower bound, the random model predicts 268 false-positive identifications at protein level among 1746 proteins with exactly two distinct peptides reported project wide, and 10 false positives among 1274 proteins with three or more distinct peptides project wide. The confidence within each class is the observed number of identifications minus predicted false positives, divided by the observed number of identifications. A lower bound on error becomes an upper bound on confidence. These upper bounds are a confidence of 85% for identifications based on exactly two peptides and 99% for those based on three or more peptides. Corresponding worst-case estimates are 70 and 97% for exactly two and for three or more peptides, respectively.

We acknowledge uncertainty in the exact value for λ. However, qualitative interpretations of the data are not sensitive to λ. For the quantity of data accumulated in this study, and throughout the range of choices for λ, the confidence in protein identifications based on four or more peptides easily exceeds 0.99 and for identifications based on exactly three peptides project wide, it varies from 0.95 ($\lambda = 0.211$) to 0.99 ($\lambda = 0.146$). Both classes achieve the traditional 95% con-

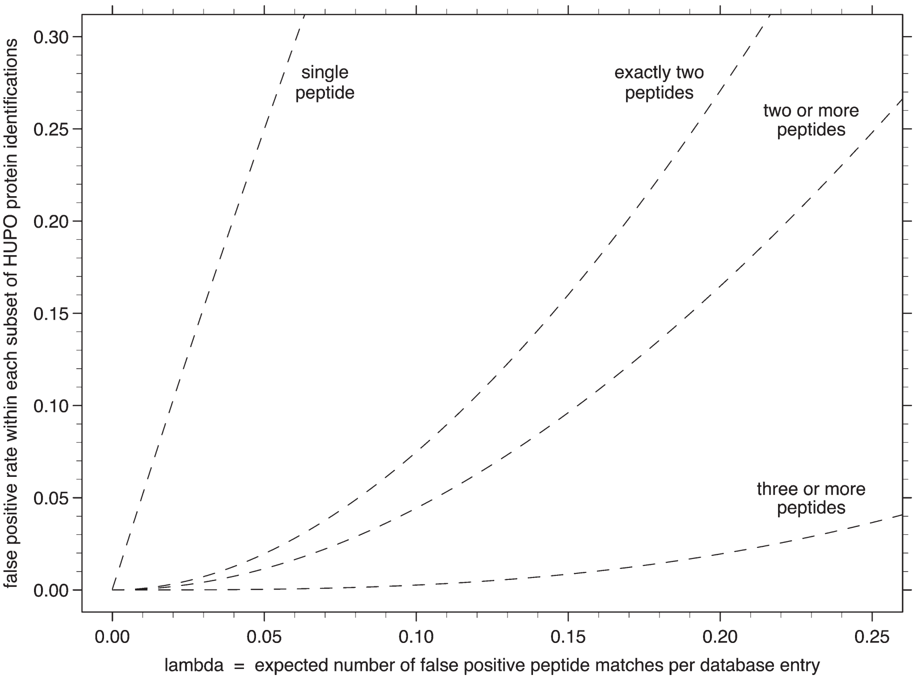

Figure 6. At protein level, false-positive identifications are strongly concentrated among the protein identifications based on a single peptide project wide. This figure shows predicted error rates (1-confidence, vertical axis) from the Poisson model as a function of λ (horizontal axis, expressed as the expected number of false-positive peptide reports *per* IPI database entry). Four curves represent the classes of protein identifications based on exactly one, exactly two, two or more, and three or more distinct peptides reported project wide.

Table 6. Allocating predicted false positives among observed identifications for $\lambda = 0.146$. Predicted total number of proteins with exactly *k* false-positive supporting peptides (right-hand column) is allocated proportionally among the observed identifications with $s \geq k$ supporting peptides (preceding columns). Each column total is the number of observed identifications with exactly *s* supporting peptides, and each row total is the number of identifications predicted to have exactly *k* false-positive supporting peptides. Only the cases where $s = k$ (main diagonal, bold type) produce false-positive identifications at the protein level

S	0	1	2	3	4	≥5	Total number of proteins with *k* false-positive peptides predicted from Poisson model
k							
0	**40 420**	1956	445.87	140.24	57.64	121.53	43 141.28
1		**4528**	1032.16	324.65	133.42	281.33	6299.56
2			**267.97**	84.29	34.64	73.04	459.94
3				**9.83**	4.04	8.52	22.39
4					**0.26**	0.55	0.82
≥5						**0.02**	0.02
Number of observed protein identifications	40 420	6484	1746	559	230	485	49 924

s, number of distinct peptides project wide; *k*, number of distinct false-positive peptides.

fidence threshold for accepting an assertion as true, regardless of λ. The confidence for identifications based on exactly two peptides project wide varies from 0.7 ($\lambda = 0.211$) to 0.85 ($\lambda = 0.146$). Again, regardless of λ, these identifications would be described in lay language as “probably correct, but by no means sure”. The majority of single-peptide identifications are false under any reasonable values for λ.

We have chosen to concentrate further analysis on the 3020 identifications made with two or more peptides project wide for two reasons. Excluding identifications based on exactly two peptides would exclude a large number of identifications that we believe are probably correct. Second, it would introduce a strong bias toward highly abundant proteins. Since the goal of the PPP is to identify a representative set of blood proteins, we chose to base subsequent analyses on the 3020 core data set, realizing that we are including a number of false-positives, but yielding a more representative view of the human plasma proteome.

Table 7. Confidence in protein identifications as predicted by the Poisson model

Number of peptides s	Reported identifications	Predicted false positives		Confidence	
		$\lambda = 0.146$	$\lambda = 0.211$	$\lambda = 0.211$	$\lambda = 0.146$
1	6484	4528	6484	0	0.302
2	1746	268	533	0.695	0.847
3	559	10	28	0.950	0.982
4	230	0.26	1.08	0.995	0.999
$\geq$2	3020	278	562	0.814	0.908
$\geq$3	1274	10	29	0.977	0.992
$\geq$4	715	0.27	1.12	0.9984	0.9996
$\geq$5	485	0.01	0.04	0.9999	0.9999

The wide range of concentrations for proteins in blood plasma and serum presents an additional complication. Clinical ELISA assays, where available, report a measurable concentration for many proteins that were never reported by MS. Almost every protein in the body is potentially present at some concentration in blood plasma or serum, whether as an intact protein or as degradation products. There is no set of proteins we can exclude as known negatives; a large number of potential positives are present at unknown but low concentrations. A similar situation is found in *Saccharomyces cerevisiae*. A recent tagging experiment [16] measured protein concentrations spanning four orders of magnitude for 4251 proteins, roughly 80% of all proteins expressed in log-phase yeast. Two separate MS/MS surveys conducted earlier [11, 17] show low concordance in protein identifications. They reported roughly 1500 proteins each, with 57% of proteins in common and 43 or 41% reported in one survey but not in the other. In yeast, as well as in this project, the reporting of low-abundance proteins is highly variable.

6 Data dissemination

The project participants accessed the database through a web-based SQL interface developed specifically for project needs. During the data submission process, before the official in-project data release, each laboratory could retrieve only its own data submitted to date. After the in-project data release, laboratories could freely access data from all the participants. The database access was limited to the project laboratories by a user and password mechanism. Each laboratory could use a set of predefined SQL queries to perform standard data requests as well as define its own, private queries for more specific tasks and save these for future use.

For the dissemination of the data gathered by the HUPO-PPP, the *ab initio* construction of a novel data structure was decided upon. Indeed, the PPP, as the first HUPO project to complete the pilot milestone, is uniquely positioned for fulfilling the pioneering role in establishing such a data (infra) structure. The finalized data are publicly available in the proteomics identifications (PRIDE) database (http://www.ebi.ac.uk/pride) (see Martens *et al.*, this issue). The results of a PRIDE web query can be visualized either as an HTML page or in the PRIDE XML format. The complete PRIDE database is also available for download in XML format. The PRIDE project site offers an Application Programmers Interface (API), which provides the tools necessary to efficiently access the PRIDE XML format and reference database implementation programmatically.

7 Discussion

The PPP integration workflow is based on a heuristic approach that the protein identifications most likely to be true are those which are supported by the largest number of independent experiments. The strength of the “independent experiment” term is gradually loosened in consecutive steps of the algorithm to select a single protein, which represents a whole cluster of equivalent identifications.

Such an optimization approach, by its nature, may not always lead to the smallest set of proteins possible. For example, let us consider a simplified problem where there are only six protein identifications in the database – A, B, C, D, E, and F. All of them are products of independent experiments. Furthermore, they are single-peptide identifications associated with distinct peptides a, b, c, d, e, and f, respectively. Searching for these peptide sequences in the protein database shows that the peptides can be found in three different proteins with overlapping sequences – p1, p2, and p3.

Figure 7 depicts the problem: rows represent the three proteins, columns the six peptide identifications. If a particular peptide can be found in a specified protein, it appears in the appropriate row.

Scoring the proteins using the algorithm results in: p1 = 4 (four different identifications), p2 = 3 (three different identifications), and p3 = 2 (two different identifications). This leads to the following assignment of the protein accession numbers to the identifications: ID A → p1, ID B → p1, ID C → p1, ID D → p1, ID E → p2, ID F → p3. Although it complies with the algorithm, the selection of protein p2 for identification E is not optimal from a mathematical point of view. If protein p3 were assigned instead of p2, the size of the set of proteins would reach its minimum. In a real experiment, the coincidence of such a particular overlap of the protein sequences and specific scoring conditions necessary to cause the algorithm to fail is very rare. Processing a subset of the HUPO PPP MS/MS and FT-ICR-MS data resulting in 9504 distinct protein identifications caused the algorithm to fail (*i.e.*, not to reach the minimum) in only ten cases.

Maximizing the number of independent supporting experiments also biases the selection of representative proteins towards those with the longest sequence, as illustrated

```
     |  id A       id B       id C       id D       id E       id F
     | <pept a>   <pept b>   <pept c>   <pept d>   <pept e>   <pept f>
     |
_____|__________________________________________________________________
     |
p1   | <pept a>---<pept b>---<pept c>---<pept d>
     |
p2   |                       <pept c>---<pept d>---<pept e>
     |
p3   |                                             <pept e>---<pept f>---
```

Figure 7. Theoretical example presenting a situation where the integration workflow may not produce the minimal possible set of proteins.

```
     |  id A       id B       id C       id D       id E       id F
     | <pept a>   <pept b>   <pept c>   <pept d>   <pept e>   <pept f>
_____|__________________________________________________________________
     |
p1   | <pept a>---<pept b>---<pept c>---<pept d>---<pept e>---<pept f>---
     |
p2   | <pept a>---<pept b>---<pept c>---<pept d>
     |
p3   |                                             <pept e>---<pept f>---
```

Figure 8. Length bias in representative protein selection. Shown in the figure are a precursor, p1, and two proteolytically cleaved products, p2 and p3. Precursor contains all the identifying peptides contained in the products. As a result, the integration algorithm will select the precursor independent of other knowledge about which form might be present in the sample.

in Fig. 8. The algorithms used to construct the IPI database also systematically select longer precursor sequences in preference to shorter forms [9].

A more sophisticated approach might incorporate additional sources of biological information in choosing a representative protein for each group. Sources of such information include protein annotation databases like GO [18] or HPRD [19]. We chose not to pursue this option because current annotation databases have limited coverage and might introduce historical biases into the protein identification process.

The integration algorithm seeks to assign the minimum number of proteins necessary to account for the observed peptide sequence lists. With no *a priori* knowledge of which proteins are present in the blood, an alternative, and equally valid, approach would be to list all proteins from which each peptide might have been derived. Figure 9 compares the results of this latter approach with the integration algorithm presented above. Note that many proteins not selected by the integration algorithm may, nevertheless, have been the source of a large number of observed peptides.

8 Concluding remarks

The pilot phase for the HUPO PPP is the first large-scale collaborative proteome project ever undertaken, and our experience highlights the challenges in data integration that are likely to be encountered in future high-throughput and collaborative proteomics studies. Several issues are identified.

A key decision was to define one recommended protein database and release, IPI 2.21 of July 2003, for all subsequent work in the project. Although this was not universally adhered to by all project participants, it simplified early data comparisons and later merging of results. However, the decision to standardize on IPI release 2.21 also complicated the annotation process. By the time the data-gathering phase of the project had concluded, this release was necessarily out of date. The process of mapping version 2.21 identifiers to version 3.01 identifiers proved to be challenging because of the large number and complex nature of the changes that have taken place in the underlying sequence collection.

We overestimated the laboratories' ability to use XML data formats. Although tools and support for XML were offered, the vast majority of laboratories chose to submit data in Word/Excel formats.

We underestimated the importance of collecting peak lists and raw spectra. The decision to collect data at the level of protein identifications rather than individual peptide identifications meant that information defined at the peptide level, such as peak lists and SEQUEST scores, were not collected.

In order to use tools like PeptideProphet and ProteinProphet [15, 7] to assess the reliability of protein identifications, search results or complete sets of peak lists are required, including those which match with extremely low scores. At the inception of the project, the decision was to perform all data analysis at the participating laboratories and to submit only protein identifications to the central repository. The initial submission forms specified only a minimal set of supporting data. As the project progressed and the data

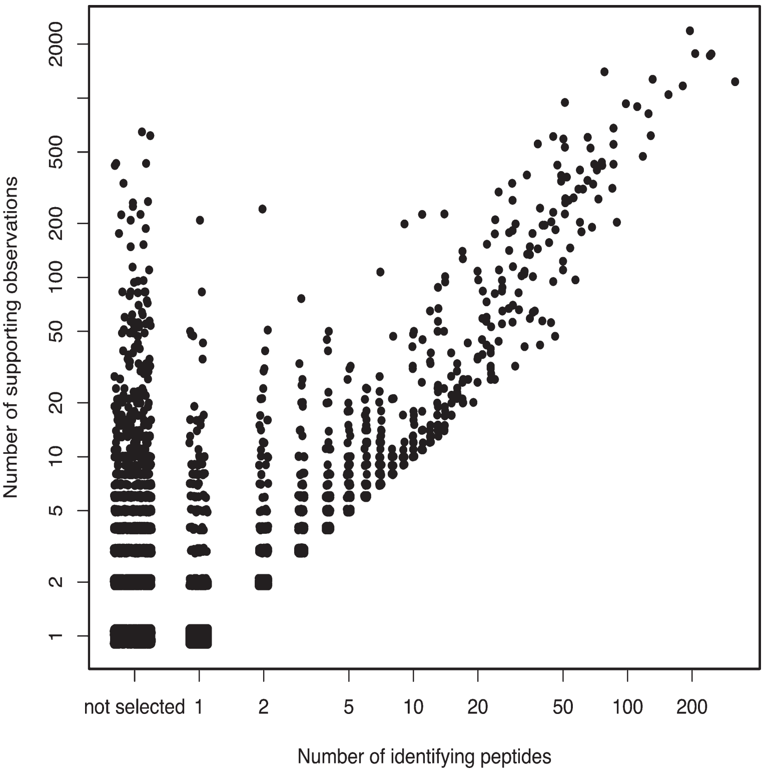

Figure 9. Number of identifying and supporting observations. This figure shows a scatterplot for all the 15 695 proteins in IPI version 2.21, which contain at least one peptide observed in the project. *X* axis is the number of distinct peptides assigned to a protein by the integration algorithm. *Y* axis is the number of distinct (laboratory × experiment × specimen) observations of a peptide which could have been derived from the protein. Note that for some proteins not selected or assigned only one peptide by the integration algorithm, a large number of supporting observations are present in the data set.

repository group assumed more responsibility for quality assurance, we requested more supporting data from the contributing laboratories including mass spectrum peak lists and full binary data files.

The decision to request a pilot round of data submissions proved invaluable in allowing the data repository group to assess the data and identify the problems described above. As a result of this pilot round of data submissions, significant changes were introduced during the project's operation. As a consequence, the data collection/integration center had to deal with the data formatted according to both the old and the new protocols, but the final product of the project was greatly enhanced.

A revised database schema for future projects has been developed; this more extensive, finer-grained schema will better serve the future needs of the PPP, and will also serve as the core for schemata tailored to meet the requirements of other HUPO tissue projects (*e.g.*, liver, brain). In this revised protocol, all entries, whether they contain new data or re-analysis of existing data, are assigned an accession number as a point of reference for use in the publications. The schema is straightforwardly extensible to accommodate additional technologies. For example, we are coordinating with project participants that generate quantitative data. Reliable quantitations, both relative and absolute, can come from a variety of methods such as differential gel electrophoresis, isotope tagging or chemical modification for MS, and protein array technologies [20].

There is also a need to "point outwards" to different resources, often done by creating a field to capture a Uniform Resource Indicator or URI (a generalized version of the familiar URL web address). Such resources include annotation resources such as UniProt (http://www.uniprot.org), EnsEMBL (http://www.ensembl.org), HPRD (http://www.hprd.org), and PeptideAtlas (http://www.peptideatlas.org) [21]. Importantly, URIs can also link to "raw" mass spectrum data repositories (the original output of a mass spectrometer scan as opposed to the heavily processed peak list); these data are increasingly in demand for in-depth analyses [22], but require special handling separate from the main project database, due to their size (see also Martens *et al.*, this issue).

In addition to its main goals of beginning the map of the human plasma proteome and assessing the power of different techniques to resolve proteins, the HUPO-PPP pilot phase has generated an extensive "real world" collection of data that will be invaluable in developing and testing enhanced software tools for proteomics. Both the structure of the revised schema and the experience gained in the pilot phase of the PPP will contribute to other HUPO proteome initiatives, in particular the Liver and Brain Proteome Projects, and the HUPO Proteomics Standards Initiative [23], which seeks to provide general standards for proteomics, both for the level of detail required when reporting work (the Minimum Information About a Proteomics Experiment, MIAPE) and the file format in which such information should be captured.

9 Computer technologies applied

The main project data repository was established with use of the Microsoft SQL server 2000™ working on a Dell Power Edge™ server running operating system Microsoft Windows 2003™. Templates of documents for the data transfer were produced with use of Microsoft Word and Microsoft Excel packages. The data submission site was established on Dell Power Edge server running Microsoft Windows 2003 and Internet Information Services. The online data submission and data access sites were created using Microsoft Visual Studio 2000™ and written in language C#. Data integration procedures were written either in C# or as stored procedures in the MS SQL server native language.

We thank all the participating laboratories for their dedication to this project, Yin Xu of the University of Michigan for data analytical support, and Rolf Apweiler of the European Bioinformatics Institute for overall guidance. We acknowledge support from the National Institutes of Health and from the multiple corporate partners of the Human Proteome Organization Plasma Proteome Project. L.M. would like to thank Kris Gevaert and Joel VandeKerckhove for their useful comments and support. L.M. is a research assistant of the Fund for Scientific Research – Flanders (Belgium), F.W.O. – Vlaanderen.

10 References

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

[2] Blaschke, C., Hirschman, L., Valencia, A., *Brief Bioinform.* 2002, *3*, 154–165.

[3] Shatkay, H., Feldman, R., *J. Comput. Biol.* 2003, *10*, 821–855.

[4] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[5] Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D., Stead, D. A. *et al.*, *Nat. Biotechnol.* 2003, *21*, 247–254.

[6] Garwood, K. L., Taylor, C. F., Runte, K. J., Brass, A. *et al.*, *Bioinformatics* 2004, *20*, 2463–2465.

[7] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., *Anal. Chem.* 2003, *75*, 4646–4658.

[8] Sadygov, R. G., Liu, H., Yates, J. R., *Anal. Chem.* 2004, *76*, 1664–1671.

[9] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[10] Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F. *et al.*, *J. Proteome Res.* 2005, *4*, 53–62.

[11] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., Gygi, S. P., *J. Proteome Res.* 2003, *2*, 43–50.

[12] Tabb, D. L., Saraf, A., Yates, J. R., 3rd, *Anal. Chem.* 2003, *75*, 6415–6421.

[13] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S. *et al.*, *OMICS* 2002, *6*, 207–212.

[14] Sadygov, R. G., Yates, J. R., 3rd, *Anal. Chem.* 2003, *75*, 3792–3798.

[15] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[16] Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W. *et al.*, *Nature* 2003, *425*, 737–741.

[17] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, *Nat. Biotechnol.* 2001, *19*, 242–247.

[18] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, *Nat. Genet.* 2000, *25*, 25–29.

[19] Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R. *et al.*, *Nucleic Acids Res.* 2004, *32 (Database issue)*, D497–D501.

[20] MacBeath, G., *Nat. Genet.* 2002, *32 Suppl*, 526–532.

[21] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. *et al.*, *Genome Biol.* 2005, *6*, Epub R9.

[22] Beer, I., Barnea, E., Ziv, T., Admon, A., *Proteomics* 2004, *4*, 950–960.

[23] Orchard, S., Hermjakob, H., Apweiler, R., *Proteomics* 2003, *3*, 1374–1376.