Enabling Computational Proteomics by Public and Local Data Management Systems

Kenny Helsens, PhD; Lennart Martens, PhD

Data management is of growing importance in proteomics research. Therefore, this review first touches on the different types of proteomics methodologies along with their associated data types and underlying proteomics informatics. Next, we discuss the variety of different standardized file formats to exchange proteomics data as well as available public and local data management solutions to handle data generated by proteomics experiments. Finally, we discuss the opportunities offered by the integration of proteomics data with other biological data sources to new create new insights into cardiovascular disease.

Advances in mass spectrometry have been crucial in enabling high-throughput proteome analysis over the past 2 decades.1 Shotgun proteomics strategies have thereby emerged as the preferred strategy to explore proteomes in combination with tandem mass spectrometry (MS/MS). Both gel-based and chromatography-based methods have been developed to separate proteins and peptides, respectively, to reduce the complexity of a proteome mixture. The oldest gel-based method is 2D gel-electrophoresis (2DGE), and it remains highly used because of the ability to visualize modified proteins and protein isoforms efficiently.2,3 Two commonly applied chromatography-based methods, MUDPIT (multidimensional protein identification technology)4 and GeLC-MS/MS (in-gel tryptic digestion followed by liquid chromatography-MS/MS),5 illustrate the capability of proteomics to create comprehensive proteome maps, rivaling even the sensitivity of tandem affinity purification and the ability of proteomics to create comprehensive proteome maps, essentially complementing the primary blind spot of shotgun proteomics.6

Although the aforementioned proteomics methods all have a qualitative nature geared toward the identification of proteins in a sample, quantitative proteomics methods add (relative) protein abundance information to a protein’s identity. Quantitative proteomics is typically carried out on samples differing in cellular phenotype (eg, benign versus malignant leukocytes), samples subjected to different stimuli (eg, control versus epidermal growth factor-stimulated samples), samples tracked through time (eg, patient treatments), and generally any situation where a differential protein distribution might be expected. Briefly, 2 groups of quantitative proteomics can be defined.7 The first group, consisting of labeled methods, relies on labeling peptides with a distinctive mass tag for each sample that allows quantitative analysis across separate samples by mass spectrometry analysis in a single analysis. Indeed, the sample-dependent mass difference for each analyte is clearly detected by the mass spectrometer. The second group foregoes labeling and consists of so-called label-free methods. Because peptides remain unlabeled, samples need to be run separately, and the acquired mass spectra need to be aligned between samples postacquisition to provide quantitative information. The following dedicated reviews provide a solid introduction to the many methods and challenges that can be found in quantitative proteomics.8–15

**Proteomics Informatics**

Proteomics experiments typically return very many data points, with the raw data acquired by the mass spectrometers...
as the connecting point between the end of the wet laboratory workflow and the start of the data analysis. Specific to gel-based proteomics are gel image processing software packages like Melanie,16 Delta2D,17 or Phoretix.18 These packages provide solid algorithms to sort out spot detection and background correction within a single gel image or spot matching across multiple experiments. Proteomics methods also use mass spectrometer-dependent and-independent processing software to process raw data, often involving signal processing steps such as peak picking, signal smoothing, noise removal, mass calibration by internal standards (often common contaminants), isotope correction, and charge deconvolution.19

The resulting mass spectrometry and MS/MS spectra can then be analyzed by secondary software tools dedicated to the identification of the peptide or protein that generated the mass spectrum in the instrument. These tools are predominantly represented by database search algorithms, such as SEQUEST,20 MASCOT,21 X!Tandem,22 OMSSA,23 or Crux,24 that initially generate an in silico peptide mixture by digesting each entry in a protein sequence database and then proceed to align each experimental MS/MS spectrum to the theoretical mass spectra computed for this in silico peptide mixture. A score is computed for each experimental-to-theoretical spectrum match, and the highest scores are assigned to the most probable peptide candidates for the experimental MS/MS spectrum.25 Although each method implements a reasonable confidence statistic to discriminate between spurious and meaningful matches, various downstream methods have been developed to complement these database search algorithms and to increase the quality of the reported peptide identifications.26–29

After the peptides have been successfully identified from the acquired MS/MS spectra, they are mapped to their parent proteins, a challenging process known as the protein inference problem.30 Briefly, this peptide-to-protein mapping is again performed using a protein sequence database, like UniProtKB/SwissProt, and (depending heavily on the database used) can result in a 1:1 mapping between peptide and protein. Such informative peptides are not always found; thus, a choice has to be made between reporting either the maximal explanatory list of proteins (ie, the list of all proteins that have at least 1 peptide matching them), or a minimal explanatory protein list (ie, the smallest protein list that can account for all observed peptides).31,32

The resulting list of peptides and proteins can then be further analyzed by an ever-growing list of tertiary software tools for functional analysis, all of which attempt to extract functional information from the lists in terms of the underlying biology. The Cytoscape tool, for instance, visualizes the identified proteins and their known interactions (as obtained from publicly available protein interaction databases) as interactive networks that can be highly customized with biological annotations.33

At this point in the typical proteomics workflow (and by extension, in any high-throughput omics technique in use today), the pivotal function of public data repositories becomes very clear. Indeed, the protein lists resulting from a proteomics experiment become far more interesting when put in context of already-established public knowledge, now readily available in abundance on the Internet.34 For instance, by integrating a list of differentially expressed proteins with Gene Ontology information and protein-protein interaction data, emerging patterns on the constitution of protein complexes or coregulation of protein clusters can be interactively surveyed.35 Reversing this logic, external research can become more valuable if the results can be put in the context of, for instance, protein abundance profiles measured in cardiac muscle. Yet, before such integrative analytics can be achieved, 2 crucial requirements must be fulfilled first: Data standards that make use of controlled vocabularies need to be established, and adequate data management systems need to be developed to handle such standardized results. These highly important topics are discussed in detail in the following sections.

**Proteomics Standards**

After years of advances in mass spectrometry and the related software, a heterogeneous set of file formats emerged for mass spectrometry data and protein identification results, which became cumbersome to handle and integrate for downstream software applications and end users alike. Indeed, in any endeavor, it is crucial to first adopt a common language to guarantee accessible information across different groups and experiments. In the field of proteomics, the Proteomics Standards Initiative of the Human Proteome Organization has developed an invaluable set of public domain, standardized file formats for mass spectrometry data and proteomics informatics results throughout the past decade (http://www.psidev.info).36 For the dissemination of gel-based proteomics results, the GeImL format has recently been published as the result of a community-driven process to capture information about gel composition, electrophoresis, protein detection, image acquisition, and spot excision in a uniform format.37

For the exchange of mass spectrometry data, the mzML standard has been published and widely adopted by most mass spectrometry and software vendors in the field.38 A standard for reporting peptide and protein identifications has also been released in the form of the mzIdentML file format, whereas its sibling format for peptide and protein quantitation, dubbed mzQuantML, is in the final stages of development at the time of this writing.39 Furthermore, a standard for the emerging technique of SRM assays is already being actively pursued in the TraML file format.

Aside from these data exchange formats, the experimental conditions are to be reported as well. For this purpose, the Proteomics Standards Initiative has also published the MIAPE (Minimum Information About a Proteomics Experiment) guidelines, which specify the appropriate level of detail to provide when describing the various aspects of a proteomics experiment and have emerged as a valid materials and methods format for manuscript submission in top-ranking proteomics journals like Proteomics, Molecular and Cellular Proteomics, and Nature Methods.40 Although the core proteomics journals recommend the use of these MIAPE standards, a considerable amount of proteomics data are published by nonproteomics journals that currently remain mostly uninvolved in these standardization initiatives. It is likely, however, that a further maturation of the corresponding infrastructure and end user tools will fuel a growing uptake of dissemination and standardized guidelines from nonproteomics journals as well.
Importantly, at the core of these exchange formats exist a set of controlled vocabularies (ontologies) with semantics that truly enable the exchange of results in a common language (Figure 1). For instance, 9606 is synonymous with human (or for Homo sapiens, for that matter) for many researchers, and this identifier is actually defined in the controlled vocabulary of the NEWT (New Taxonomy Database) that combines taxonomy data specific to UniProtKB/Swiss-Prot with information provided by the National Center for Biotechnology Information taxonomic database.41 One of the strengths of controlled vocabularies, or their more elaborate cousins called ontologies, is that the simple numbers they use to represent a concept allow for many synonyms in use in the life sciences to be grouped together without conflict (eg, human and Homo sapiens). Furthermore, the semantic relationships that exist between individual terms within a controlled vocabulary or ontology immediately provide a staggering amount of relevant contexts to any single term. For instance, in the NEWT database, 9606 (human) IS A (the relationship type) primate, which in turn IS A multicellular eukaryote (using the transitivity of the many layers of relationships to skip a large number of ever-coarser taxonomic subdivisions). The NEWT controlled vocabulary thus enables a unified language to communicate highly efficiently about species in an evolutionary context. Similarly, the BRENDA (Braunschweig Enzyme Database) tissue controlled vocabulary provides a unified language and corresponding semantic relationships for eukaryote tissues.42 The BRENDA term 0000199, for instance, is synonymous with cardiac muscle, and the semantic relationships of this entity reveal that this tissue is PART OF both heart and the muscular system. On the other hand, 0000199 is synonymous with cardiac muscle, which is again, PART OF both heart and the muscular system yet represents a completely different type of tissue. Another controlled vocabulary has been established by the Proteomics Standards Initiative for reporting mass spectrometry data, enabling researchers to accurately annotate their experimental results, such as describing the type of ion source or the software used to generate the proteomics results. Data annotated in this way dramatically facilitate downstream integration of multiple proteomics experiments, such as comparing protein identification lists per mass spectrometer type or ion source.

The best-known controlled vocabulary might well be the Gene Ontology, which represents the biological process, cellular location, or molecular function of genes and their products.43 The cardiovascular Gene Ontology annotation initiative is a main contributor to the project and specifically annotates genes relevant to the cardiovascular system based on experimental literature.44

Obviously, one does not need to learn controlled vocabulary identifiers by rote. Controlled vocabularies have been successfully established for various other biomedical domains and are typically managed by the Open Biomedical Ontology Foundry.45 All the controlled vocabularies supported by this foundation can be searched or navigated using the Ontology Lookup Service so that accurate descriptions can be easily located to annotate the experiment.46

The benefits of standardized data formats and controlled vocabularies for annotation are realized at 2 distinct stages. First, within a laboratory, the standardization of data allows...
Managing Proteomics Data

A more generic consideration in any proteomics data analysis workflow is data management, and quite a few solutions have been developed to handle this task. Two levels of data management can be distinguished: local and public management (Table 1).

The first level typically involves a laboratory information management system implementation. Such systems track and register data actions (e.g., storing MS/MS spectra and peptide identifications) to create a queryable, historical log for the experimental results. Such a log can in turn be used to monitor who performed which action at what time, a pivotal feature for data provenance. Moreover, a convenient access point is thereby created for future data and results retrieval. Through automation, data management greatly facilitates the implementation and adoption of standardized data processing workflows, which in turn result in a substantial net increase in productivity. Finally, organizing data from different experiments in an identical manner enables powerful meta-analyses among different experimental data sets.47–49 An excellent summary of currently available proteomics data management systems for local use can be found in 2 recent reviews.50,51

The second level to data management involves the propagation of proteomics results into public repositories. This level is of growing importance because journals are increasingly demanding that proteomics data be made publicly available through such repositories in support of submitted manuscripts, thereby enabling both public evaluation and integration of the reported results. A variety of repositories exists today for managing both gel-based and chromatography-based proteomics data. The World-2DPage Constellation plays a key role in the dissemination of gel-based proteomics results through the World-2DPage portal, which connects a federation of Make2D-DB implementing databases.52 Furthermore, the World-2DPage repository enables data propagation to a central server and thereby allows researchers to avoid the overhead of setting up a local database. Gel images in the system are annotated by the protein identifications and brief identification information, and the Web site allows the user to search for proteins across many experiments.

The proteomics mass spectrometry databases handle instrument data and their derived peptide and protein identifications. The most prominent ones are the Proteomics Identification repository (PRIDE) at the European Bioinformatics Institute in Cambridge, United Kingdom,53 and PeptideAtlas at the Institute for Systems Biology in Seattle, Washington.54 PRIDE experiments can reference protein identifications to gel image coordinates in World-2DPage, which is a convenient example of synergy across public data repositories. Whereas PeptideAtlas reanalyzes the mass spectrometry data for each release, the PRIDE database stores the protein and peptide identifications as submitted by the user. Although some local data management systems support direct output to one or more public repositories,55 overall export functionality remains scarce. The popular PRIDE Converter application features a simple, wizard-like interface to funnel proteomics data from various software formats into the PRIDE format and allows annotation of the experiment using the aforementioned controlled vocabularies.56 Such conversion solutions remain suboptimal, however, because good local data management systems capture a considerable amount of additional, highly relevant data and metadata annotations on a project and thus comprise the best possible point of origin for a submission to a public repository that fully adopts the recently established standards and controlled vocabularies. It is therefore certain that this integration between local and public data management systems must further evolve in the near future because this will enormously benefit the community over time.

Different software packages from different vendors or academic laboratories to work on the same data without conversion or incompatibility issues. It also allows structured storage and retrieval of locally produced data (see next section). Second, it allows publicly shared data to be efficiently localized, thoroughly understood, and effectively reused.

### Table 1. Overview of Local and Public Data Management Systems and Online Coordinates

<table>
<thead>
<tr>
<th>Local data management</th>
<th>Web Site</th>
</tr>
</thead>
</table>

ms-lims indicates mass spectrometry laboratory information management system; MASPCTERAS, Mass Spectrometry Analysis System; SBEAMS, Systems Biology Experiment Analysis Management System; PRIDE, Proteomics Identification repository; LIPAGE, Laboratory Information Management System for 2D PAGE-Based Proteomics Workflow; gpmdb, Global Proteome Machine Database.

different software packages from different vendors or academic laboratories to work on the same data without conversion or incompatibility issues. It also allows structured storage and retrieval of locally produced data (see next section). Second, it allows publicly shared data to be efficiently localized, thoroughly understood, and effectively reused.

---

**Managing Proteomics Data**

A more generic consideration in any proteomics data analysis workflow is data management, and quite a few solutions have been developed to handle this task. Two levels of data management can be distinguished: local and public management of proteomics data (Table 1).

The first level typically involves a laboratory information management system implementation. Such systems track and register data actions (e.g., storing MS/MS spectra and peptide identifications) to create a queryable, historical log for the experimental results. Such a log can in turn be used to monitor who performed which action at what time, a pivotal feature for data provenance. Moreover, a convenient access point is thereby created for future data and results retrieval. Through automation, data management greatly facilitates the implementation and adoption of standardized data processing workflows, which in turn result in a substantial net increase in productivity. Finally, organizing data from different experiments in an identical manner enables powerful meta-analyses among distinct experiments.47–49 An excellent summary of currently available proteomics data management systems for local use can be found in 2 recent reviews.50,51

The second level to data management involves the propagation of proteomics results into public repositories. This level is of growing importance because journals are increasingly demanding that proteomics data be made publicly available through such repositories in support of submitted manuscripts, thereby enabling both public evaluation and integration of the reported results. A variety of repositories exists today for managing both gel-based and chromatography-based proteomics data. The World-2DPage Constellation plays a key role in the dissemination of gel-based proteomics results through the World-2DPage portal, which connects a federation of Make2D-DB implementing databases.52 Furthermore, the World-2DPage repository enables data propagation to a central server and thereby allows researchers to avoid the overhead of setting up a local database. Gel images in the system are annotated by the protein identifications and brief identification information, and the Web site allows the user to search for proteins across many experiments.

The proteomics mass spectrometry databases handle instrument data and their derived peptide and protein identifications. The most prominent ones are the Proteomics Identification repository (PRIDE) at the European Bioinformatics Institute in Cambridge, United Kingdom,53 and PeptideAtlas at the Institute for Systems Biology in Seattle, Washington.54 PRIDE experiments can reference protein identifications to gel image coordinates in World-2DPage, which is a convenient example of synergy across public data repositories. Whereas PeptideAtlas reanalyzes the mass spectrometry data for each release, the PRIDE database stores the protein and peptide identifications as submitted by the user. Although some local data management systems support direct output to one or more public repositories,55 overall export functionality remains scarce. The popular PRIDE Converter application features a simple, wizard-like interface to funnel proteomics data from various software formats into the PRIDE format and allows annotation of the experiment using the aforementioned controlled vocabularies.56 Such conversion solutions remain suboptimal, however, because good local data management systems capture a considerable amount of additional, highly relevant data and metadata annotations on a project and thus comprise the best possible point of origin for a submission to a public repository that fully adopts the recently established standards and controlled vocabularies. It is therefore certain that this integration between local and public data management systems must further evolve in the near future because this will enormously benefit the community over time.

Given that proteomics standards and repositories have been established only relatively recently, we believe that the field is actually progressing quite well. The field had to overcome initial problems with the existence of overlapping standards (notably mzData and mzXML for mass spectrometry) and an immature submission process to the data repositories. Meanwhile, how-
ever, mzData and mzXML were merged into a single standard called mzXML, and tools like PRIDE Converter were created to assist data annotation and submission. Still, as new methodologies continuously arise from innovative proteomics research, growing bodies of controlled vocabularies will remain essential to elaborate experiment annotation. Additionally, the increasing amount of data produced by state-of-the-art mass spectrometers will put a strain on network capacity and data handling. Regardless, with public data sharing currently recommended by core proteomics journals, data dissemination will undoubtedly increase and further drive the maturation of these processes until most inaccuracies are resolved. Ultimately, the goal is that the results from proteomics experiments become readily accessible, even by nonproteomics researchers.

## Integrating Proteomics Data With Cardiovascular Disease

In the past, there have been various studies in which public proteomics results have been integrated with particular relevance to cardiovascular disease. McGregor and Dunn\(^2\) reviewed multiple studies on dilated cardiomyopathy and suggested that altered protein expression be categorized into 3 functional groups by cytoskeletal organization, energy production, and stress response. Alternatively, Bousette et al\(^{57}\) recently integrated proteomics experimental proteomics data with public transcriptome resources to define clusters of protein expression that are specific to heart tissue.

To estimate the number of cardiovascular public proteomics resources, we have manually evaluated PRIDE. Making use of the experimental meta-information that accompanies PRIDE experiments, we were able to select experiments from heart tissue, blood, and cardiac muscle cells from humans and other model organisms (Table 2). Because blood can be easily extracted from humans, both serum and plasma sample preparations represent by far the largest volume of human cardiovascular data. On the other hand, tens of analyses are available for heart tissue from a variety of model organisms. Yet, given that >500 manuscripts have been published on cardiac proteomics since the first publication of the PRIDE repository, one can say that proteomics data dissemination from the cardiovascular community is still in its infancy.\(^{58}\) As described previously, a collection of standards and software tools have been developed since then, and the next generation of submission tools is under active development at the time of this writing. Therefore, we expect that when all tools and submission processes are in place, proteomics data repositories will gradually become more complete and will better reflect the number of cardiovascular proteomics manuscripts published each year.

<table>
<thead>
<tr>
<th>Tissue/Cell</th>
<th>Animal</th>
<th>Experiment Count</th>
<th>Spectra Sum</th>
<th>Peptide Sum</th>
<th>Protein Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart (tissue)</td>
<td>Homo sapiens (human)</td>
<td>1</td>
<td>0</td>
<td>3974</td>
<td>413</td>
</tr>
<tr>
<td></td>
<td>Bos taurus (bovine)</td>
<td>6</td>
<td>2785</td>
<td>2785</td>
<td>1171</td>
</tr>
<tr>
<td></td>
<td>Rattus norvegicus (rat)</td>
<td>18</td>
<td>6458</td>
<td>6458</td>
<td>2301</td>
</tr>
<tr>
<td></td>
<td>Mus musculus (mouse)</td>
<td>3</td>
<td>161 854</td>
<td>44 119</td>
<td>6344</td>
</tr>
<tr>
<td>Cardiac muscle (cell)</td>
<td>Rattus norvegicus (rat)</td>
<td>156</td>
<td>566 107</td>
<td>384 373</td>
<td>251 891</td>
</tr>
<tr>
<td>Blood (tissue)</td>
<td>Homo sapiens (human)</td>
<td>725</td>
<td>9 706 103</td>
<td>2 154 103</td>
<td>479 926</td>
</tr>
<tr>
<td></td>
<td>Mus musculus (mouse)</td>
<td>7</td>
<td>360</td>
<td>3117</td>
<td>304</td>
</tr>
</tbody>
</table>

Species and sample types are combined and display the total count of PRIDE experiments, tandem mass spectrometry spectra, and peptide and protein identifications. It is clear that most human data are derived from blood samples, and various heart tissue data sets are available for distinct model organisms. PRIDE indicates Proteomics Identification repository.

### Conclusion

Proteomics methodologies have come a long way in the past decade, having now firmly established a robust platform for
high-throughput discovery and validation of promising targets and markers involved in disease and fundamental biology. Furthermore, the required (bio-)informatics infrastructure has matured alongside the methods and instruments, providing sophisticated data processing approaches that can be used efficiently and effectively to turn the large amounts of acquired mass spectrometry data into reliable and confident peptide and protein lists and corresponding quantifications. Finally, with standardization and public data dissemination now established, the results obtained can be studied in a larger context through public repositories, allowing the spanning of multiple levels of biological information, moving seamlessly from the genome to the proteome. With such solid backing, proteomics is certain to play a key role in the future of the biomedical sciences.

Acknowledgments
We thank Dr Joël Vandekerckhove for his support. Dr Helsens would like to thank the European Molecular Biology Organization for its short-term fellowship support.

Sources of Funding
Dr Helsens is a postdoctoral fellow of the Research Foundation-Flanders (FWO-Vlaanderen).

Disclosures
None.

References


**Key Words:** proteomics ■ computational biology ■ database management systems