Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics

Eric W. Deutsch,1 Henry Lam,1 and Ruedi Aebersold1,2,3,4

1Institute for Systems Biology, Seattle, Washington; and 2Institute of Molecular Systems Biology, ETH Zurich, 3Faculty of Sciences, University of Zurich, and 4Center for Systems Physiology and Metabolic Diseases, Zurich, Switzerland

Submitted 20 December 2007; accepted in final form 18 January 2008

Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiol Genomics 33: 18–25, 2008. First published January 22, 2008; doi:10.1152/physiolgenomics.00298.2007.—Data processing is a central and critical component of a successful proteomics experiment, and is often the most time-consuming step. There have been considerable advances in the field of proteomics informatics in the past 5 years, spurred mainly by free and open-source software tools. Along with the gains afforded by new software, the benefits of making raw data and processed results freely available to the community in data repositories are finally in evidence. In this review, we provide an overview of the general analysis approaches, software tools, and repositories that are enabling successful proteomics research via tandem mass spectrometry.

Although earlier techniques such as peptide mass fingerprinting demonstrated that mass spectrometers were useful instruments for exploring the protein content of biological samples, it was the advent of tandem mass spectrometry (MS/MS) that enabled the identification of large numbers of proteins in a high-throughput manner. The technique is often called shotgun proteomics, as it is reminiscent of genomic shotgun sequencing.

The typical experimental workflow begins with isolation of proteins from the sample or samples of interest. In gel-based workflows, the proteins are then separated on either a one-dimensional or a two-dimensional gel via electrophoresis. In a somewhat labor-intensive process, spots or lanes of interest are then cut out of the gel, and the proteins within are digested into shorter peptides with an enzyme such as trypsin. Newer, high-throughput techniques instead rely on high-performance liquid chromatography coupled directly to the mass spectrometer. The protein mixture is simply digested with an enzyme, and the resulting peptides are separated in one or more liquid chromatography columns. In the most typical setup, fractionation by ion exchange or isoelectric focusing is followed by reverse-phase chromatography. It is important to use such separation techniques to reduce the sample complexity so that only a handful of peptides are introduced into the mass spectrometer any given time, lest the instrument be overwhelmed by the most abundant species and be unable to measure the less abundant ones. These separated peptides are then most commonly injected into the instrument via electrospray ionization from a reverse-phase liquid chromatography column or by laser pulses on a matrix-assisted laser desorption ionization (MALDI) plate onto which the analytes have been spotted.

As the peptides are injected into the mass spectrometer, the instrument first acquires a precursor ion scan, wherein each intact peptide ion produces a peak in the mass spectrum. The instrument then dynamically selects one or more of those peaks to isolate and subject to collision-induced fragmentation. A mass spectrum of the fragment ions, known as a tandem mass (MS/MS) spectrum, is obtained for each selected precursor. All the mass spectra are written to a file or loaded into a database and are then further analyzed, using the informatics techniques described below to identify the peptides and proteins present in the sample.

Several recent articles have reviewed the basics of shotgun proteomics techniques and instrumentation (14, 15, 35), as well as identification and validation informatics of MS/MS data (42). A great resource for finding software tools for proteomics is the www.proteomecommons.org web site. Here we discuss the entire life span of MS/MS data, from raw mass spectrometer output, through analysis and validation of the data, and finally transfer to public proteomics data repositories, highlighting some successes and lessons learned that may well be relevant to other disciplines of biomedical informatics.

Tandem Mass Spectrometry Proteomics Workflow

Although there are a number of variants and side steps that are used for some experiments, optimal analysis of shotgun proteomics experimental data will usually involve most or all of these eight steps: 1) conversion to and processing via open data formats, 2) spectrum identification with a search engine, 3) validation of putative identifications, 4) protein inference, 5) quantification, 6) organization in local data management systems, 7) interpretation of the protein lists, 8) transfer to public data repositories. The overall workflow is represented in Fig. 1.

For each component in the workflow, the rationale is first described, demonstrating its significance and importance for proteomics, followed by a brief overview of the open-source...
tools available. For several components, general lessons learned will be drawn with the hope that they may provide insight for the challenges faced by researchers analyzing other kinds of life sciences experimental data.

**Conversion to and Processing via Open Data Formats**

Nearly every mass spectrometer vendor has developed its own proprietary data format for storing the data acquired in a mass spectrometer. It is impractical for software tools developed for general use to support all these different formats. Thus the approach in proteomics has been to employ open, XML-based formats and develop the software to work with these formats. It is also common to extract the fragmentation spectra (only) into peak lists in plain text formats such as dta, pkl, and mgf before processing them with a sequence search engine, although most newer engines work directly from the XML files.

However, these plain text files do not contain precursor scans, and thus when quantification using precursor scans is used, these formats are insufficient.

To build quantification software that would work for data from instruments from multiple vendors, an open XML format for storing data generated by mass spectrometry runs (typically of several thousand fragment ion spectra generated over 1–2 h) was developed (47) and was quickly adopted by many in the community, primarily because of the software tools that were made available at publication, including tools to convert vendor format files to mzXML such as Mascot Distiller (12) plus ReAdW, massWolf, and mzWiff (40), and software libraries that simplify the development of programs that read these formats such as RAMP and JRAP (40). At the same time, the mzData format (38) was developed with the primary intent of creating an open format for the durable archiving of raw mass spectrometry data in public repositories. This format was slower to catch on, primarily because few tools supported it when it was released. It is thus a critical lesson that standards or open formats must be accompanied by software tools that can read and write them if they are to gain wide usage. Nevertheless, because of the openness of the mzXML and mzData formats, software tools implementing either or both of these formats are essentially vendor neutral, a significant step forward for the field.

However, the community now has two different formats for essentially the same purpose, which caused considerable confusion about which format to use, as well as extra programming effort to support both formats. It was decided that a new format would be developed with full participation by the community, merging the best aspects from these two original formats into a single format that would be used by everyone. The new format, named mzML, is nearing completion and is expected to be released in early 2008 (39). Even as it is still in beta stage, there are several converters to mzML and validators available to ensure that the format is correctly used. Since there will already be software available for it when the format is released, its adoption will likely be rapid.

Such a circuitous path of reaching a consensus data format with community-wide support is by no means unique. In the field of mRNA microarrays, several formats were in wide use before the community came together and developed a single data model, the Microarray Gene Expression Object Model (MGEXOM) (46). Conversion to and Processing via Open Data Formats

![Fig. 1. Schematic overview of the typical workflow of the proteomics informatics processing of a data set. Not all data sets go through all steps, but most data sets will benefit from some aspect of processing from most of the steps.](image-url)
Spectrum Identification With A Search Engine

The tandem mass spectra generated in a shotgun proteomics experiment are typically identified to their respective originating peptides by a method called sequence searching. In this method, a sequence search engine tries to match these observed spectra with theoretical fragment ion spectra generated from a list of all possible protein sequences. The usual approach for sequence search engines is to iterate over each observed MS/MS spectrum and, for each spectrum, scan through a FASTA file of protein sequences, selecting only those peptides that would have the same precursor mass-to-charge ratio (m/z) within a given tolerance. For this subset of candidate peptides, theoretical MS/MS spectra are generated and compared to the observed spectrum and a score is calculated to quantify how well each theoretical spectrum matches the observed spectrum. These scoring methods vary considerably between engines. The top scoring peptide is then reported together with various similarity and significance metrics unique to each search engine.

A list of these tools, as well as their references, is provided in a recent review of MS/MS analysis and validation strategies (42). Approximately half are commercial offerings, and the other half are free and open-source software (FOSS) tools. The most widely used sequence search engines are Mascot (48), SEQUEST (18), and X! Tandem (5), the first two of which are commercial products. Regardless of the search engine used, the fraction of spectra identified is usually quite low, with 10–20% being typical, although some high-quality experiments do yield as high as 50% identifications. This is largely dependent on the quality of the sample, the type of mass spectrometer used, and the techniques of the instrument operator. Most of the search engines appear to agree on ~80% of identifications that are made, but for the remaining 20% different search engines will identify different spectra with high scores.

However, sequence searching is largely a memoryless and brute-force process, in which all putative peptides in all protein sequences are considered candidates for matching, and the theoretical spectra to be compared are generated anew every time by simple rules of thumb. This process is therefore very inefficient, as it does not make use of the information available from previous experiments. Recently a new class of search engine has become available, instances of which attempt to match new observed spectra with a library of consensus spectra derived from previous identifications. These spectral search engines can typically identify a greater number of spectra with much higher sensitivity at a given false discovery rate (FDR), in a fraction of the time taken in sequence searching. By limiting the candidates to only previously observed peptides, the spectral search engine is much more efficient; by matching against real previously observed spectra rather than simplistic theoretical ones, it is also more precise and accurate than sequence search engines (29). Of course, in this approach, one can only identify spectra that have been observed before. However, with the advent of proteomics data repositories and large-scale spectrum library building efforts, one can easily and very quickly identify a large number of spectra within an experiment, leaving the smaller number of remaining spectra to be searched with traditional sequence searching tools. The currently available spectral search engines are all FOSS: BiblioSpec (20), Bonanza, SpectraST (29), and X! Hunter (6). SpectraST is able to use the spectrum libraries distributed by the National Institute of Standards and Technology (NIST) and is also integrated with the TPP.

A third class of search engine is a de novo sequencing engine. This type of software attempts to identify an MS/MS spectrum without any prior knowledge, neither previously observed spectra nor even input protein sequences. A de novo tool simply tries to read off the sequence directly from the spacing of the peaks. However, for this to work reliably, the MS/MS spectra typically need to be of high resolution and exceptional quality, and such tools are commonly used only in special circumstances where no protein sequences are available, or in conjunction with other techniques. Such hybrid approaches generate short sequences (called tags) of a few consecutive amino acids that can be determined de novo and then use these tags to drastically limit the search space for a conventional sequence search. In other words, only peptides that fall within the correct mass range and contain one of the possible tags determined from the de novo part will be considered.

As additional search engines become available and computers get faster, it becomes practical to search the same data with multiple search engines and combine the results. Different search algorithms have different strengths and seem to excel in identifying some spectra where others fail. Thus a more complete result can be achieved by searching with multiple engines, although integrating these multiple results can be a challenge.

Validation of Putative Identifications

Unfortunately, the top-scoring peptide for each spectrum determined by the search engines described in the previous section is not necessarily the correct identification. The remaining challenge is to determine whether the putative identification is in fact correct. Five years ago, the standard procedure was to apply an arbitrary score cutoff and manually inspect spectra and judge correctness. This was extremely labor intensive and subject to interpreter variability and thus not repeatable. Recently several main techniques have emerged for validating the search results and assigning a FDR for a given threshold.

One popular technique is decoy searching, which has recently been examined in detail by Elias and Gygi (16). There are several variants, but one representative technique is to
reverse all the protein sequences in a protein database and append the reversed decoy proteins to the target proteins. The search engine will thus be challenged with answers that cannot possibly be correct, and the frequency of matches to decoy proteins allows one to estimate the specificity of the search. When one then applies a score threshold it is straightforward to count the number of identifications to decoy peptides, and then, with the assumption that there will be a similar number of incorrect identifications to target peptides, the FDR can be calculated. This approach is easy to implement and fairly robust, although disagreement remains as to how best to generate decoy protein databases. A notable shortcoming of this approach is that the search space and the search time are doubled.

Some search engines include a statistical scoring mechanism within the search engine itself. For example, Mascot uses the probability-based Mowse scoring algorithm (45), which yields a score based on the probability that the top hit is a random event. Given an absolute probability that the top match is random, and knowing the size of the sequence database being searched, the engine calculates an objective measure of the significance of the result. X! Tandem similarly calculates an expectation value that the top match for a spectrum is a random event.

A third technique is to model the distribution of scores as a mixture of two populations, correct and incorrect assignments. This requires that the search scores allow significant separation of the two distributions. Then, based on the mixture model, probabilities of correctness to all identifications can be calculated, together with global FDRs for a given probability threshold. PeptideProphet (27) is the most popular of the tools to take this approach. It has the advantage that decoy sequences are unnecessary; the resulting smaller search space decreases search time and makes it more likely that the search engine will find the correct match. Furthermore, PeptideProphet incorporates many additional factors into the models including the number of enzymatic cleavage termini, the number of missed internal cleavage sites, the observed mass deviation from the expected value, and, more recently, retention time and isoelectric point information. This additional information allows better discrimination and thus increases the sensitivity at a constant FDR.

In the early days of proteomics, statistical validation of search results was often overlooked, making it difficult to assess the quality of published data. Consequently it becomes problematic to combine or compare experiments to evaluate protocols or to gain biological insight. Today, postsearch validation still does not enjoy universal application, but its importance has been recognized by most researchers and codified in the editorial policies of some leading journals. The growing emphasis on statistical validation of data also mirrors that in the field of genomics.

It is usually the case that the analysis of a single biological sample will yield multiple files before the searching step and must be combined into a single data set, usually before the validation step. Most samples are too complex to be introduced into an instrument in a single run, and therefore the sample is often separated into multiple fractions, typically via cation exchange, isoelectric focusing, or protein molecular weight stratification. Furthermore, performing several technical replicates will increase the completeness of the protein list detected for a sample, since any single analysis without replicates will usually yield an incomplete result. Biological replicates are, of course, usually part of a good experimental design.

### Protein Inference

The process described thus far yields identifications of peptides, but in order to understand the nature of the sample, it is necessary to infer which proteins are present in the sample. For simple organisms where most peptides map uniquely to one protein, this is straightforward. However, in higher eukaryotes frequent occurrences of homologous proteins, protein families, multiple isoforms, and even redundant entries in the reference database yield complex and often indiscernible mappings between peptides and proteins. This difficulty is exacerbated by our incomplete understanding of the transcriptome and proteome of most species. Several strategies are used to perform this inference, the most advanced of which is the ProteinProphet software (41), which is also part of the TPP. ProteinProphet calculates probabilities of each protein being present in the sample based on probabilities of the associated peptides being present and uses Occam’s razor rules to reduce the protein list to the minimal set that can explain all the observed peptides.

An important step when determining the final list of proteins in the sample analyzed is calculation of an accurate FDR at the protein level. This is calculated by ProteinProphet as part of its modeling of the distribution of correctly and incorrectly identified proteins. The decoy approach may also be used, by which one can simply count the number of decoy proteins that pass the selection criteria or threshold and thereby calculate the FDR.

### Quantification

After the list of proteins in a sample is determined, it is often critical to measure and compare protein abundances in order to study perturbations of a biological system or discern differences between samples and controls. Several techniques for labeling of peptides are available [see review by Mirza and Olivier in *Physiological Genomics* (37a)], and several software tools have emerged to properly analyze the resulting data (42), nearly all FOSS. The whole field of quantitative mass spectrometry in proteomics has recently been reviewed by Bantscheff et al. (1).

Several quantification tools are integrated in the TPP (26). The XPRESS tool was the first to support ICAT labeling, in which the cysteinyl residues in two different samples are labeled with either a light or a heavy isotope. Given an identification of an MS/MS spectrum, the XPRESS program integrates the intensities in the elution profile of both the heavy and light forms of the peptide to derive a relative intensity ratio. A subsequent and more advanced tool, ASAPRatio (32), provides similar functionality but adds support for additional labeling technologies such as SILAC (43, 44). ASAPRatio incorporates features to automatically detect and combine ratio measurements for multiple charge states, even when only one of the charge states was positively identified. Furthermore, the abundance ratios for several peptides could be combined to derive a final abundance ratio for the parent protein. An interactive component allows users to examine the integrated profiles and manually adjust the integrated region if necessary.
Similar functionality for $^{18}$O labeling is provided by the Zoom-Quant software (22). Isobaric mass tagging such as with the iTRAQ reagent, which allows up to eight separate samples to be tagged and combined into a single run, can be analyzed with the Libra tool, also distributed with the TPP.

Several label-free techniques for quantification are available, which can be broadly separated into techniques that use integrated intensities from elution profiles of peptide ions and those that use spectral counts for each species to derive abundances. Quantification by measuring the intensity profiles of eluting peptides is most often used with single-stage mass spectrometry (MS1), as described below. Spectral counting has recently gained popularity (21, 33) and essentially involves counting the total number of identified spectra attributed to a given protein. Several adjustments (34) are applied to the raw counts since some peptides are easily detected by typical instrumental setups but other peptides are never detected even if they are quite abundant in the sample. In addition, the dynamic exclusion lists used in most instrumental setups explicitly try to prevent repeated fragmentation of the same peptide, working against the premise of this technique. Nonetheless, it has been shown to be crudely effective. One of its advantages is that special software is not required; one merely counts the number of spectral identifications attributed to each protein. An alternative approach is to use the sum of the probabilities of all the identifications (e.g., from the output of ProteinProphet) rather than just giving each spectrum a weight of 1.

**Organization in Local Data Management Systems**

Storing a few experiments on a file system is fine for a small laboratory, but with the increasing size of experiments and teams of researchers, it becomes important to organize and annotate data within a local laboratory information management system (LIMS) or analysis database. Most of the tools described above are file based, and it is common to keep raw data and files from intermediate processing steps in an ordinary file system. Unfortunately, it is also common for data to become lost and forgotten this way because of the difficulty of accessing and querying the data. Moreover, access control for multiple researchers is often difficult with this file-based informatics system.

To address these problems, several proteomics data organization and analysis applications based on relational database systems have emerged. Both commercial and FOSS applications are available. Because of the complexity of such systems, however, none of the systems has achieved widespread installation across many sites. Nearly all were developed around the workflow commonly practiced at the originating institution and therefore incorporate different features. CPAS (53), Proteus (52), SBEAMS (37), Elucidator (17), and YPED (57) are designed to work in conjunction with the TPP processing workflow. Proteios has good support for gel-based proteomics (30). DBParser (63) and numerous other applications (62) provide a simple interface for a specific search engine or workflow. Most of these applications provide similar functionality, and one can choose which to install based on the subset of offered features that seem most relevant. Note that while this is presented as the sixth step here, it is common for the data to be shepherded into the LIMS even before the conversion to open formats and all of the steps listed occur within the database system.

**Interpretation of the Protein Lists**

Once the above data processing steps are complete, the data may then be used to address the biological questions of relevance in the study at hand. This may involve drawing conclusions from the nature of the derived protein list, from the protein abundance analysis, or from comparisons with other studies. There are a myriad of different ways in which this interpretation can take place, and it would be impossible to cover all avenues here, especially since many employ custom software to perform the analysis. However, a few possibilities are explored below.

A common approach is to take the list of proteins, or a subset of proteins that exhibit differential abundances in the system, and analyze the list based on the functional classifications of those proteins. For example, one can upload a list of proteins to PANTHER (60) or DAVID (23) via a web interface and view the proteins organized into groups of protein families, molecular functions, biological processes, and pathways to discover common threads underlying the proteins of interest.

Another approach is to visualize the protein lists and abundances in the context of known molecular interactions to discover regulation mechanisms. The STRING tool (61) provides a large database of known interactions and web-based tools to map out the interactions for proteins of interest. Cytoscape (54) is a desktop graphical software tool that allows the user great flexibility in visualizing data in the context of molecular interactions or indeed any kind of relationships. It cooperates as part of the Gaggle (55) suite of tools, each of which implements a different type of visualization of functionality with a common mechanism of passing information between the tools. The Protein Function Workbench (49), a new tool designed to manage protein lists and interact with Gaggle applications, is just becoming available.

**Transfer to Public Data Repositories**

Five years ago, the fate of proteomics data, after being analyzed, was to decay gradually and become largely forgotten on the file systems of the laboratory in which they were generated. However, it was recognized that there was considerable value to the community in making the raw data publicly available (51), and the first public proteomics data repository, Open Proteomics Database (OPD) was made available. It contains both raw data and processed results for experiments generated at the University of Texas.

Subsequently, several other repositories have emerged. The PRoteomics IDEntifications database (PRIDE) (25, 36) is designed to capture all the protein and peptide identification lists that accompany published articles in a single queryable resource; even unpublished identifications lists may be submitted. The PeptideAtlas database (9–11, 28) accepts only raw output of mass spectrometers, and all raw data are processed through a uniform pipeline of search software plus validation with the TPP. The results of this processing are coalesced and made available to the community through a series of builds for different organisms or sample types.
The Tranche repository (19, 59) is a distributed file system into which any sort of proteomics data may be uploaded. It is then distributed to other Tranche nodes across the globe and may be downloaded by anyone who has access to the hash key identifiers for the data, which may be kept private or publicly released. The GPMdb (7) collects spectra and identifications that have been uploaded by researchers to a GPM analysis engine and presents the summarized results back to the community. Users who upload data can compare their experimental data with identifications that have been previously made by other scientists.

The benefit of these repositories has been significant. They promote greater accessibility to data during the review process of manuscripts. Additional data analysis research projects are enabled by free access to large amounts of data (59), and users can easily access data sets to compare with their new data. An additional benefit demonstrated in the field of transcriptomics data in public microarray repositories, such as ArrayExpress (4, 46), GEO (2), and SMD (8, 56), is the tendency for articles describing data in such repositories to receive greater numbers of citations (50).

Conclusion

There has been great progress in the field of tandem mass spectrometry proteomics informatics just within the past 5 years. The analysis techniques of the latest generation of software are able to identify and validate those identifications in an automated way better than ever before.

However, there does remain a barrier to many research groups in using the latest tools. In many cases the software works extremely well in the hands of the experts who wrote them or those with access to those experts. However, the software is sometimes not easy to install or use and feels unapproachable to researchers without programmers to assist them. To break down those barriers, more time must be spent making the software accessible by making installations easy and focusing on usability of the user interfaces. This is time-consuming work and often requires different skill sets than the original developers have. This effort can come either from companies aiming to commercialize such software or in the form of grants aimed at improving existing software with a clear potential benefit to the community. Funding agencies such as NIH support such work for software tools with a demonstrated user base.

We have described here the typical life span of tandem mass spectrometry proteomics experimental data. There are other, related types of mass spectrometry proteomics experiments that employ different techniques to generate data. Notably, of current great interest is the single-stage mass spectrometry (MS1) technique, often practiced as a biomarker discovery platform, wherein multiple samples are run through as serial MS runs without triggering MS/MS fragmentation. The resulting map of eluting features is compared among the runs in order to discover changes in protein abundance that might signify important biological differences among the samples. Follow-up analyses may be done with MS/MS techniques to identify the interesting features. Several FOSS tools are available for this kind of analysis, including msInspect (3), PEPeR (24), and SpecArray (31).

A final technique is now just emerging: targeted proteomics. Whereas MS/MS proteomics provides an incomplete survey of the most abundant proteins present in the sample, somewhat analogous to EST sequencing in the transcriptomics field, targeted proteomics can provide quantitative measurements for all the desired proteins relevant to a study, analogous to microarrays in transcriptomics. This technique is called selected reaction monitoring (SRM, or equivalently multiple reaction monitoring, MRM). It is performed by first compiling a list of targets (proteins or peptides) to be measured, along with their expected fragmentation patterns. Then the mass spectrometer is provided a list of pairs of precursor ion m/z values and product (fragment) ion m/z values (the combination of which is called a transition) uniquely identifying these targets to continuously monitor instead of obtaining full MS/MS spectra on the most abundant precursors. For each transition the instrument measures intensity over time, and multiple transitions per peptide and multiple peptides per protein yield replicate measurements for each protein of interest. Targeted proteomics also has its own informatics challenges, although primarily in optimal target selection. See Reference 14 for further details on this strategy.

Many of the lessons learned during the evolution of the MS/MS proteomics informatics field are worth noting for bioinformatics researchers and software developers in other biomedical fields. Most notably this includes 1) conversion of vendor or ad hoc formats to open XML-based formats along with development of good FOSS reader and writer programs/libraries; 2) automated model-based validation of results with the goals of minimizing expensive and variable manual validation; 3) open data repositories where raw and processed data can be stored permanently and redistributed; and 4) free and open-source software that encourages the best techniques to be used by everyone in the field and allows continued evolution of the best analysis techniques. Applying these lessons to other disciplines will surely promote advances similar to those that have flourished in MS/MS proteomics informatics.

GRANTS

This work has been funded in part with funds from the National Heart, Lung, and Blood Institute, under contract no. N01-HV-28179, and from PM50-GM-076547/Center for Systems Biology.

REFERENCES


