Challenges and Opportunities in Proteomics Data Analysis*

Bruno Domon‡§ and Ruedi Aebersold‡¶‖

Accurate, consistent, and transparent data processing and analysis are integral and critical parts of proteomics workflows in general and for biomarker discovery in particular. Definition of common standards for data representation and analysis and the creation of data repositories are essential to compare, exchange, and share data within the community. Current issues in data processing, analysis, and validation are discussed together with opportunities for improving the process in the future and for defining alternative workflows. ***Molecular & Cellular Proteomics 5:1921–1926, 2006.***

Proteomics has undergone tremendous advances over the past few years, and technologies have noticeably matured. Despite these developments, biomarker discovery remains a very challenging task due to the complexity of the samples (*e.g.* serum, other bodily fluids, or tissues) and the wide dynamic range of protein concentrations. To overcome these issues, effective sample preparation (to reduce complexity and to enrich for lower abundance components while depleting the most abundant ones), state-of-the-art mass spectrometry instrumentation, and extensive data processing and data analysis are required. Most of the serum biomarker studies performed to date seem to have converged on a set of proteins that are repeatedly identified in many studies and that represent only a small fraction of the entire blood proteome. At the 2005 American Society for Mass Spectrometry Asilomar Conference several speakers stressed the bottlenecks of current proteomics strategies and the imminent need for standards (and base-line criteria) to allow benchmarking the various approaches and to compare results obtained by different laboratories. The need for thorough data generation was emphasized. It requires rigorous analytical chemistry tools, the use of instrumentation that ensures high data quality, and consistent and transparent analysis of the generated data. Furthermore experimental design should be improved to generate statistically meaningful results, namely by avoiding overfitting data or by using distinct training and validation data sets. Therefore, a successful biomarker discovery program relies on the experimental design, its execution using high performance instrumentation, and the processing and analysis of the data using refined tools. Currently data analysis remains a major bottleneck.

Historically proteomics analyses have focused on the identification of proteins in the context of a specific experiment typically in a single laboratory. More recently, the need for more global, quantitative, and comparative studies has been recognized, and the value of comparing proteomics data across studies and laboratories has been highlighted especially in biomarker studies affecting different disease sites. However, the meaningful comparison, sharing, and exchange of data or analysis results obtained on different platforms or by different laboratories remain cumbersome mainly due to the lack of standards for data formats, data processing parameters, and data quality assessment. The necessity of an integrated pipeline for processing and analysis of complex proteomics data sets has therefore become critical. Here we briefly describe current art in proteomics data analysis and its integration into a continuous linear pipeline while underscoring current issues and pointing out opportunities for the near future.

AN OPEN SOURCE PROTEOMICS DATA ANALYSIS PIPELINE

Processing and analysis of proteomics data is indeed a very complex, multistep process (1–3). The consistent and transparent analysis of LC/MS and LC/MS/MS data requires multiple stages as indicated in Fig. 1. It includes processing of the raw data to extract the relevant signals and information, database searches to assign the spectra to peptide sequences, reassembly *in silico* of the identified peptides into proteins, validation of the search results at the peptide and protein levels, and annotation and storage of the results in a database. This process remains the main bottleneck for many larger proteomics studies. Ideally modules solving each one of these tasks should be integrated into a linear process like the Trans-Proteome Pipeline, which allows smooth processing of the data through the different stages (4).

Several factors can lead to significant improvements in the data analysis, including the full leverage of latest instrument capabilities to generate high quality data, the definition of platform-independent data format, and the use of standardized, transparent, and generally available data analysis protocols and tools that will produce consistent and comparable results. Despite the complexity of the process, it is likely to

From the ‡Institute of Molecular Systems Biology, ETH Zurich and ¶Faculty of Sciences, University of Zurich, CH-8049 Zurich, Switzerland and ‖Institute for Molecular Systems Biology, Seattle, Washington 98103

Received, June 16, 2006, and in revised form, August 8, 2006
Published, MCP Papers in Press, August 9, 2006, DOI 10.1074/mcp.R600012-MCP200

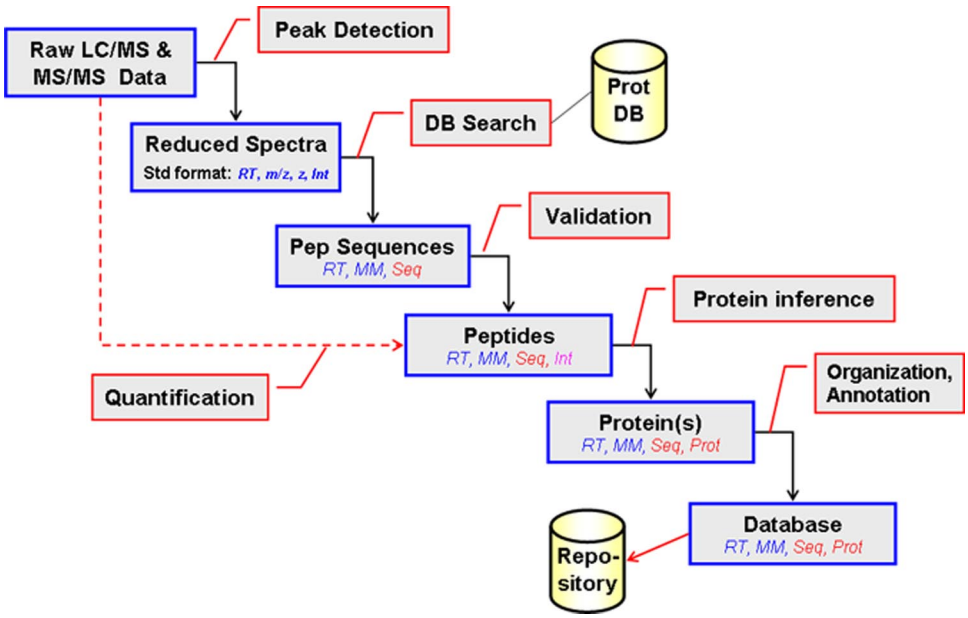

FIG. 1. **Schematic representation of the different modules constituting a data analysis pipeline.** *RT*, retention time; *z*, charge state; *Int*, signal intensity; *Seq*, peptide amino acid sequence; *Prot*, protein accession number and sequence; *DB*, database; *Std*, standard; *MM*, molecular mass.

improve rapidly as the analysis pipeline evolves and data quality increases.

DATA PROCESSING

An important, frequently underestimated element of an integrated data analysis system is the data acquisition and signal preprocessing. A related issue is the format used to capture and store the data.

Signal Processing—Often instruments are operated as a black box and are not always used to the maximum of their performance while data preprocessing is often performed in default mode. For instance, data quality increases significantly (at the expenses of data volume) by acquiring data in profile mode and by subsequently (postacquisition) using more elaborate algorithms to determine signal and noise and derive more accurate measurements. Peak detection (peak picking) is a key element, often neglected, in the data analysis. It is usually part of the instrument software, and the users have limited control over it. This step is often performed automatically during the data acquisition, and the critical parameters are not always explicitly documented. However, high quality data combined with effective preprocessing tools (*i.e.* algorithms for noise reduction, peak detection, and monoisotopic peak determination) are the basis of a reliable data analysis. The maxim "*garbage in-garbage out*" *retains* its full meaning is this context. High quality raw data (*e.g.* profile mode) together with refined peak detection algorithms allow reliable determination of charge state, monoisotopic *m/z* values, and signal intensities of peptide ions. In practice, much better results are obtained by collecting the data first and then reprocessing them off line to fully take advantage of the capabilities of modern instrumentation, which can drastically improve both identification and quantification. It includes higher sensitivity, high resolution, and high mass accuracy, which should be fully retained and exploited during the downstream data analysis. High quality data combined with advanced data processing tools are critical for a deeper insight into proteomics samples in general and in serum or plasma samples more specifically.

Data Format—A large variety of instrument platforms (ion traps, quadrupole/time-of-flight, ion cyclotron resonance, time-of-flight/time-of-flight, etc.) from various manufacturers are available for proteomics studies. Each instrument type will generate spectra with its own characteristics (signal-to-noise ratio, resolution, accuracy, etc.) usually in a proprietary data format. Data processing algorithms are not fully documented and usually are restricted to one instrument platform, thus limiting portability to other data processing tools and comparison of results.

The definition of a generic mass spectrometric data format such as mzXML[1] (5) and the Human Proteome Organization's Proteomics Standards Initiative (6) have been first steps to overcome this problem. The use of a standardized file format allows analyzing data within a pipeline that is independent of the instrument platform. Although the conversion into the mzXML format requires additional computing resources and may increase the file size, a generic format broadly accepted by the community, including the manufacturers, will foster sharing and exchanging data in the future. In this context, the

[1] The abbreviations used are: XML, extensible markup language; MRM, multiple reaction monitoring.

concrete plan to merge the mzXML and mzDATA formats into a single unified file format is encouraging.

PEPTIDE IDENTIFICATION AND VALIDATION

The second main step consists of the assignment of MS/MS spectra to peptide sequences by submitting the MS/MS spectra to a database search using one of several engines available (*e.g.* Sequest, Mascot, Comet, X!tandem, etc.). Most approaches are matching and scoring large sets of experimental spectra with predicted masses of fragment ions of peptide sequences derived from a protein database. Results are scored according to a scheme specific to each search engine that also depends on the database used for the search. Usually tools are linked to one specific platform or were optimized for one instrument type. The various search engines do not yield identical results as they are based on different algorithms and scoring functions, making comparison and integration of results from different studies or experiments tedious. To compare results (and to some extent search engine performances) searches have to be performed under well defined and comparable parameters. Recently published guidelines for database searching have addressed this issue (7, 8). It is critical at this point that the parameters used for the search are fully tracked and documented. Peptide identification via database searches is very computationally intensive and time-demanding. High quality data allow more effective searches due to tighter constrains, *i.e.* tolerance on precursor ion mass and charge state assignment, which will drastically reduce the search time in case of an indexed database. In addition, accurate mass measurements of fragment ions further simplify the database searches and add confidence to the results.

Once the initial output of the database search engine has been obtained, it is essential that the reliability of the assignments of spectra to peptide sequences is statistically validated. Such analyses generate reliable estimates of the false positive and false negative error rates, values that are critical to meaningfully compare results from multiple experiments or platforms. The PeptideProphet algorithm (9) has be designed to achieve this goal.

The error rates in a data set can also be estimated by performing a search using a "reversed database" (*i.e.* a database in which the sequences were scrambled to produce only false positive identifications and thus ascertain the false positive error rate (10)). In contrast to more specialized tools, reversed database search results do not estimate the false negative error rate of a dataset.

An alternate strategy consists of storing MS/MS data in a library at an earlier stage in the identification process. Comparison of MS/MS data occurs by comparing experimental spectra with those previously measured and stored in a database using a spectra-matching algorithm. Such an approach was proven to be very effective for several decades in the small molecule area. Stein *et al.* (11) have explored that route by building a library of consensus peptide spectra (*i.e.* a set of consistent spectra derived from multiple experimental data sets measured on quadrupole ion traps and quadrupole time-of-flight instruments). An extension of the library together with the ability to produce easily comparable results requires some normalization of the parameters used for the data generation (in particular collision energy). Despite this limitation, identification of already known (observed) peptides is much easier and faster than conventional database searches. With adequate search speed, one could even envision on-the-fly deployment of such a tool on current instruments with fast data acquisition to make data-dependent decisions and exclude ions based on their MS/MS signature rather that a single parameter such as the mass of the precursor.

A spectral matching approach is likely to be less biased because of the search engine or the protein database used. The limitations of such an approach are the number and quality of the spectra included in the library, the level of confidence in the peptide sequence assignments (emphasizing the need for a curation mechanism), and the performance of the spectral matching algorithm in minimizing false positive and false negative calls. Such libraries will also be valuable resources for multiple reaction monitoring approaches as they will simplify the selection of the transitions, *i.e.* precursor ions/fragment ion pairs (see below).

PROTEIN IDENTIFICATION AND VALIDATION

The association of identified peptides with their precursor proteins is a very critical and difficult step in shotgun proteomics strategies as many peptides are common to several proteins, thus leading to ambiguous protein assignments. Therefore it becomes critical to have an appropriate tool that is able to assess the validity of the protein inference and associate a probability to it. ProteinProphet combines probabilities assigned to peptides identified by MS/MS to compute accurate probabilities for the proteins present (12).

ProteinProphet weights peptides that have a reliable score/probability and from all corresponding proteins derives the simplest list of proteins that explains the observed peptides (13). Obviously proteins with multiple peptide matches have a much greater confidence in their assignment than proteins identified by one single peptide. In fact, there is a massive amplification of the false positive error rate at the protein level compared with the peptide level. This emphasizes the importance of this step and the need for a tool that is able to predict peptide and protein association (13).

QUANTIFICATION

Quantification is a further critical step in biomarker studies because the primary focus is on peptides (proteins) that show differences in expression between two sets of samples; peptides that are invariant present much less interest. Systematic quantification of all peptides across multiple data sets is

actually a very demanding task that has not yet been fully resolved. Strategies emphasizing the quantitative aspect tend to decouple identification and quantification and perform two independent experiments.

Basically two main approaches have been applied. The first is based on stable isotope labeling and requires derivatization of the peptides from the various samples sets with different reagents that have different isotopic composition. The resulting products are then pooled together and analyzed in one single LC/MS/MS experiment. The relative quantity of a specific analyte is then determined from the relative signal intensity of the signal in the full spectrum. Subsequently the analyte in question is identified by database searching of the corresponding MS/MS signal.

The second approach, which is more relevant to larger biomarker studies (*i.e.* analysis of a larger sets of samples from normal (control) and disease (or treated) patients), analyzes each sample individually and then compares the multiple LC/MS runs subsequently. Performing data acquisition under rigorously controlled conditions and in an unbiased manner is essential for this method. The processing of the multiple data sets raises several important issues, including control of instrumental drifts (mass calibration and elution times) over longer periods of times, correction for shifts in elution times, and normalization of ion abundances to adjust for variation in sample amounts or instrument (ionization or detection) performances. None of these are trivial to solve. In essence, a series of LC/MS patterns are acquired, peaks (or more precisely peak clusters) are detected, and data are merged together. The main step consists in matching the ions observed (*i.e.* within specific *m/z* and retention time tolerances) across all experiments. It is a critical task as the high density of features within a window might results in mismatches that might jeopardize any downstream analysis. Thus high quality data (*i.e.* high mass accuracy and reproducible elution times) are critical to this process. It is typically reflected by a high number of "isolated" features, which are observed in only one or a few experiments. If features are properly matched, the quantitative analysis can be performed in a relatively straightforward manner; a number of tools have been described to perform such analyses (14, 15).

DATA REPOSITORIES

We have already mentioned the importance of standards to store, retrieve, and exchange data and results. Typically proteomics experiments are carried out in isolation by one single laboratory often in an uncoordinated way, thus making sharing and comparison of results tedious if not impossible. The lack of common standards and protocols has lead to this situation and often resulted in duplication of efforts.

Results were usually reported as a set of identified proteins (*i.e.* list of peptides identified and associated proteins) with minimal supporting data. Obviously the large volume of such data sets has made publication of detailed results using classical mechanisms very challenging. Sharing and exchange of data and results requires the definition of standard formats for the data at all levels (including raw mass spectrometric data, processed data, and search results) as well as a better definition (and/or standardization) of the parameters used for the data processing or the database searches. In this way, a broad range of data and results generated by a variety of tools can be captured into widely accessible repositories, including results of database search engines, *de novo* sequencing algorithms, and statistical validation tools (*e.g.* PeptideProphet and ProteinProphet). In all cases metadata describing the parameters used for the analysis are essential and have to be included. Additional data, including annotation and clinical information about the samples, ought to be incorporated in the database as well. Data and results repositories such as PeptideAtlas (Institute for Systems Biology) (16), Global Proteome Machine Database (Beavis Informatics) (17), or Proteomics Identifications (PRIDE) Database (European Bioinformatics Institute) (18) facilitate exchange of results and information. Initially limited to peptide sequences and proteins with a minimum of parameters (*i.e.* retention/elution time, mass, charge state, and signal intensity) these databases are rapidly expanding to include actual spectra, links to the associated proteins, and genomics data.

One can envision expanding the repertoire of parameters captured as long as robust normalization methods are defined. For instance, elution times have limited value unless they are translated into normalized values that can be generalized across the entire database. Defining broad standards that the community can agree to and developing integrated platforms (or at least platforms that can integrate the different modules/tools) are currently two of the main challenges in integrating results from multiple platforms and laboratories. Despite major advances toward the standardization and sharing of data already achieved, issues regarding the protein annotation and accession numbers remain to be addressed.

NEW STRATEGIES

Current shotgun proteomics-based approaches, relying on (random) data-dependent peptide sequencing, have resulted in limited numbers of identified peptides and proteins despite major efforts. Furthermore these peptides and proteins are identified with high redundancy. This has prompted the development of new strategies to overcome this limitation. More directed approaches taking advantage of the quantification tools described above are currently used to generate inclusion lists and then selectively sequence peptide ions that are part of such a list. This, at least in part, overcomes the limitations of data-dependent approaches and allows sequencing peptides with much greater depth.

Proteomics-integrated data analysis pipelines were initially designed to automatically identify peptides and proteins in larger data set. More recently, the interest in important protein functional information has prompted the development of an-

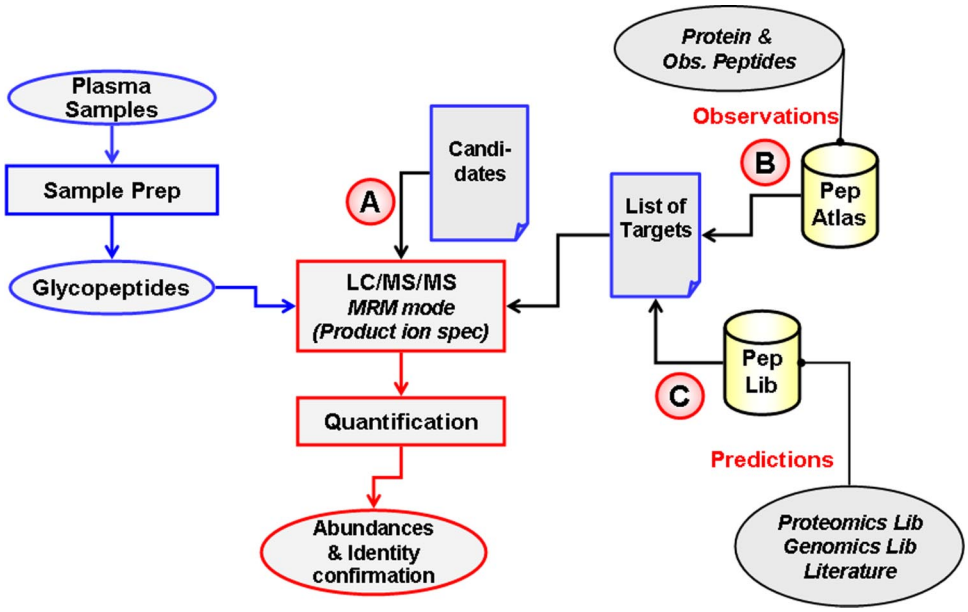

FIG. 2. **Schematic representation of new paradigm for hypothesis-driven proteome analysis.** *A*, conventional MRM approach to evaluate biomarker candidates that were previously identified. *B*, hypothesis-driven screen (discovery) of putative biomarkers based on prior knowledge and information present in proteomics database. *C*, generalized discovery screen based on more general knowledge and predictions of protein (peptides/ions) to be observed. *Prep*, preparation; *Obs.*, observed; *Pep*, peptide; *spec.*, spectra; *Lib*, library.

alytical strategies that also focus on the detection and identification of post-translational modifications (*i.e.* nature of the modification and its localization on the peptidic backbone). Currently most approaches rely on extended database searches that incorporate predefined modifications in the search space. More refined approaches based on *de novo* sequencing or integrating other experimental designs (neutral loss scans and MS^3 experiments) are gaining importance. Also sample preparation to specifically isolate the peptides of interest is usually a critical step in post-translational modifications. It will be critical to expand the current proteomics data repositories and databases to also include high confidence information on modified peptides.

To overcome some of the limitations of current proteomics strategies in regard to the dynamic range of peptides detected and the undersampling of MS/MS spectra that restrict the ability to comprehensively analyze proteomes, alternative mass spectrometry-based approaches are being explored. Conventional data-dependent acquisition is limited by the detection of a signal in full scan mode to trigger a product ion spectrum. In contrast, targeted strategies exemplified by multiple reaction monitoring (MRM) detect, quantify, and possibly collect a product ion spectrum to confirm the identity of a peptide with much greater sensitivity because the precursor ion is not detected in the full mass spectrum. Performed on a triple quadrupole instrument (or hybrid quadrupole/linear ion trap instrument) the data are acquired by setting both mass analyzers to predefined *m/z* values corresponding to the multiply protonated ion and one specific fragment ion of the peptide of interest. The two-level mass filtering drastically increases the selectivity while the non-scanning nature of this experiment accounts for an increased sensitivity. It thus allows detection and quantification of low abundance analytes in complex biological matrices. Obviously MRM experiments differ from conventional approaches in that they are hypothesis-driven (*i.e.* screening for known or putative entities) and are primarily quantitative experiments or possibly confirmatory of identity through matching to already known information (*e.g.* elution time or observed or predicted MS/MS spectra).

Multiple ion monitoring was demonstrated to be a valuable approach to quantify with high sensitivity and selectivity peptides in complex mixtures such as a tryptic digests of plasma samples (19) or a specific subproteome, *e.g.* glycopeptides (20). The recent advances in the software of instrument control and data acquisition now provide the capability to analyze larger number of peptides (>500 transitions) in one single LC/MS run (17). This truly opens new paths toward in-depth analysis of proteomes at dramatically reduced redundancy using a hypothesis-driven approach as illustrated in Fig. 2. The strategy integrates high sensitivity mass spectrometric measurements in the MRM mode and the proteomics knowledge already acquired that is available in databases (*e.g.* PeptideAtlas).

Existing data captured in databases are used to predict proteotypic peptides (*i.e.* sequences that are unique to one single protein), and their corresponding fragmentation patterns are used to define MRM transitions. This enables large screens to detect and identify putative peptides (and thus the inferred proteins) as illustrated in Fig. 2. In such approaches, one starts by generating a list of proteins of interest, the

corresponding proteotypic peptides are derived, and the associated fragment ions are predicted (or extracted from a database) to define the MRM transitions. Relative estimation of the elution times might also be made. All that information will then be used to generate a list of targeted peptides. It is thus possible in one single LC/MS run (and thus with minimal effort) to detect and quantify these peptides.

CONCLUSION

Biomarker discovery projects (as well as many other proteomics studies) are often large experiments generating large data sets, and results might be obtained from concerted efforts of several laboratories. It is essential that data exchange and sharing becomes a transparent process. Standardization of data through wide use of common formats and use of transparent tools for data processing and analysis with well defined parameters is essential.

The data processing and analysis bottleneck can nowadays be overcome through integration of the entire suite of tools into one linear pipeline. This allows processing of data from different instrument platforms in a reliable way through the different steps while maintaining consistency of results. It eases comparison between multiple platforms or laboratories. The organization, annotation, and sharing of data in public repositories will greatly facilitate the data exchange within the community. It should also prevent the replication of experiments that have already been carried out. In this way, ambitious programs such as the creation of a full proteome map of tissues or cell types will be achieved more readily.

The focus of this account has been on the integration of proteomics data. It thus represents the first step for the incorporation of the results in a more global, systems biology framework that includes data from various platforms, including genomics, metabolomics, and physiology. New quantitative proteomics strategies that leverage recent mass spectrometry technologies combined with advanced data analysis tools are anticipated to play a crucial role in the global analysis of biological systems.

* This study was supported in part with federal funds from the NHLBI, National Institutes of Health, under Contract N01-HV-28179. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ To whom correspondence should be addressed: ETH Zurich, Hoenggerberg HPT E73, CH-8093 Zurich, Switzerland. Tel.: 41-44-633-2088; Fax: 41-44-633-1051; E-mail: domon@imsb.biol.ethz.ch.

REFERENCES

1. Kearney, P., and Thibault, P. (2003) Bioinformatics meets proteomics—bridging the gap between mass spectrometry data analysis and cell biology. *J. Bioinform. Comput. Biol.* **1,** 183–200
2. Listgarten, J., and Emili, A. (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434
3. Baldwin, M. A. (2004). Protein identification by mass spectrometry. *Mol. Cell. Proteomics* **3,** 1–9
4. Keller, A., Eng, J., Zhang, N., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1,** 2005.0017
5. Pedrioli, P.G.A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. A., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, Jr R. K., Kapp, E., McComb, M., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. A. (2004) Common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22,** 1459–1466
6. Orchard, S., Kersey, P., Hermjakob, H., and Apweiler, R. (2003). Proteomics Standards Initiative meeting: towards common standards for exchanging proteomics data. *Comp. Funct. Genomics* **4,** 16–19
7. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* **3,** 531–533
8. Bradshaw, R. A., Burlingame, A. L., Carr, S., and Aebersold R. (2006) Reporting protein identification data: the next generation of guidelines. *Mol. Cell. Proteomics* **5,** 787–788
9. Keller, A., Nesvizhskii, A., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
10. Elias, J. E., Hass, W., Faherty, B. L., and Gygi, S. P. (2005). Comparative evaluation of mass spectrometry platforms used in large scale proteomics investigations. *Nat. Methods* **2,** 667–675
11. Stein, S., Kilpatrick, L., Neta, P., and Roth, J. (2005) Building and using reference libraries of peptide mass spectra, in *Proceedings of the 53rd ASMS Conference on Mass Spectrometry, San Antonio, TX, June 5–9, 2005*, A051573, American Society for Mass Spectrometry, Santa Fe, NM
12. Nesvizhskii, A., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
13. Nezvizhsii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomics data: the protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440
14. Lie, X.-J., Kemp, C. J., Zhang, H., and Aebersold, R. (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* **4,** 1328–1340
15. Rauch, A., Bellewm, M., Eng, J., Fitzgibbon, M., Holzman, T., Hussey, P., Igra, M., Maclean, B., Lin, C. W., Detter, A., Fang, R., Faca, V., Gafkenm P., Zhang, H., Whitaker J., States, D., Hanash, D., Paulovich, A., Martin, W., and McInstoshm, M. W. (2006). Computational Proteomics Analysis System (CPAS): an extensible open source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome. Res.* **5,** 112–121
16. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A. Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6,** R9
17. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data *J. Proteome Res.* **3,** 1234–1242
18. Jones, P., Cote, R. G., Martens, L., Quinn, A. F., Taylor, C.F ., Derache, W., Hermjakob, H., and Apweiler, R. (2005). PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Proteomics* **5,** 3537–3545
19. Anderson, N. L., and Hunter, C. L. (2006) Quantitative mass spectrometric MRM assays for major plasma proteins. *Mol. Cell. Proteomics* **5,** 573–588
20. Domon, B., Stahl-Zeng, J., and Aebersold, R. (2006) Novel strategy for rapid screening and quantification of biomarkers in serum, in *Proceedings of the 54th ASMS Conference of Mass Spectrometry, Seattle, WA, May 28–June 1, 2006*, in press, American Society for Mass Spectrometry, Santa Fe, NM