MRMaid 2.0: Mining PRIDE for Evidence-Based SRM Transitions

Jun Fan,¹ Fady Mohareb,¹ Nicholas J. Bond,² Kathryn S. Lilley,² and Conrad Bessant¹

Abstract

Selected reaction monitoring (SRM) is becoming the tool of choice for targeted quantitative proteomics. The fundamental principle of proteomic SRM is that, for a given protein of interest, there is a set of peptides that are unique to that protein. The characteristic retention time (RT), and intact peptide m/z of these so-called proteotypic peptides are then programmed into the mass spectrometer, along with the m/z of high-intensity product ions for targeted quantitation. The particular combination of RT, peptide m/z , and product m/z for a given peptide is referred to as a transition. Selection of the most appropriate set of transitions for a given set of proteins is crucial to any SRM experiment. We previously developed the web-based MRMaid tool, which suggested the optimal transitions for a given human protein by mining spectral evidence from a small in-house database. In this article we present a completely new implementation of MRMaid, which offers substantial improvements over the original. The new version, MRMaid 2.0, uses spectra from the EBI's PRIDE database, which massively increases the coverage and quality of transitions. Transition lists can now be generated for multiple proteins simultaneously, edited within the web browser, and exported for laboratory use.

Introduction

THE TARGETED QUANTITATIVE PROTEOMICS TECHNIQUE of multiple reaction monitoring (MRM), technically known as selected reaction monitoring (SRM), continues to grow in popularity, thanks to its high sensitivity and the fact that it can be used to quantify multiple proteins in a single chromatography run on a relatively inexpensive triple quadrupole mass spectrometer (MS) (Anderson and Hunter, 2006). Applications range from validation of biomarkers discovered in shotgun proteomics experiments (Elschenbroich et al., 2011), through the detection of protein markers in clinical practice (Ang et al., 2011), and the detection of allergens in food products (Johnson et al., 2011). It is also an important tool for systems biology studies, as it can be used to monitor the protein elements of a specific pathway of interest.

The fundamental principle behind SRM is that, for a given protein of interest, a set of peptides (usually tryptic peptides) can be found that are unique to that particular protein. The characteristic retention time (RT) and intact peptide m/z of these so-called proteotypic peptides are then programmed into the MS, along with the m/z of high-intensity product ions for targeted quantitation of the protein of interest. By focusing on specific masses at specific retention times, sensitivity is maximized.

The particular combination of RT, peptide m/z , and product m/z for a given peptide, is referred to as a transition, and selection of the most appropriate set of transitions is the key element in the design of any SRM assay. In recent years a number of freely available software tools have been developed to help experimentalists design the optimal set of transitions (Cham Mead et al., 2010), including our own web-based MRMaid service (Mead et al., 2009). A key feature of many of these tools is the use of previously collected spectral data as evidence for predicting which peptides are likely to be detectable by MS, and which will produce ion series conducive to SRM.

In this article we present a new implementation of MRMaid that brings substantial improvements over the original MRMaid and significant advantages compared to the other available tools. The most important change is that MRMaid now uses spectral evidence from the European Bioinformatics Institute's PRIDE database (Vizcaino et al., 2009), which massively increases the coverage and quality of MRMaid's transitions, ultimately allowing support for species other than the typical trio of human, mouse, and yeast. The user experience has also been substantially improved by pre-computing transitions to improve response time, and implementing a new user interface that supports the generation of transition lists for multiple proteins, and allows this list to be exported

¹Cranfield Bioinformatics Group, Cranfield Health, Cranfield University, Bedfordshire, U.K.

²Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge, U.K.

for use in the laboratory. The culmination of this work is an SRM assay design tool that can be used to produce complex transition lists with minimal effort, with coverage that will grow automatically as the volume and quality of data in PRIDE increases.

Materials and Methods

MRMaid architecture

MRMaid now consists of three main components. The first component is the core of the system—a relational database designed to store transitions. This database is periodically populated with transitions by the second component, a transition builder pipeline. Finally, a web front end at www.mрмаid.info provides a visual interface through which users can retrieve transitions from the database. A schematic overview of MRMaid 2.0 is shown in Figure 1, and the three main components are described in more detail below.

Transition database

The transition database is a relational database implemented in MySQL, with tables for (1) entities including

precursor, product ion, peptide, protein, and instrument type; (2) relationships among entities including peptides in proteins, pairs of precursors, and product ions, the instrument types and the PRIDE experiments from which transitions were observed; and (3) pre-calculated metrics including peptide score, probability of peptide and product ions being observed, and relative product ion intensities (methods for calculating these metrics are described in the next section).

Transitions are stored in the database according to the type of instrument on which the PRIDE evidence for the transition was generated. It was not possible to take the instrument types directly from PRIDE, as we found experiments tagged with 53 individual instrument identifiers. We therefore manually partitioned PRIDE's instrument identifiers according to ionization method (ESI or MALDI), and detector type (time of flight or ion trap), with any other instruments or ambiguous identifiers being relegated to a further, "unclassified", group. Since there is no MALDI-Trap data, the list of instrument types supported by MRMaid at present is ESI-ToF, MALDI-ToF, ESI-Trap, "unclassified," and "all." The latter type is used to label transitions that have been derived from all PRIDE data, regardless of instrument type (i.e., it is a superset of the other four types).

FIG. 1. Schematic overview of MRMaid 2.0. Dark-colored boxes linked by solid arrows indicate components that have been implemented and are described in this article. Lighter boxes and dotted lines show future developments. The core of the new MRMaid implementation is the transition database. This is populated by the transition builder pipeline, using data from PRIDE and UniProt. Data can currently be retrieved from MRMaid's transition database via the web interface, which includes the capability to export transition lists in comma-separated values (CSV) format. In the future we plan to support other formats, including the proposed Proteomics Standards Initiative (PSI) TraML format, and provide an application programming interface (API) for developers to include MRMaid functionality within their own software.

Transition builder pipeline

The transition builder pipeline is used to populate the MRMAid transition database, by mining PRIDE for data from proteotypic peptides identified from UniProt (UniProt Consortium, 2011). This program is designed to run periodically to accommodate updates from PRIDE and UniProt, performing the transition building process automatically. The pipeline comprises four main steps.

Preparation of proteotypic peptide sequences. The human proteomic sequence data is retrieved from UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions). All protein sequences are digested *in silico* to produce tryptic peptides, among which only unique peptides (proteotypic peptides) are retained. Peptides having more than 24 amino acids are then discarded, as these will have masses outside the detection range of MS/MS instruments.

Retrieval of spectral data from PRIDE. The proteotypic peptides are searched in a locally installed PRIDE BioMart (Haider et al., 2009) to find which of those peptides are present in PRIDE, and retrieve their corresponding experiment and spectrum references. The spectral data and related metadata for the mapped proteotypic peptides are retrieved from the public PRIDE MySQL database. In the original MRMAid we used a relatively small but high quality in-house spectral dataset generated by our GAPP identification pipeline (Shadforth et al., 2006). Moving to the PRIDE database gives us access to several orders of magnitude more spectra, but quality is not guaranteed, as the majority of data in PRIDE are not curated. To overcome this issue we filter PRIDE spectra by removing any spectrum with ambiguous metadata (e.g., a single spectrum assigned to more than one peptide), with an inappropriate fragmentation method (we are interested in *b* and *y* ions, so only collision-induced dissociation spectra are considered), with modification (modified peptides are not considered at the moment), and with inconsistent precursor information (which we define as the theoretical molecular weight of a peptide deviating more than 5% from the mass value calculated by multiplying the precursor *m/z* value by the precursor charge).

Conversion of spectral data into transition database. The theoretical *m/z* values for all *b* and *y* ions are calculated from the peptide sequence. For each instrument type, the spectra are parsed to map the theoretical *m/z* values to detect the *b* and *y* ions. For any detected ion, the transition database is populated with the corresponding data for product ion, precursor, instrument type, and peptide.

Calculation of transition metrics. To quantify the suitability of peptides and product ions to be monitored for proteins of interest, several quantitative metrics have been designed. These are:

Peptide score: Called “transition score” in the original MRMAid (Mead et al., 2009), this is a weighted sum of the factors of spectral evidence and characteristics of the peptide sequence, designed to indicate the expected performance of the peptide in SRM. The algorithm for calculating the peptide score is almost the same, except in Equation 2 of the original

article std_{peaks} has been replaced by the root of least square sum of *y* ion numbers. This modification of the scoring was shown to increase calculation efficiency and produce peptide scores that better reflect the suitability of the peptide for SRM. Full details of the peptide scoring algorithm are provided in the supplementary materials (Supplementary Material 1; see online supplementary material at <http://www.liebertonline.com>).

Peptide probability: Denoted as P_{peptide} , this value represents the probability of observing a particular peptide when the parent protein is present. This is calculated by simply dividing the times a peptide has been identified in PRIDE by the number of times the parent protein has been identified. Calculating the number of times a protein has been identified is not trivial because protein presence is inferred from peptides, and the presence of a single protein can be inferred from multiple different proteotypic peptides. The number of times a protein has been identified is therefore estimated as the sum across all PRIDE experiments (for the chosen instrument type) of the maximum times in an experiment that any proteotypic peptide has been seen for the protein in question. This undoubtedly means that P_{peptide} underestimates the probability of observing a given proteotypic peptide, but when used as a relative measure for ranking peptides it proves to be a valuable tool.

Product ion probability: Denoted as P_{product} , this is calculated from the filtered PRIDE data in a similar way to P_{peptide} . It represents the probability of observing a given product ion when the parent peptide is observed across all PRIDE experiments for the chosen instrument type.

Product ion relative intensity: This represents the height of the ion peak in the spectrum. It is calculated by averaging the normalized ion intensity from each spectrum for the peptide in question, across all spectra from the chosen instrument type.

Hydrophobicity is also calculated for each peptide during this process. This is used later to calculate estimated RT by the SSRCalc method (Krokhin et al., 2004).

In summary, the new transition builder algorithm is similar to the original MRMAid algorithm (Cham Mead et al., 2010) at the peptide level, but now also suggests product ions to monitor.

Web interface

The MRMAid web interface has been completely rewritten using the latest web technologies to give users a more intuitive interface with enhanced functionality. Rather than having to specify search options and filters prior to performing the transition search, the search is initiated using only the list of proteins and species of interest. The results are then displayed as a table that can be dynamically refined using interactive filters. The user interface was implemented using a model-controller-viewer infrastructure. The model communicates with the MRMAid database via Java Persistence API (JPA), and is represented as a hierarchical tree having the levels of protein-peptide-instrument-product ion-transition, and Java Server Faces (JSF) serves the role for both controller and viewer. All information in the results table is taken directly from the database, except RT, which is calculated on the fly from hydrophobicity as described earlier.

TABLE 1. VALIDATED TRANSITION LISTS USED FOR BENCHMARKING

Source	Tissue	Instrument	Allocated instrument type	Proteins			
				Total	Uniquely identified	Modified	To be searched
Anderson and Hunter, 2006	Plasma	4000 QTRAP NanoSpray	ESI Trap	53	52	0	52
McKay et al., 2007	Plasma	4000 QTRAP NanoSpray	ESI Trap	18	18	0	18
Kay et al., 2007	Serum	4000 QTRAP NanoSpray	ESI Trap	15	14	0	14
Keshishian et al., 2009	Plasma	4000 QTRAP NanoSpray	ESI Trap	9	8	0	8
Keshishian et al., 2007	Plasma	4000 QTRAP NanoSpray	ESI Trap	1	1	0	1

ESI, electrospray ionization.

Benchmarking comparison with validated MRMs

To test the efficacy of the new PRIDE-based MRMAid implementation we benchmarked MRMAid's results against validated transitions for a number of human proteins that have been published in the literature (Anderson and Hunter, 2006; Kay et al., 2007; Keshishian et al., 2007, 2009; McKay et al., 2007). The lists of validated SRM transitions from these articles were retrieved from the SRMATlas Transition Lists section (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/GetTransitionLists?_subtab=2) of PeptideAtlas (Farrah et al., 2011). The identified proteins in the transition list were searched against UniProt, followed by BLASTing of the peptide sequence if no match was found in UniProt to determine the protein name to be searched for in MRMAid. Transitions for proteins with peptide modifications were not considered because these are not supported in the current MRMAid implementation. The existence and ranks of the validated peptides and the product ions in MRMAid's results were recorded for both the ionization/detector combination used in the papers (all used an ESI-ToF instrument), and "all" instrument type.

Results

In the downloaded UniProt version (December 2010) of the human proteome, there were 20,252 proteins, in which 522,958 unique tryptic peptides with no more than 24 amino acids were found. Within PRIDE, we found 1,061,445 spectra that had been assigned to 50,068 of these proteotypic peptides. These peptides mapped to 10,188 unique proteins. After filtering the spectra, the number of peptides covered was reduced to 23,569, which mapped to 7109 unique proteins. Transitions for these peptides were built with the transition-building pipeline and deposited in MRMAid's transition database.

Web interface

The MRMAid 2.0 web interface is freely accessible via www.mrmaid.info. On the front page there is a drop-down list containing available species and a text area for entering names or accession numbers of proteins of interest. After clicking the "MRMAid Search" button, MRMAid retrieves the relevant transitions from the database and presents these in a dynamic table. In this table, each row represents a product ion for a proteotypic peptide from one of the proteins of interest. The suggested product ions are by default ranked according to the order in which the proteins were entered in the search, the suitability of the peptide for SRM, and the suitability of the

product ion to be monitored, but the rows can be ordered differently by clicking on the column headings. Hovering over a column heading reveals an explanation of that column's content. Clicking on a value in the PRIDE Data column links to the experimental evidence in PRIDE.

The content of the table can be filtered according to instrument type, and there are text fields in the table heading where the user can type values to filter based on the various transition metrics, and by peptide composition. The HPLC conditions used to predict peptide retention times can be changed using controls above the table. It is also possible to specify the maximum number of peptides to be displayed per protein, and the maximum number of product ions to be displayed per peptide. As a rule of thumb it is considered good practice to use five peptides per protein and three transitions per peptide, so these are programmed as the default values. The leftmost column of the table contains check boxes. These are used to select transitions of interest, in preparation for downloading them as a single file via the "Export Selection" button at the foot of the table.

Benchmarking

A total of 67 unique proteins were identified from five lists of validated transitions downloaded from SRMATlas (Tables 1 and 2). From the 67 proteins, 99 out of 109 (91%) of peptides were found to be present in the MRMAid database, which relate to 171 transitions in the list. Of the 10 peptides that were not found, three are not tryptic peptides, three were not unique tryptic peptides within the human proteome used in this article, three had no related spectra in PRIDE, and the remaining peptide was discarded during the filtering process due to inconsistent precursor mass information.

The MRMAid transitions for the 67 proteins are summarized in the large table in the supplementary materials (Supplementary Material 1; see online supplementary material at <http://www.liebertonline.com>). The reported transitions were benchmarked at the peptide and product ion levels. To a potential MRMAid user, the most important observation from the table is that the majority of the validated SRM peptides rank in the top five peptides recommended for each protein. As explained earlier, the peptide probabilities seem to underestimate the chance of detecting each peptide, but it is still a useful metric for ranking. The table also shows that a peptide score of 20 can be empirically considered to be a useful threshold for filtering transitions, because over 90% of the validated transition peptides exceed this score.

TABLE 2. SUMMARY OF BENCHMARKING RESULTS

	Peptide									Product ion				
	Probability			Peptide score				Rank		Probability			Rank	
	>0.5	>0.9	0.95	>20	>30	>40	>50	Top 5	Top 3	>0.5	>0.9	>0.95	Top 5	Top 3
All instruments	10	2	0	96	74	32	3	64	48	150	90	67	156	138
ESI Trap	10	2	1	95	74	34	3	61	47	139	96	78	148	125

The values indicate the number of peptides and product ions from the validated lists in Table 1 that appear in MRMAid results when filtered according to the criteria shown. The total numbers of validated peptides and product ions are 99 and 171, respectively. For each criteria category, the criteria become stricter from left to right.

ESI, electrospray ionization.

At the product ion level, there is excellent agreement between MRMAid's results and the validated transitions. The top five ranked product ions covered around 86% of those used in the validated transitions. The probabilities of observing the product ions also appear to be more realistic at this level, as these can be computed directly from the PRIDE data.

It is interesting to note that there is little difference between the results for the designated instrument type (ESI Trap) and the results obtained when spectra from all instruments are considered. This suggests that MRMAid is robust enough to predict suitable transitions, even if there are no data in PRIDE for the user's instrument.

Discussion

To our knowledge, MRMAid 2.0 is one of the first projects to perform large-scale mining of the PRIDE database, and certainly the first SRM transition tool to use PRIDE data. Mining PRIDE has revealed some interesting issues such as the wide variety of instrument identifiers used and the presence of ambiguous metadata, but by applying simple filters we have shown that it is possible to extract valuable methodological information from PRIDE on a large scale. Although we have only concentrated on humans to date, we believe there is already enough data in PRIDE to create viable MRMAid builds for at least 10 other species, and adding these is a priority for future work.

The drawback of building transitions from data in a public repository such as PRIDE is that the data come from a wide range of experiments with diverse aims, different sample preparation methods, and a range of sample matrices. Since PRIDE captures experimental metadata it should be possible to use this information to design transitions for specific applications, but we have found that the quality of the metadata and breadth of coverage is not yet sufficient to achieve this in practice. The only way to develop application-specific transitions at the moment is to gather your own shotgun proteomics dataset and use software such as Skyline (MacLean et al., 2010) to suggest a list of transitions from that dataset.

However, the benchmarking against validated transitions has shown the overall performance of MRMAid 2.0 to be excellent, in that it covers the vast majority of the validated peptides and product ions, and most are ranked highly in MRMAid's results. The results for prediction of the best product ions to monitor is particularly striking, as fragmentation patterns have hitherto been regarded as being poorly conserved across different experimental set-ups.

Conclusions

MRMAid 2.0 represents a substantial improvement on the original implementation of MRMAid, and in its current form is a valuable tool for assisting proteomics practitioners in the design of SRM assays. Perhaps more importantly, the reimplementation of MRMAid has put in place a valuable foundation for future development. As the spectral data in PRIDE continue to increase in volume, quality, and coverage, so MRMAid will improve as the transition database is periodically rebuilt. Furthermore, thanks to the pre-computed nature of the transitions, the increase in available data will not significantly decrease the responsiveness of the MRMAid website to users.

Acknowledgments

MRMAid is part of ProteoSuite, funded by Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/I00095X/1. For more information, visit www.proteosuite.org. We would like to thank the PRIDE team at the EBI, and Dr. Luca Bianco for helpful discussions.

Author Disclosure Statement

No competing financial interests exist.

References

- Anderson, L., and Hunter, C.L. (2006). Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Molec Cellular Proteomics* 5, 573–588.
- Ang, C.S., Rothacker, J., Patsiouras, H., Gibbs, P., Burgess, A.W., and Nice, E.C. (2011). Use of multiple reaction monitoring for multiplex analysis of colorectal cancer-associated proteins in human feces. *Electrophoresis* 32, 1926–1938.
- Cham Mead, J.A., Bianco, L., and Bessant, C. (2010). Free computational resources for designing selected reaction monitoring transitions. *Proteomics* 10, 1106–1126.
- Elschenbroich, S., Ignatchenko, V., Clarke, B., et al. (2011). In-depth proteomics of ovarian cancer ascites: Combining shotgun proteomics and selected reaction monitoring mass spectrometry. *J Proteome Res* 10, 2286–2299.
- Farrah, T., Deutsch, E.W., Omenn, G.S., et al. (2011). A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Molec Cellular Proteomics* 10, M110.006353.
- Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). BioMart central portal—unified access to

- biological data. *Nucleic Acids Res* 37(Web Server issue), W23–W27.
- Johnson, P.E., Baumgartner, S., Aldick, T., et al. (2011). Current perspectives and recommendations for the development of mass spectrometry methods for the determination of allergens in foods. *J AOAC International* 94, 1026–1033.
- Kay, R.G., Gregory, B., Grace, P.B., and Pleasance, S. (2007). The application of ultra-performance liquid chromatography/tandem mass spectrometry to the detection and quantitation of apolipoproteins in human serum. *Rapid Commun Mass Spectrometry* 21, 2585–2593.
- Keshishian, H., Addona, T., Burgess, M., Kuhn, E., and Carr, S.A. (2007). Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Molec Cellular Proteomics* 6, 2212–2229.
- Keshishian, H., Addona, T., Burgess, M., et al. (2009). Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Molec Cellular Proteomics* 8, 2339–2349.
- Krokhin, O.V., Craig, R., Spicer, V., et al. (2004). An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: Its application to protein peptide mapping by off-line HPLC-MALDI MS. *Molec Cellular Proteomics* 3, 908–919.
- MacLean, B., Tomazela, D.M., Shulman, N., et al. (2010). Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* (Oxford, England) 26, 966–968.
- McKay, M.J., Sherman, J., Laver, M.T., Baker, M.S., Clarke, S.J., and Molloy, M.P. (2007). The development of multiple reaction monitoring assays for liver-derived plasma proteins. *Proteomics Clin Appl* 1, 1570–1581.
- Mead, J.A., Bianco, L., Ottone, V., et al. (2009). MRMAid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. *Molec Cellular Proteomics* 8, 696–705.
- Shadforth, I., Xu, W., Crowther, D., and Bessant, C. (2006). GAPP: A fully automated software for the confident identification of human peptides from tandem mass spectra. *J Proteome Res* 5, 2849–2852.
- UniProt Consortium. (2011). Ongoing and future developments at the universal protein resource. *Nucleic Acids Res* 39 (Database issue), D214–D219.
- Vizcaino, J.A., Cote, R., Reisinger, F., et al. (2009). A guide to the proteomics identifications database proteomics data repository. *Proteomics* 9, 4276–4283.

Address correspondence to:
Conrad Bessant
Cranfield Bioinformatics Group
Cranfield Health
Cranfield University
Bedfordshire MK43 0AL, U.K.

E-mail: c.bessant@cranfield.ac.uk