Integrated workflow to design methods and analyze data in large-to-extremely-large scale SRM experiments

Amol Prakash[1], Reiko Kiyonami[2], Alan Shoen[2], Huy Nguyen[2], Scott Peterman[1], Andreas Huhmer[2], Mary Lopez[1], Bruno Domon[3]
[1]Thermo Fisher Scientific, Cambridge, MA; [2]Thermo Fisher Scientific, San Jose, CA; [3]Institute for Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

Thermo SCIENTIFIC

Overview

Purpose: Developing SRM assays for identifying sensitive and selective transitions with a high throughput is a challenging, iterative labor-intensive task.

Methods: A novel software algorithm was developed to facilitate the various iterative steps involved, performing all these steps in matrix. It provides a mix of automated as well as interactive options allowing the user to make the optimal decision at every step. It also allows the user to save the various iterations for later examination.

Results: Using the novel software, we built a quantitative and qualitative assay for 1000 proteins from yeast.

Introduction

SRM-based targeted assays have the potential of becoming the choice technology for many diagnostic and therapeutic purposes. SRM based assays provide high sensitivity, selectivity and throughput, altogether providing a breakthrough quantification methodology. However, developing such an assay can be challenging, as it requires multiple refinement steps, each of which is quite complex and involved. Moreover, in almost all cases, the selection of peptides, transitions, and optimization is performed using synthetic peptides in a matrix which is typically much simpler than the actual matrix in which we aim to quantify (e.g., plasma or serum) the targeted peptides. The choice of transitions/peptides and the optimized retention time and instrument parameters may completely change depending on the interferences caused by the complex background matrix. Moreover, improvement in sensitivity leads to loss of selectivity and throughput, and vice versa, leading to unwanted trade-offs.

In this report, we present novel software that can streamline all the steps involved providing an iterative mode of development in the actual matrix. For a target protein, the workflow usually starts with a choice of peptides and transitions that are unique and "proteotypic" for that protein. Next, empirical and heuristic based approaches are applied to define the initial set of transitions. Further steps require data acquisition using the initially defined set of transitions with subsequent assessment of the data quality to eliminate less than optimal transitions and therefore confirm the next set. When analyzing target peptides in a background complex matrix such as plasma or serum, data analysis can be very complex necessitating novel algorithms and validation strategies. As a result, the new, optimized, set of transitions may contain transitions and peptides from the previous set in addition to others. In this approach, the various iterations are maintained in the algorithm memory allowing fully automated development of optimization steps. A further challenge is presented with multiplexing, where a multiple set of proteins/peptides are targeted for a single assay. In order to provide efficiency for these processes, the algorithm provides intelligent resource management. Finally, once the peptides and transitions are selected, they must be optimized for instrument parameters to provide the best signal-to-noise in the assay. This task is fully automated in the final matrix. Details on assay building, data analysis and calibration and in-depth details are presented in the figures below.

Methods

Sample preparation

Serum samples were thawed and prepared by diluting serum 1:4 v/v with 8 M GuHCl/150 mM Tris/10 mM DTT pH 8.5. Samples were incubated at 37 °C for 60 minutes and then cooled to room temperature. Alkylation was performed by addition of 500 mM iodoacetic acid/1 M Tris pH 8.5 to each sample to give a final concentration of 40 mM and incubated in the dark at room temperature for 60 minutes. The reaction was quenched with the addition of 2M DTT to a final concentration of 5mM. Post quench, 30µg samples were diluted to 120 µL with the addition of 50 mM Tris/5 mM $CaCl_2$ pH 8.0 and 0.3µg of sequencing grade enzyme (Promega) was added to each sample and incubated at 37 °C for 24 hours. The digestion reaction was quenched with the addition of TFA to 1% and heavy (R or K) peptide standards were spiked into each sample. Samples were then desalted by SPE on 96 well plates (ThermoFisher).

High resolution LC-MS/MS

High resolution LC-MS/MS analysis was carried out on an LTQ-Orbitrap XL mass spectrometer (ThermoFisher Scientific). Samples in 5% (v/v) acetonitrile 0.1% (v/v) formic acid were injected onto a 75µm x 25cm fused silica capillary column packed with Hypersil Gold C18AQ 5µm media (ThermoFisher Scientific), in a 250µL/min gradient of 5%(v/v) acetonitrile, 0.1% (v/v) formic acid to 30% (v/v) acetonitrile, 0.1% (v/v) formic acid over the course of 180 minutes with a total run length of 240 minutes. The LTQ-Orbitrap was run in a top 5 configuration at 60K resolution for a full scan, with monoisotopic precursor selection enabled, and +1, and unassigned charge state rejected. The analysis was carried out with CID fragmentation modes.

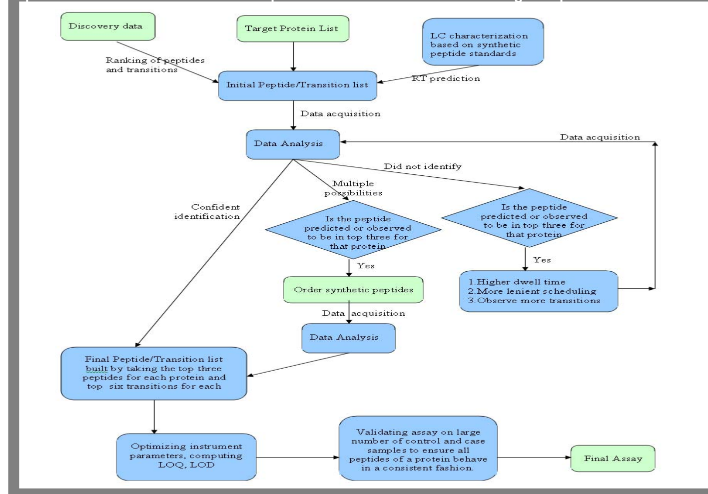
Label-free, differential analysis and protein identification

Bioinformatic analysis was carried out using the SIEVE (ThermoFisher Scientific) algorithm and SEQUEST.

SRM assays

SRM assays were developed on a Thermo Scientific TSQ Vantage and TSQ Quantum Ultra triple quadrupole mass spectrometers (Fig. 2), Surveyor MS pump, Micro autosampler and an IonMax Source equipped with a low flow metal needle. Samples were analyzed in triplicate. Reversed phase separations were carried out on a 1 mm X 50 mm Thermo Scientific Hypersil Gold 1.9µm C18 particle. Solvent A was LC-MS grade water with 0.2% (v/v) formic acid, and solvent B was LC-MS grade 30% (v/v) acetonitrile with 0.2% (v/v) formic acid (Optima grade reagents, Thermo Fisher Scientific).

FIGURE 1. Flowchart detailing how our we choose the best peptides and transitions for a set of proteins. Blue colored boxes are steps that the software automates along with provision of manual



Step 1 : iSRM using discovery data

Figure 1 shows the optimized flowchart using iSRM (intelligent SRM) for selecting the best peptides and transitions for each of the targeted proteins. The blue boxes are steps that are automated by the software (user option of manual analysis) including correction and feedback. The green boxes show user input/output. The workflow starts with a list of targeted proteins, for example proteins involved in a particular pathway, or any list of proteins found to be differentially expressed in a biomarker discovery study.

The first step is to utilize all the available discovery data for the proteins of interest to help initiate the analysis. Usually, most proteins that are being targeted have been seen previously in some discovery experiments. This information can be extremely useful in helping to estimate the best peptides and transitions, as well as for increasing throughput. While the iSRM software can accept the discovery data in multiple formats, for the present study we utilized data acquired from human serum samples digested with trypsin and analyzed on an Orbitrap instrument with CID and HCD as the fragmentation modes. All the MS/MS spectra were analyzed with SEQUEST and the data were quantified with SIEVE differential expression software (ThermoFisher). SIEVE annotates every protein/peptide observed with an MS1 peak integrated intensity and observed LC retention time. Figure 2 shows the SIEVE analysis (identification and integration) for one of these protein/peptides.

Step 2 : LC characterization to enable T-SRM

In order to extrapolate elution behavior of peptides from the discovery LC workflow to the targeted LC workflow, 10 heavy peptides were spiked in samples and run on both platforms. The observed retention times from both systems were used to generate a retention time plot comparing the linearity of the discovery LC versus the targeted LC workflows. These peptides were also used to further plot the relationship between hydrophobicity factors and observed retention time on the targeted platform. This allowed peptides that had not been seen on the discovery platform to be scheduled. Figure 3 plots these two relationships. The high R-square value demonstrates the strong correlation between the two workflows confirming the feasibility of predictive scheduling.

Step 3a : Initial Peptide/transition list

Using the discovery data, peptides for each protein can be ranked based upon their observed intensities. We use this criteria to choose peptides. The top five peptides and the top eight y-ions observed in the discovery data are chosen as the initial list of transitions. Typically, the top two product ions are chosen as the primary SRM transitions and the next 6 as the secondary SRM transitions.

Step 3b : Scheduling

A predictive retention time window based on LC characterization is scheduled for each transition. This provides an estimate of the dwell time per transition, based on the number of transitions being scheduled at any given time (Figure 4A). If the dwell time is too low, the software algorithm splits the acquisition into multiple runs (Figure 4A), and the acquired data are then smoothly stitched back together eliminating the need for manual file intervention.

FIGURE 2. High-resolution LC-MS/MS Orbitrap were processed with SIEVE example peptide analysis below. As part of the quantitative analysis, proteins are identified with SEQUEST. In addition, each peptide is annotated with its MS1 peak integration intensity observed retention times.

FIGURE 3. LC retention time plot for 10 heavy peptides comparing the (a) discovery (Orbitrap XL hybrid trap) and targeted (Vantage triple quadrupole) platforms. (b) hydrophobicity factors and targeted (Vantage triple quadrupole) platform Peptides were chosen based upon their predicted elution times across the gradient. The observed linear correlation between the two LC workflows was excellent.

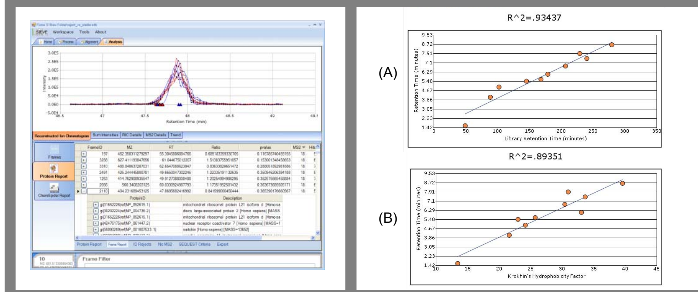
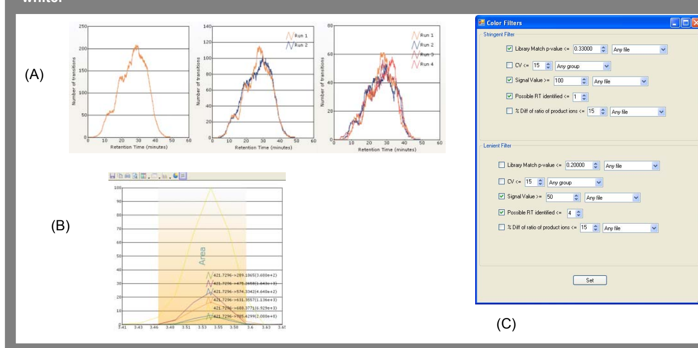


FIGURE 4. (A) Plot of the number of transitions versus time across the SRM LC gradient. This information allows for the estimation of dwell time and data quality before data acquisition and the run can be split into multiple runs if required. (B) The software identifies the retention time of the target peptides by identifying the co-elution times of multiple fragment ion transitions. The figure shows the co-eluting transitions for the peptide GPFFVWAAK (precursor m/z = 421.7296). (C) Various scoring filters and threshold to help in automated data analysis, which is then translated into color. All peptides passing the stringent filter are green, all peptides passing the lenient are yellow, and the remaining are white.



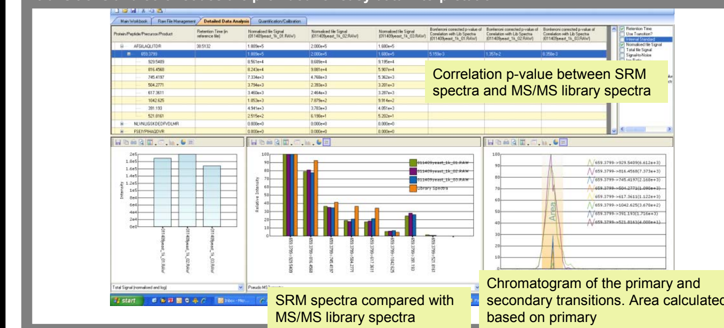
Step 4a: Data analysis and scoring

Once the data are acquired, the software automatically identifies peptides based on co-eluting primary transitions (Figure 4B). If multiple SRM channels produce signal at multiple retention times, the software identifies all of them, and ranks them based on their correlation p-value with the library spectra (Figure 5).

The software also computes peak integrated areas of all primary ions and coefficient-of-variance percentages.

Based on user-defined filters and thresholds (Figure 4C), the software scores all peptides giving a measure of confidence in their identification. These scores are represented as color codes: green (confident identification), yellow (multiple identifications) and white (not identified). The analysis for 1000 yeast proteins is shown in Figure 5.

FIGURE 5. Example analysis of 1000 proteins in yeast lysate using iSRM. The software chooses the optimal retention time for each of the peptides based upon the correlation between the primary/secondary SRM transitions and the library spectra. Quantification is based upon the primary transitions and color codes are provided for easy data interpretation.



Step 5: Peptide/Transition list refinement

Color coding of peptides and transitions (Figure 5) provides for rapid interpretation. Green-coded peptides are those that were confidently identified exactly at an LC retention time in the predictive scheduling window, or have an identified retention time that has a very high correlation score with the discovery MS/MS spectra. These peptides are ready to be finalized for the final assay. Yellow-coded peptides have multiple potential elution times, and require synthetic peptides for further confirmation. White-coded peptides are those that were not identified. Intensity-based ranking of the peptides on the targeted platform allows user choice for the next iteration of either only the high confidence green-coded peptides or the inclusion of yellow and white scored peptides for further exploration by configuring more lenient scheduling and/or choosing different transitions.

Step 6: Building an extremely large-scale, high-throughput assay

In an experiment designed to scale the described iSRM process into a large-scale, high-throughput assay, a yeast cell lysate digest was analyzed on a triple quadrupole mass spectrometer equipped with nanoLC pump and nanospray source (See Poster #T138). In that study, prior data used for empirical predictions was collected on multiple linear-ion-trap and triple-quadrupole instruments. SRM transitions were prioritized into primary (top 2) and secondary (next 6), with primary transitions leading to a data-dependent scan in which all primary and secondary ions were quantified. This strategy in combination with the predictive scheduling strategy described herein allowed the development of an extremely large scale, high-throughput assay of 1000 peptides. The total transition set was split into 4 runs to provide high dwell times (~20ms) for better quantification across peaks. Data acquired in one run were used to further refine the next acquisition.

Conclusions

- Novel software facilitates iterative building of SRM assays, automating the choice of the optimal peptides and transitions therefore making instrumentation and analysis time much more efficient.
- The strategy described herein makes possible the development of extremely large scale, high-throughput assays of up to 1000 peptides and transitions.

References

(1) V. Krokhin, R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, J. A. Wilkins
An improved model for prediction of retention times of tryptic peptides in ion-pair reverse-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. Molecular and Cellular Proteomics 2004 3:908-19

ICAT and SEQUEST are registered trademarks of the University of Washington.

Intel and Pentium are registered trademarks of Intel Corporation. Microsoft and Windows are registered trademarks of Microsoft Corporation. All other trademarks are the property of Thermo Fisher Scientific Inc. and its subsidiaries.