DOI 10.1007/s12014-009-9023-6

# Critical Evaluation of Product Ion Selection and Spectral Correlation Analysis for Biomarker Screening Using Targeted Peptide Multiple Reaction Monitoring

Jian Liu · Johannes A. Hewel · Vincent Fong ·
Michelle Chan-Shen-Yue · Andrew Emili

Published online: 14 February 2009

**Abstract**

*Introduction* Tandem mass spectrometry (MS/MS) has emerged as a cornerstone of proteomic screens aimed at discovering putative protein biomarkers of disease with potential clinical applications. Systematic validation of lead candidates in large numbers of samples from patient cohorts remains an important challenge. One particularly promising high throughout technique is multiple reaction monitoring (MRM), a targeted form of MS/MS by which precise peptide precursor–product ion combinations, or transitions, are selectively tracked as informative probes. Despite recent progress, however, many important computational and statistical issues remain unresolved. These include the selection of an optimal set of transitions so as to achieve sufficiently high specificity and sensitivity when profiling complex biological specimens, and the corresponding generation of a suitable scoring function to reliably confirm tentative molecular identities based on noisy spectra.

*Methods* In this study, we investigate various empirical criteria that are helpful to consider when developing and interpreting MRM-style assays based on the similarity between experimental and annotated reference spectra. We also rigorously evaluate and compare the performance of conventional spectral similarity measures, based on only a few pre-selected representative transitions, with a generic scoring metric, termed $T_{\mathrm{corr}}$, wherein a selected product ion profile is used to score spectral comparisons.

*Conclusions* Our analyses demonstrate that $T_{\mathrm{corr}}$ is potentially more suitable and effective for detecting biomarkers in complex biological mixtures than more traditional spectral library searches.

**Keywords** Clinical proteomics · Biomarker · Validation · Tandem mass spectrometry · Peptide identification · Spectral correlation · Multiple reaction monitoring · Bioinformatics

## Introduction

Clinical proteomics depends on the identification of putative biomarkers that correlate with disease status and the subsequent confirmation of differential abundance of candidates in large numbers of patient samples. Currently, tandem mass spectrometry coupled with high performance liquid chromatography (LC-MS/MS) has become the prevailing technique for both high throughput shotgun identification and subsequent verification of potential biomarkers (1–3). Associated software tools and statistical measures have likewise become essential for interpreting the resulting large spectral datasets so as to accurately determine the identities of the corresponding cognate proteins (4).

In general, two types of automated computational procedures have been developed to reconstruct peptide sequences from MS/MS spectra. The first class of techniques attempts to solve this problem by interpreting the

Jian Liu and Johannes A Hewel contributed equally to this study.

**Electronic supplementary material** The online version of this article (doi:10.1007/s12014-009-9023-6) contains supplementary material, which is available to authorized users.

J. Liu · J. A. Hewel · V. Fong · M. Chan-Shen-Yue · A. Emili
Banting and Best Department of Medical Research,
Terrance Donnelly Center for Cellular and Biomolecular
Research (CCBR), University of Toronto,
Toronto, ON, Canada

A. Emili (✉)
CCBR—160 College Street,
Toronto, ON M5S 3E1, Canada
e-mail: andrew.emili@utoronto.ca

entire set of product ions recorded in the experimental spectra. Such tools can be further divided into two major subcategories. De novo methods, such as the PEAKS algorithm (5), attempt to deduce either part (i.e., sequence tags) or the complete peptide sequence based solely on the observed spectral features based on finding the informative b and y ion fragment series characteristically embedded in experimental peak lists. Conversely, database search tools utilize scoring algorithms to compute the similarity (closeness of fit) between each experiment spectra and theoretical representations of peptide product ion patterns generated in silico as a pragmatic solution for large-scale proteomic profiling. For instance, the popular SEQUEST program (6) uses a correlation function to compute the spectral similarity of predicted patterns of b and y ions to the experimentally observed peaks; its variants have been recently developed (7, 8) that improve search speed by novel algorithms. Determining the confidence associated with a tentative match is paramount. For example, the Mascot database search program (9) assigns a score based on the probability of obtaining a close match by chance alone. Hybrid approaches (10, 11) have recently been introduced combining the merits of de novo and database search techniques. However, information about specific ion intensities is not exploited in the scoring functions of most database-search algorithms. While software tools, such as MassAnalyzer (12), can predict ion intensity patterns in silico, this has not been widely implemented in common database-search routines. Uncertainty in such intensity predictions may bias a spectral correlation or render the score meaningless.

The second school of peptide identification tools is based on the direct comparison of experimental spectra. The rationale is that spectra are representative, and hence should generally be highly similar for the same peptide even if obtained independently, assuming the data are produced by analogous instruments under similar conditions. In this paradigm, a compendium of annotated reference spectra is first established, based either on synthetic peptides or recombinant proteins or from available high confidence identifications. Then, unassigned spectra are compared to the library and subsequently identified based on a highly similar match. Pairwise similarity between two experimental spectra can be computed as the normalized inner product (13) or variants such as the correlation coefficient (14). Popular tools in this class include Pep-Miner (15), MS2Grouper (16), X! Hunter (17), and BiblioSpec (18). Such methods have the advantage of offering fast, and potentially more reliable, identifications. The matching process has even been extended to document or verify instances of posttranslational modifications (19, 20).

Unlike clinical discovery studies, wherein large numbers of proteins must be identified, usually in a few biological specimens, biomarker validation is typically restricted to evaluating a subset of promising leads across many samples. Multiple reaction monitoring (MRM) is particularly well suited to this focused form of proteomic detection, since it offers the potential for high throughput quantization in a clinical context (21). In a typical MRM screen using a triple quadrupole instrument, the first mass analyzer is used to preselect one or more target peptides of interest based on predefined precursor mass-to-charge (*m/z*) ratios. These molecules are then fragmented in a second analyzer, resulting in multiple product ions. In a typical MRM experiment, only a few b/y product ions are usually transmitted by the third analyzer to the detector. These reporter ions are often selected with the goal of maintaining a high duty cycle required for accurate quantization in mind. Consequently, protein identifications are based solely on detecting these signature transitions. The relative intensity profiles of these selected fragment ions could likewise potentially be incorporated into a descriptor. In this respect, we (22, 23) and others (24) have also introduced an LC-MS/MS-based screening procedure, termed targeted peptide monitoring (TPM) (22, 23) or product ion monitoring (PIM) (24), that combines selective precursor data acquisition on an ion trap instrument with spectral library comparisons to achieve highly sensitive and specific protein identifications.

Computational tools have been developed to facilitate such MRM/TPM pipelines. For example, the software suite TIQAM (25) was introduced to assist in the selection of suitable signature ions to assay target peptides of interest. A public spectrum library (http://thegpm.org) also allows for searching of signature ions in existing protein spectra based on intensity and uniqueness measures. However, defining which transitions (precursor–product ion combinations) are in fact optimal for screening clinical samples depends on diverse empirical criteria, which are often not rigorously evaluated in a relevant biological context. Here, we study various practical and statistical issues relating to this problem. As a recent study (26) has established the advantages of database-dependent peptide identifications based only on correlating b/y ions, we have explored this same direction from a quantitative MRM/TPM assay perspective, proposing an effective, generalizable “$T_{\mathrm{corr}}$” (for *T*ransition *corr*elation) scoring metric. The potential practical advantages and general feasibility of applying such a scoring mechanism for routine biomarker validation is addressed.

## Methods

### Tandem Mass Spectrometry of Synthetic Reference Peptides

We acquired experimental reference spectra for 384 distinct synthetic peptides (JPT technologies, Berlin, Germany) using the TPM concept of continuous targeted precursor isolation and collision-induced dissociation fragmentation using predefined inclusion lists in three complementary manners:

First, we performed continuous direct infusion of individual peptides in a linear ion trap LTQ instrument to capture a comprehensive collection of spectra at three different charge states (+1, +2, and +3) over a 3 *m/z* isolation window. Two microliters of each crude peptide (0.2–2.0 pmol) was individually injected using an EASY-nLC autosampler/HPLC system (Proxeon, Odense, Denmark) with a nano-electrospray source at an isocratic flow rate of 600 nL/min using an unpacked 100 um I.D. microcapillary fused silica column (Polymicro, Pheonix, AZ, USA) with an ~10 m tip opening generated using P-2000 laser puller (Sutter, Carlsbad, CA, USA). The LTQ was programmed to monitor precursors generated by the peptides in each of three charge states (+1, +2, and +3) to a maximal range of 2,000 *m/z*. Total runtime for each peptide was 8.2 min, with 5 min of spectral recording in centroid mode. For each peptide, about 800~1,000 spectra were produced. The resulting MS/MS spectra of 384 peptides, acquired in ~52 h of data acquisition, were used to construct a reference spectra library.

Second, an analogous TPM dataset was acquired using targeted LC-based MS/MS, however, monitoring only the +2 charge precursor m/z. A microcapillary column was packed with Luna 3 µm reverse-phase media (Phenomenex, Torrence, CA, USA) over a length of 10 cm. Sample loading and organic gradient chromatography [98% buffer A (95% water, 5% acetonitrile and 0.1% formic acid) to 80% buffer B (95% acetonitrile, 0.1% formic acid, in water) over 45 min] at a flow rate of 300 nL/min was driven by the Proxeon EASY-nLC system. Spray voltage was at +2.5 kV. Twelve precursors of 12 peptides were monitored in parallel over a 1- to 2-s duty cycle, with an average chromatographic peak width of 6- to 15-s FWHM. MS/MS spectra, 6227, were acquired for 12 different peptides mixed together and injected for one targeted LC/MS/MS experiment.

Third, we performed TPM-style LC-MS/MS in the same way as described above after spiking the sets of 12 peptides (approximately 800 fmol) into a tryptic digest of a mouse embryonic stem cell cytosolic extract (5 ug total protein) serving as a complex biological background.

The MS/MS spectra were all mapped to a sequence database composed of the 384 peptide targets by SEQUEST (v2.7).

## Precursor–Product Ion (Transition) Correlation Analysis

Given a list of presumably prominent product ions $T=<M_1, M_2, ,M_n>$, a spectrum can produce an intensity vector $f=<I_1, I_2, ,I_n>$, where $M_{ii}$ are the *m/z* and intensity of *i*th selected daughter ions. To determine the intensity, the peak list is partitioned into a sequence of equal-sized bins (1 Da by default). Each peak is assigned to the bin covering its *m/z.* If a bin contains multiple peaks, the intensities are summed up. For a given product ion, the intensity is the total intensity in the associated bin. The pairwise correlation between two spectra *s*1 and *s*2 under the given product ion series is formally defined as:

$$T_{\text{corr}}(s1, s2) = \frac{\langle f_1 \cdot f_2 \rangle}{\|f_1\| \|f_2\|} \qquad (1)$$

where *f*1 and *f*2 are the intensity vectors for spectra *s*1 and *s*2 under predefined transition list, respectively.

Other similar statistics for pairwise correlation have also been proposed to calculate the spectral pairwise similarity. For instance, correlation coefficient is also an effective alternative. However, although correlation coefficient has the merits to return value of 1 when two vectors have a perfect linear relationship, even with certain base shifts, our previous study (14) suggests that, in practice, it is only marginally better than the dot product. In this current work, we chose the inner product for implementation of the spectral correlation. The reason is mainly twofold. First, it is much more computationally efficient. Second, this statistic and its variants have already been widely adopted in many widely accepted software engines, such as SEQUEST (6), X!Tandem (27), and spectral comparison tools (13, 15, 16). As the peak intensities for different ions may vary over a few orders of magnitude, square rooting transform was applied to stabilize the intensity variance (14) as a preprocessing step. Mathematically, the $T_{\text{corr}}$ can be deemed as a special version of dot product-based pairwise spectral similarity. The main difference is that the latter is based upon ions across the entire *m/z* range, while $T_{\text{corr}}$ is limited to preselected product ions. For the sake of brevity and to avoid confusion, the conventional dot product similarity of whole spectra is referred to as $S_{\text{corr}}$ hereinafter.

Moreover, the well established $X_{\text{corr}}$ scores of SEQUEST (6) is also a special version of inner product of two spectra. As SEQUEST is a database search tool, the $X_{\text{corr}}$ is based on the matching between the experimental spectrum and a theoretical one predicted for each candidate peptide. The $X_{\text{corr}}$ formula is described as:

$$X_{\text{corr}}\left(s_{\text{exp}}, s_{\text{theo}}\right) = \left\langle s_{\text{exp}}, s_{\text{theo}} \right\rangle - \frac{\sum_{\tau=-75}^{75} \left\langle s_{\text{exp}}, s_{\text{theo}}(\tau) \right\rangle}{151} \qquad (2)$$

Where, $\tau$ is a displacement value, i.e., the spectrum is displaced by adding $\tau$ Da to the *m/z* value of each peak. Empirically, the displacement values ranges from −75 to 75, which leads to best results for an experiment. To speed up the computation, it is implemented via fast Fourier transform analysis. In this study, we mainly compare results of these three distinct correlation measures.

### Monte-Carlo Simulations to Generalize the Performance of $T_{\mathrm{corr}}$

In a spectrum library search, typically a pool of candidate peptides falls within the predefined mass error tolerance to a given precursor. The discriminating capability of a correlation score therefore reflects the distributions of true- and false-positive matches. To obtain such data, we used the collection of spectra acquired for the synthetic peptides that were correctly identified by the search engine SEQUEST. For each annotated peptide, multiple spectra were available. Therefore, pairwise comparison of all these representative spectra offers a population of correlation scores for true positives, denoted as $C_{\mathrm{s}}$. Likewise, spectra corresponding to the different peptides with approximately the same precursor mass were used to derive background correlation scores, denoted as $C_{\mathrm{d}}$.

In general, and as expected, scores (using either $T_{\mathrm{corr}}$ or $S_{\mathrm{corr}}$) in $C_{\mathrm{s}}$ were higher than those from $C_{\mathrm{d}}$. To generally evaluate the discriminating capacity, especially when the size of candidate peptide pool varies, receiver operating characteristic (ROC) curves were inferred from the underlying $C_{\mathrm{s}}$ and $C_{\mathrm{d}}$ distributions. Assuming that the size of candidate peptide pool for library search with precursor mass constraint is $N$, for a given cutoff correlation score $S_{\mathrm{c}}$, sensitivity was calculated by:

$$P(s > s_{\mathrm{c}}, s > \max(s'i)), i = 1, 2, \cdots, n - 1. \quad (3)$$

where $s \in C_{\mathrm{s}}$, $s'i \in C_d$, respectively. Similarly, the specificity was computed as:

$$P(s_{\mathrm{c}} > \max(s'i)),\ i = 1, 2, \cdots, n \quad (4)$$

where $s'i \in C_{\mathrm{d}}$, with random selection.

As there is not straightforward way to model $C_{\mathrm{s}}$ and $C_{\mathrm{d}}$ in closed form, the theoretical computations (Eq. 3 and 4) are difficult. Hence, a Monte Carlo simulation (28), used in our previous study (14), was adopted to estimate the values. Briefly, as Monte Carlo methods are based on iterative random sampling, they are mostly suitable when deterministic results are computationally infeasible or prohibited. In our case, a group of scores are randomly drawn from the two populations $C_{\mathrm{s}}$ and $C_{\mathrm{d}}$. Such procedure repeats for sufficient times (in our cases 5,000 iterations). Then, the above probabilities are approximated by the observed frequencies satisfying the above conditions (Eq. 3 and 4) and considered to be the estimates of sensitivity and specificity.

We also considered the impact of noise in the spectra. Given the null hypothesis that a $T_{\mathrm{corr}}$ score is achieved by spurious matching to noisy peaks (i.e., a false positive), it is necessary to estimate the significance of the measurement. For this purpose, we generated stochastic intensities for the transition ions, either based on experimental spectra or artificially, to produce a control $T_{\mathrm{corr}}$ population, $C_{\mathrm{R}}$, mapping to noise. Based on this distribution, corresponding $P$ values were estimated for a given $T_{\mathrm{corr}}$ value $s$, proportional to the $C_{\mathrm{R}}$ above a given threshold.

### Impact of Transition Ion Set Sizes on $T_{\mathrm{corr}}$ Performance

It is evident from Eq. 1 that when more transitions are matched (i.e., more ions detected for monitored transitions), the correlation scores tend to be higher, given that all other conditions are the same. As is well known, peptide fragmentation is a scholastic process. Although some approaches, such as PepHMM (30), MassAnalyzer (12), apply sophisticated models to unveil the intrinsic relations between different ions, most popular database search tools, such as SEQUEST (6) and MASCOT (9), adopt simplified models, which assumes that matching of ions are independent events. In this study, we also assume that the observation of specific transition ions follows a binominal distribution, with probability $p$ to detect any individual product ion. Hence, the probability to detect exactly $k$ ions out of total $n$ product ions is:

$$p(k, n, p) = C_n^k p^k (1 - p)^{n-k} \quad (5)$$

Then the cumulative probability of observing more than $k$ ions is:

$$P' = 1 - \sum_{j=0}^{k} p(k, n, p) \quad (6)$$

A detailed table of such cumulative distributions can be found in (29). Although the correct peptides have a higher expectation value of $p$, other false-positive peptides in the database or library also have a chance to match such product ions. This work also provides an empirical study on how the sizes of selected transition sets can impact the $T_{\mathrm{corr}}$ performance (More details can be found in the following "Experimental Results" section.).

## Experimental Results

### Constructing a Reference Spectrum Library

A critical question for biomarker detection is establishing a statistical significant cutoff value for distinguishing biologically correct transitions from potentially spurious false-positive correlations. To produce a rigorous statistical framework, we first constructed a high-quality reference spectral library composed of three independent annotated datasets, each consisting of thousands of representative MS/MS spectra for 384 distinct synthetic peptides derived from a variety of mammalian transcription factors that were analyzed individually by targeted TPM ion trap fragmentation (see “Methods” section).

The first dataset (DIMS) consisted of about 800 to 1,000 MS/MS spectra acquired for each peptide at three precursor charge state (+1/+2/+3) through direct infusion nano-ESI-MS/MS[3]. We used this dataset to construct a spectrum library. For the second dataset, batches of 12 synthetic peptides were mixed together and the +2 charge state precursors subsequently fractionated using microcapillary scale reverse phase chromatography and analyzed by MS/MS over a 45-min chromatographic gradient. We used this dataset to investigate the specificity of fragment-ion intensity patterns across an entire LC-MS/MS experiment. To create the third “test” spectral dataset, we increased the complexity of sample by spiking the 12 peptide sets into a complex mixture consisting of a tryptic digest of a mouse embryonic stem cell cytoplasmic fraction. We used this dataset to evaluate the specificity needed to confidently identify specific target peptides.

The spectra of the first dataset were subsequently mapped against a sequence database composed of the 384 peptide targets using the standard SEQUEST search tool. To generate a reliable core reference set, we picked the top ten highest-scoring positive spectra, sorted according to the SEQUEST $X_{\mathrm{corr}}$ score, for each individual peptide. We then computed the pairwise correlation scores of spectra corresponding to different (nontarget) or the same (target) peptides. To derive discrimination scores to empirically estimate the false discovery rate, all matches to different (nontarget) sequences were deemed false positives, while matches to same (target) sequences were deemed true positives.

## Transition Feature Selection and Correlation Analysis

A key issue for achieving reliable peptide identifications by targeted MS/MS is the appropriate number and types of transition ions chosen for monitoring. We reason that more inclusive lists will generally result in better performance, at least in terms of accuracy, since empirically the chances of observing an equivalent number of product ions with false positives will be substantially reduced with longer lists. Nevertheless, the leading edge of $T_{\mathrm{corr}}$ for true positives diminishes quickly when the number of false positives becomes large (see Eq. 2 and 3 in the “Methods” section). Hence, for our subsequent analyses, we also examined more closely how the number and choice of reporter product ions influence detection sensitivity and specificity, producing some empirical guidelines for transition ion selection.

The first dataset (see the “Methods” section) is used to construct a spectrum library and derive $T_{\mathrm{corr}}$ and $S_{\mathrm{corr}}$ populations. The pairwise comparison was conducted among spectra of the same peptides. As b/y ions are typically most abundant, the full list of b/y ions was used for the $T_{\mathrm{corr}}$ computation. We gauged the relative effectiveness of these two measures by cross-validation and by investigating their performance at peptide detection when applied to complex biological mixtures, with the expectation of detecting the targets as well-defined chromatographic peaks with characteristic retention times.

Overall, we obtained correlation scores for all the tandem mass spectra in the first reference (DIMS) dataset. Histograms of the resulting correlation scores were generated for spectra matching the correct target peptides (Fig. 1a, b, right panels) and those that mapped to a different (nontarget) peptide (left panels). We observed a much higher frequency for correlation values >0.8 for the spectra ($S_{\mathrm{corr}}$) or corresponding transition intensities ($T_{\mathrm{corr}}$) matching the correct sequence than for nontarget matches. The maxima of the distributions were also at ~1.0 and 0.9 for correct cognate sequences using $T_{\mathrm{corr}}$ and $S_{\mathrm{corr}}$, respectively, versus only 0.25 for nontarget matches. Interestingly, the histogram of the false-positive $T_{\mathrm{corr}}$ matches was much broader compared to that for $S_{\mathrm{corr}}$, suggesting less effective discrimination. However, $T_{\mathrm{corr}}$ had a far higher incidence for nearly perfect correlation values for true-positive matches than was seen with $S_{\mathrm{corr}}$.

As $T_{\mathrm{corr}}$ is a special version of $S_{\mathrm{corr}}$ theoretically, we next investigated correlation between $T_{\mathrm{corr}}$ and $S_{\mathrm{corr}}$ metrics in the real application. As seen in a scatter plot analysis of the $S_{\mathrm{corr}}$ and $T_{\mathrm{corr}}$ results based on the spectrum library (Supplementary Figure 1). Supplementary Figure 1a shows the very clear correlation (correlation coefficient 0.879) between the values obtained for true-positive spectra. Conversely, as shown in Supplementary Figure 1b, while the values obtained for false-positive matches were consistently lower, the overall correlation coefficient was only 0.492. It implies that when true positives are identified, $S_{\mathrm{corr}}$ and $T_{\mathrm{corr}}$ have similar values; whereas they can be very different for false positives.

As described in the “Methods” section, for the real application of library search, a one-to-many comparison is conducted (i.e., a pool of candidate peptides with the same precursor mass constraints must be searched). Depending on the search space, the sensitivity and specificity can vary considerably. Therefore, we conducted Monte Carlo simulation to estimate the performance with the search space increases. To make a fair comparison of the discriminating capacity for true and false positives, we derived ROC curves for $T_{\mathrm{corr}}$ and $S_{\mathrm{corr}}$ separately using different pool sizes (10, 100, and 1,000 spectra). Performance area-under-the-curve values are shown in Table 1. $T_{\mathrm{corr}}$ gradually

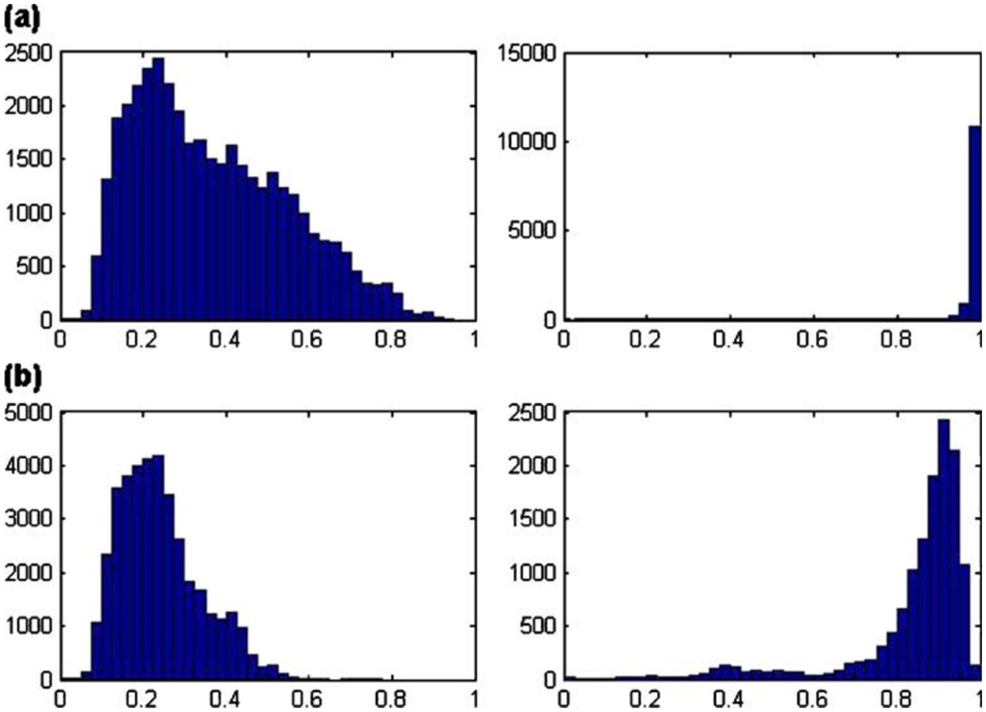

**Fig. 1** Histogram of correlation scores for pairwise spectral comparisons. A subset of high quality spectra from 384 synthetic peptides were compared pairwise. **a** The histogram of $T_{\mathrm{corr}}$ scores generated for comparisons of different (*left panel*) or the same (*right panel*) peptides. **b** Analogous $S_{\mathrm{corr}}$ histograms

outperformed $S_{\mathrm{corr}}$ when examining high-quality spectra as the size of candidate peptide pool increased.

## Statistical Modeling to Estimate $T_{\mathrm{corr}}$ Reliability to Noisy Spectra

We next estimated the statistical significance of $T_{\mathrm{corr}}$ scores in the presence of noisy peaks. In other words, we consider the circumstance that all the peaks from an experimental spectrum for selected ions are spurious. Currently, there are different approaches to model the intensity patterns of noisy fragment ion peaks. For instance, PepHMM (30) relies on a uniform distribution, while PeakSelect (31) assumes a Gaussian distribution. In order to make a solid conclusion, we tested three different schemes. First, we examined our experimental spectra to generate an empirical noise intensity distribution after removing the main (b/y) product ions in the peak lists from consideration. The resulting skewed histogram of observed noise intensities is shown in Fig. 2 (with relative values linearly re-scaled to a [0,1] range). Clearly, the intensity distribution from this set of experimental MS2 spectra does not follow either normal or Gaussian distribution. Therefore, we artificially generated markedly different normalized noise intensities with Gaussian and uniform distributions (data not shown). Then, for each peptide in the library, we repeatedly computed $T_{\mathrm{corr}}$ scores for the corresponding reference spectra to the artificially generated ion intensity series by each of these noise models. Figure 3a shows the resulting distributions of $T_{\mathrm{corr}}$ values. Given such a distribution, $P$ value for a specific $T_{\mathrm{corr}}$ $X$ against noisy spectra is the proportion of $T_{\mathrm{corr}}$ above the threshold $X$. Figure 3b illustrates the $P$ values derived using each of these three models. Despite the overall differences in the curves, the $T_{\mathrm{corr}}$ based on randomized intensities were unlikely to achieve high scores (>0.7), making $T_{\mathrm{corr}}$ relatively immune to noisy spectra, boosting the confidence of peptide identification.

## $T_{\mathrm{corr}}$ Application to LC/MRM/MS

One potential application of $T_{\mathrm{corr}}$ is for peptide identification by MRM using a standard triple quadrupole instrument. In this scenario, only predefined product ions are examined as signature transitions. We have established in a previous study (32) that the overall intensity patterns of MS/MS fragment ions acquired on a linear ion trap (LTQ) are typically highly similar to CID patterns acquired in MRM mode on a triple stage quadrupole. Consequently, the key issue is to choose an optimal transition set, which is unique or most distinct for a given target peptide. Although

**Table 1** Performance comparison of $S_{\mathrm{corr}}$ and $T_{\mathrm{corr}}$ as determined by ROC area under various sizes of candidate peptide pool

| Size of candidate peptide pool | ROC area | |
|---|---|---|
| | $T_{\mathrm{corr}}$ | $S_{\mathrm{corr}}$ |
| 10 | 0.937 | 0.939 |
| 100 | 0.916 | 0.910 |
| 1,000 | 0.907 | 0.871 |

The entire set of putative b and y ions from DIMS library spectra were used for the $T_{\mathrm{corr}}$ calculations

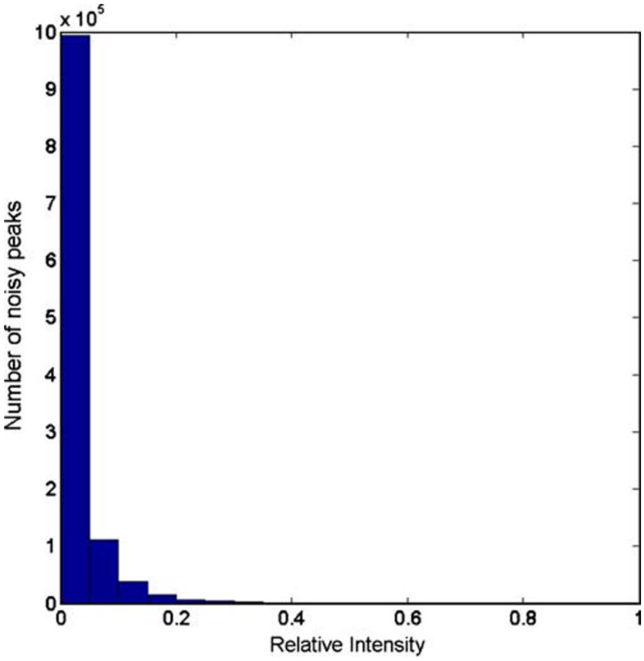

**Fig. 2** Histogram showing the frequency distribution of "noise" peak intensities based on recorded *m/z* ratios not associated with primary product b/y ions. The ions were derived from DIMS high-quality reference spectra

rules to derive robust transitions remain uncertain, software tools, such as TIQAM (21) and P3 (http://proteome.gs.washington.edu/software/P3/), generally select large y ions (above half the mass of the precursor), since these are typically most pronounced. For instance, certain ions may experience multiple fragmentations, producing smaller satellite ions hence, high *m/z* ions are supposed to be more distinct.

In order to validate this strategy, we tested correlation performance using either four y ions, or four b ions, or a combination of the two as target transitions for correlation analysis. For the purpose of comparison, we first chose the smallest ions above the half precursor mass threshold in the respective ion series. In the same manner as we did for the full b/y ion series, we derived the score distributions and generated ROC curves after Monte Carlo simulations. Table 2 provides a summary of the results. Indeed, as expected, ions in the upper *m/z* range appeared to be more discriminating, with y ions outperforming the b ions.

On the other hand, these standard selection criteria entirely ignored the actual intensities of the fragment ions. Low *m/z* but nevertheless high intensity peaks may also have the potential to serve as discriminating ions. Hence, as an alternate comparison, we selected the subsets of four b and/or y ions with the most intensity as targets and conducted similar experiments in the same manner. The results shown in Table 3 confirm that selection of topmost

**Fig. 3** $T_{\mathrm{corr}}$ score distribution for mapping noise to reference spectra and the significance of $T_{\mathrm{corr}}$. Three different intensity modeling for noise were used to compute the randomized the $T_{\mathrm{corr}}$. **a** The empirical frequency of $T_{\mathrm{corr}}$ for each condition. **b** The *P* values of $T_{\mathrm{corr}}$ based on the above randomized $T_{\mathrm{corr}}$ distributions, respectively

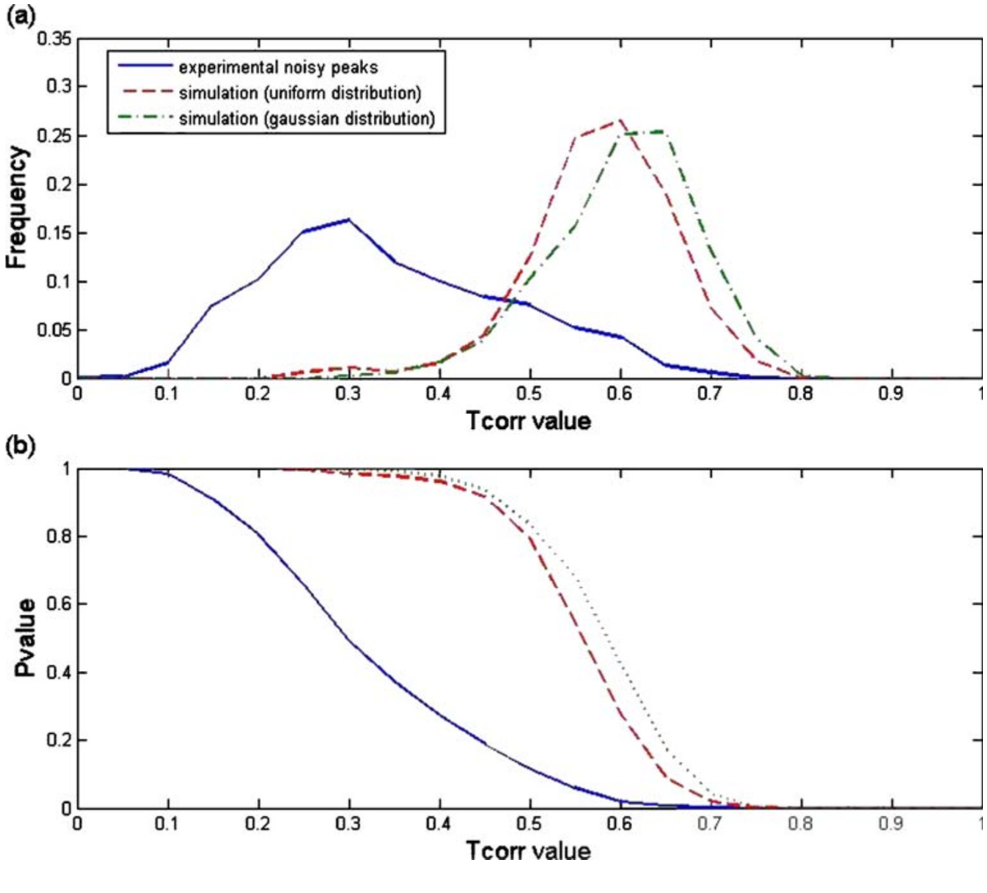

**Table 2** $T_{corr}$ performance using predefined subsets of b and y ions

| | ROC area | | |
|---|---|---|---|
| $T_{corr}$ performed using b and y ions with smallest m/z values *above* half the precursor mass as transition ions | | | |
| Size of candidate peptide pool | 4 y ions | 4 b ions | 4 b and 4 y ions |
| 10 | 0.844 | 0.807 | 0.917 |
| 100 | 0.511 | 0.381 | 0.835 |
| 1,000 | 0.042 | 0.008 | 0.541 |
| $T_{corr}$ performed using b and y ions with smallest m/z values *below* half the precursor mass as transition ions | | | |
| 10 | 0.848 | 0.571 | 0.888 |
| 100 | 0.503 | 0.147 | 0.674 |
| 1,000 | 0.039 | 0.013 | 0.311 |

The ROC area data were based on the first (DIMS) spectrum library of 384 peptides

intensive ions as transitions is capable of outperforming the standard selection criteria (c.f. Tables 2 and 3).

Figure 4 provides a more general comparative view of how increasing the size of transition ion sets gradually improves the performance in terms of ROC area. This sort of analysis can be used to determine the trade-off between the number of transitions monitored and the size of candidate peptides in the library to achieve satisfactory sensitivity and specificity.

## Cross-Validation to Spectra Acquired by LC-MS/MS

Ultimately, since we are interested in developing clinical MRM-type assays, we examined the effectiveness of the $T_{corr}$ metric using a second batch of 6,227 MS/MS reference spectra obtained by LC-MS/MS-based profiling experiments. We randomly mixed batches 12 of the synthetic peptides in equal volumes and analyzed these on a linear ion trap mass spectrometer that was configured to continuously capture MS/MS spectra on doubly charged precursor ions over a chromatographic elution time of 45 min. As a control, the spectra were first categorized into two groups by mapping against a sequence database of the 384 target peptides using the SEQUEST search algorithm. The first group consisted of spectra that mapped to an appropriate batch target, which were presumed to be true positives, while the rest were considered false positives and put into the second group. The two sets of spectra were then correlated using both $T_{corr}$ and $S_{corr}$ against the reference spectra library built from the first (DIMS) dataset.

Figure 5 shows a histogram of correlation scores obtained for these two groups, while Supplementary Figure 2 provides a Venn diagram of the overlap obtained based on from results obtained with $T_{corr}$, $S_{corr}$, and the SEQUEST search engine. Although $S_{corr}$ identified more putative correct targets, the discrimination between the true- and false-positive scores was far more pronounced with $T_{corr}$ than for $S_{corr}$, or $X_{corr}$ (SEQUEST) for that matter.

To analyze the performance of $T_{corr}$ in a more biologically meaningful context, we re-performed the same entire analysis after spiking in 12 synthetic peptides into a complex background consisting of soluble mouse embryonic stem cell cytoplasmic proteins. Figure 6 and Supplementary Figure 3 demonstrate highly similar overall patterns, both in terms of coverage and sensitivity, despite that the slightly diminished discriminating performance of the three correlation metrics.

## Cross-Validation to Spectra Acquired by MALDI Mass Spectrometry

In essence, $T_{corr}$ is based on the assumption that major product ion patterns are most reproducible. Therefore, its performance is comparable to, or often even better, than full spectra comparisons (i.e., $S_{corr}$). As a further independent validation of effectiveness, we also tested a batch of 64 MS/MS spectra generated using a MALDI ToF-ToF instrument, using our LTQ DIMS spectra for the reference ion signatures. $T_{corr}$ and $S_{corr}$ correctly identified 56 and 50 of the targets, respectively. In addition to higher accuracy, Table 4 provides statistical evidence supporting the advantages of $T_{corr}$- over $S_{corr}$-type spectral comparisons. First, calculation of a $T$ test shows that the mean score of the $T_{corr}$

**Table 3** $T_{corr}$ performance using either the four most intensive y ions only, or the four most intense ions observed (regardless of b or y ion status)

| Size of candidate peptide pool | ROC area | |
|---|---|---|
| | 4 y ions | 4 b/y ions |
| 10 | 0.674 | 0.755 |
| 100 | 0.228 | 0.356 |
| 1,000 | 0.009 | 0.144 |

The ROC area data were based on the DIMS spectrum library of 384 peptides

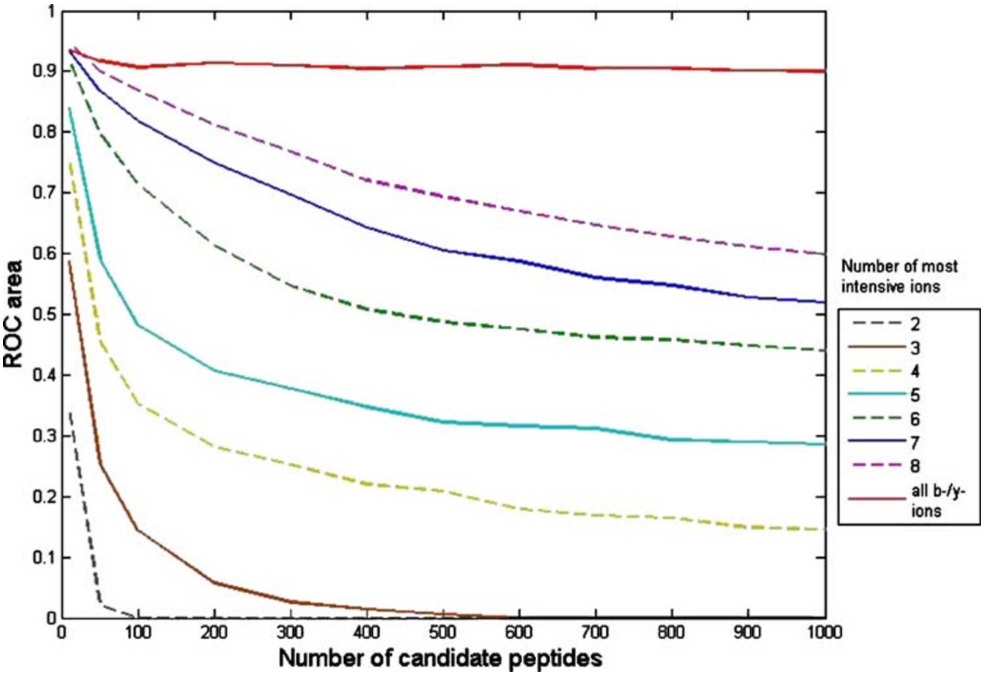

**Fig. 4** $T_{\text{corr}}$ performance using various numbers of highest intensity b and/or y product ions. The ROC area-under-the-curves were estimated by Monte Carlo simulation based on the reference DIMS library spectra

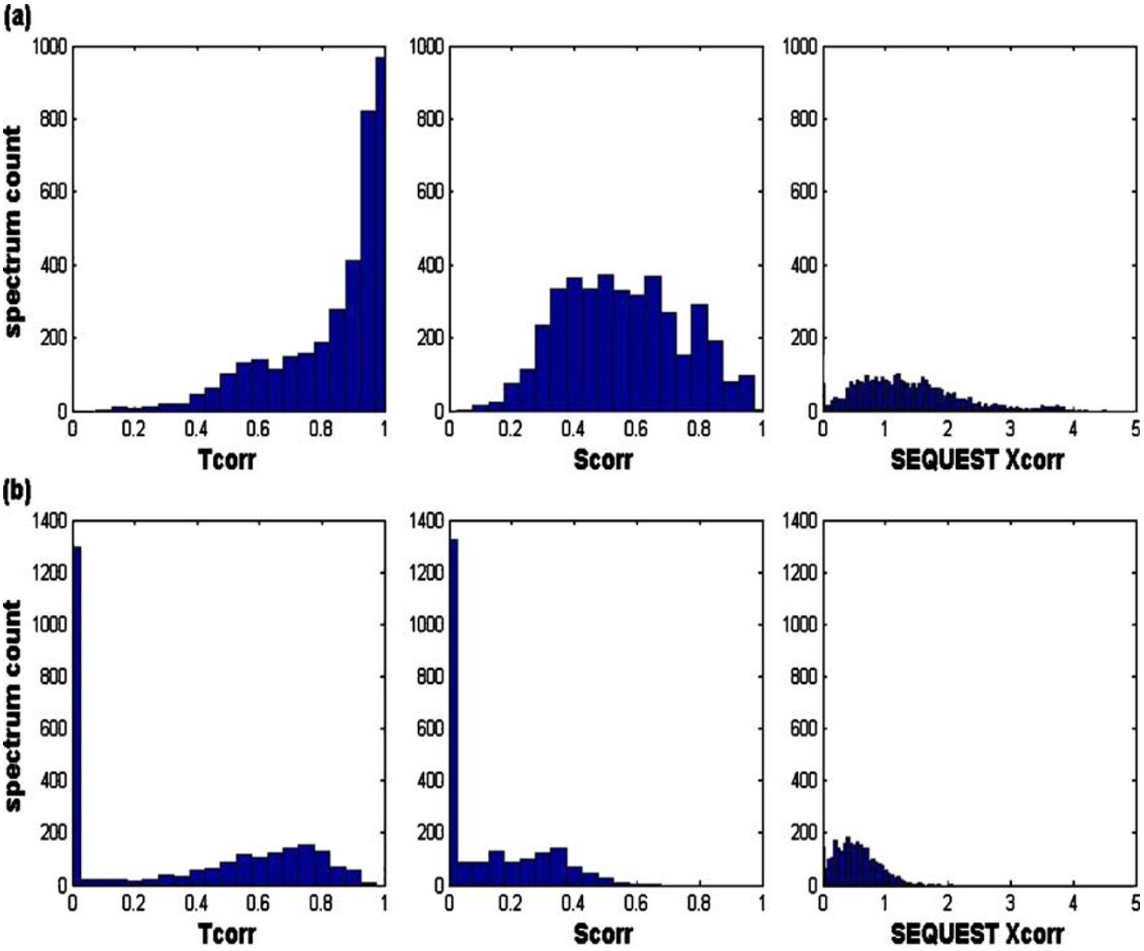

**Fig. 5** Cross-validation. The 6,227 spectra were separated into two groups for each search engine, depending on whether they were assigned to one of the target 12 peptides. **a**, **b** shows the score histograms for true and false positive matches, respectively

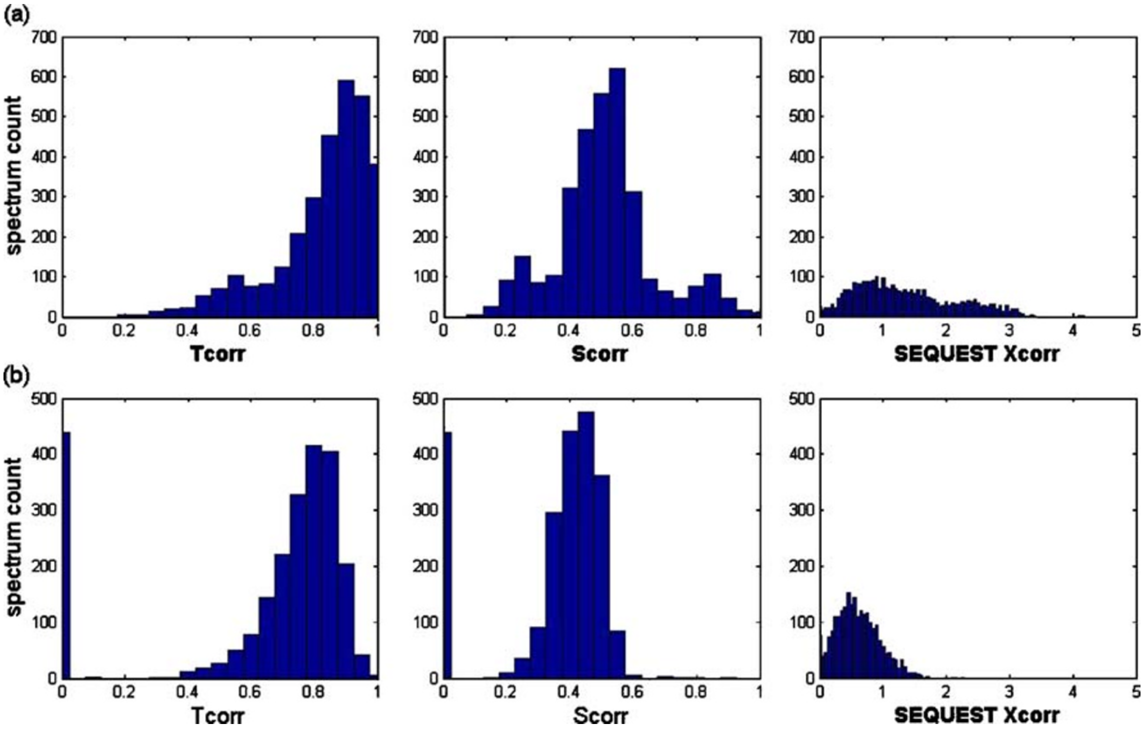

**Fig. 6** Comparative analysis of three different search engines. The analyses involved 5,541 spectra generated by LC-MS/MS of batches of 12 synthetic peptides spiked into a complex biological sample. **a**, **b** Histograms of the score obtained for properly matched and incorrectly identified spectra, respectively

population is significantly higher from that obtained using $S_{\mathrm{corr}}$. Moreover, $T_{\mathrm{corr}}$ was more robust as the mean difference of $T_{\mathrm{corr}}$ values between the correct (i.e., same) and different (mismatched) peptides was larger than that with $S_{\mathrm{corr}}$, implying a more distinct separation between true and false positives.

## Discussions and Conclusion

Clinical proteomics remains critically dependent both on the successful discovery and the subsequent verification of protein biomarker correlation with a disease phenotype under study. Validation, in turn, depends on fast, accurate, sensitive, and robust analytical methods. Currently, few rigorous reports have been published on how to optimize MRM-transition selections, and existing software tools, like TIQAM (21), use heuristics to pick product ions using fixed statistical criteria, regardless of the actual experimental spectra.

To alleviate such a dilemma, we have introduced $T_{\mathrm{corr}}$, a similarity metric for determining spectral similarity to facilitate peptide identification via various modes of MS/MS. Although spectra correlation-based library searching has becoming popular in proteomics studies, most existing tools correlate entire peak lists. In contrast, we explore the feasibility of applying robust correlation methods based on observing much smaller sets of signature ions. This functionality is critical for the general suitability for MRM- or TPM-style targeted mass spectral analyses, in contrast to standard MS/MS spectral comparisons, due to the far more limited ion species coverage obtained by MRM/TPM assays optimized for high throughput protein screening.

**Table 4** Statistics comparison for $T_{\mathrm{corr}}/S_{\mathrm{corr}}$ over the DIMS and MALDI spectra

| | Spectra pairs of same peptides | | | Spectra pairs of different peptides | | |
|---|---|---|---|---|---|---|
| | Mean | SD | $\alpha$ for $t$ test | Mean | SD | $\alpha$ for $t$ test |
| DIMS spectra | | | | | | |
| $S_{\mathrm{corr}}$ | 0.394 | 0.285 | <0.001 | 0.241 | 0.099 | <0.001 |
| $T_{\mathrm{corr}}$ | 0.517 | 0.308 | | 0.364 | 0.182 | |
| MALDI spectra | | | | | | |
| $S_{\mathrm{corr}}$ | 0.442 | 0.100 | <0.001 | 0.392 | 0.074 | <0.001 |
| $T_{\mathrm{corr}}$ | 0.759 | 0.115 | | 0.645 | 0.149 | |

For each dataset, the statistics were derived from scores over spectra pairs of same and different peptides. The $\alpha$ (columns 4 and 7) is the significance level of $t$ test on $T_{\mathrm{corr}}$ and $S_{\mathrm{corr}}$; and $T_{\mathrm{corr}}$ values are significantly higher than $S_{\mathrm{corr}}$s

In a series of objective experiments, we have established that high intensity transitions, regardless of the actual product ion identity, are a particularly effective alternative approach for peptide identification in the domain of spectral comparisons. Based on our results, we argue that ad hoc choices of most significant ions may have advantage over such existent strategy. Such an evaluation is particularly important, since ideally, one wants to reduce the number of transitions used in an MRM-experiment in order to assign and maintain high duty cycle to achieve reliable LC peak area quantitation.

Our results also demonstrate a consistent, albeit modest, enhancement of performance using $T_{\mathrm{corr}}$ versus the spectral standard dot product, across different experimental scenarios in three benchmark datasets. Though ROC performance may be low when only a few product ions are monitored (i.e., <4), particularly if the candidate peptide search space is large, $T_{\mathrm{corr}}$ performance is not significantly degraded when all putative y and b ions are considered. Hence, we conclude that our correlation algorithm can handle complex samples in a fairly robust manner, while detecting peptide compounds with added confidence. This characteristic of maintaining good performance with biological specimens has two major benefits: (1) biomarkers can potentially be detected with greater sensitivity due to reduction in possible interference by atypical or unspecific fragment ions and (2) analytical run times can be shortened as they need not necessarily have to be extremely highly resolved to establish stringent pattern matching. Short run times, with accurate results, are crucial for high-throughput clinical sampling of human tissue or body fluids samples from large patient cohorts.

The $T_{\mathrm{corr}}$ correlation distributions of true positive and false positive differ significantly when matching all spectra against the reference library of 384 peptides. The high frequency of true-positive hits at correlation scores 0.98–1.0 suggests that we have achieved almost perfect matching. Such correlations cannot be achieved when including all fragment ions, as with $S_{\mathrm{corr}}$. Even though this dataset is fairly small, the intensity pattern of a predefined selection of fragment ions seemed to be highly specific to each of the 384 different peptides in the database. This is especially interesting from the perspective of dealing with potentially overlapping (i.e., isobaric) precursor ions, since on the computational level, we treated the transition sets as being acquired simultaneously.

In this work, we used a subset of the highest scoring spectra, based on a standard SEQUEST database search, as references for our library comparisons. We have shown that other spectra derived from these same peptides usually have high pairwise spectral similarity in $S_{\mathrm{corr}}$ and particularly in the $T_{\mathrm{corr}}$ metric. In practice, noisy spectra can produce lower scores in library searches (10). However, we note that biased selection of reference spectra (i.e., pristine fragmentation patterns) may not generalize well, potentially resulting in degraded sensitivity in real world applications (i.e., with less than ideal MS/MS scans). Recent studies (11, 16) have proposed methods to derive more robust representative compilations from a group of related spectra mapping to the same peptide. We are currently conducting studies to determine whether such an integrative procedure can be imported to good effect with our $T_{\mathrm{corr}}$-based library search strategy.

Like other database search approaches, the performance of $T_{\mathrm{corr}}$ may not be satisfactory when the search space is large, i.e., there are a sizeable number of peptides sharing the same precursor *m/z* ratio. Figure 4 illustrates such a trend. The fundamental reason is that the false positives have greater chances to get high scores, regardless of scoring scheme, when the candidate peptide pool increases. On the other hand, additional information about peptide-specific properties, like chromatographic retention time, can alleviate this problem by substantially shrinking the search space. Recently, accurate retention time prediction (33–35) has been documented, and knowledge of retention time (36) is becoming a useful tool in peptide identification. Our additional experimental observations, as shown in Supplementary Fig. 4, demonstrate that peptide retention times are highly reproducible across different LC/MS runs. Our ongoing efforts aim at exploiting such information to boost the accuracy of $T_{\mathrm{corr}}$-based peptide identification and quantification.

In summary, this work demonstrates that empirically driven signature ion selection criteria combined with the $T_{\mathrm{corr}}$ correlation measure can be an effective approach to peptide identification through spectral library searching. Although we have based our conclusions on an analysis of spectra derived by ion trap TPM, our study also provides a solid rationale for applying $T_{\mathrm{corr}}$ to targeted biomarker tracking by MRM on a triple quadrupole, which produce spectra that are even more selective and therefore potentially less affected by product ion interferences. Our results of the DIMS experiment underline that specificity of matching intensity patterns of selected fragment ions is extremely high, when high-quality reference spectra are used. Nevertheless, we expect that including retention time, we can substantially, provide satisfactory specificity and sensitivity. Moreover, these highly specific proteomic data based on tandem mass spectrometry can be utilized for biomarker discovery and especially biomarker verification in respect of early stage detection of critical diseases such as cancers in complex sample matrices, where highest sensitivity is needed.

**Acknowledgement** The authors are grateful to Kento Onishi and Peter Zandstra for generously providing mouse stem cell extract. We also thank Clement Chung for his thoughtful discussions on computational-related issues. This study was supported in part by funds to A.E. from the Ontario Research Fund.

## References

1. Hoffmann DE, Stroobant V. Mass Spectrometry: principles and applications. 2nd Edition. John Wiley; 2001.
2. Verberkmoes NC, Bundy JL, Hause JL, Asano KG, Razumovskaya J, Larimer F, Hettich RL, Stephenson JL. Integrating top-down and bottom-up mass spectrometric approaches for proteomic analysis of *Shewanella onneidensis*. J Proteome Res. 2002;1:239–52.
3. McDonald WH, Yates JR. Shotgun proteomics and biomarker discovery. Dis Markers. 2002;18:99–105.
4. Vesvizhskii AI, Aebersold A. Interpretation of shotgun proteomic data: the protein interference problem. Mol Cell Proteomics. 2005;4:1419–40.
5. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003;17:2337–42.
6. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral database of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976–89.
7. Park CY, Klammer AA, Käll L, MacCoss MJ, Noble WS. Rapid and accurate peptide identifications from tandem mass spectra. J Proteome Res. 2008;7:3022–2027.
8. Eng JK, Fischer B, Grossmann J, MacCoss MJ. A fast SEQUEST cross correlation algorithm. J Proteome Res 2008;7:4598–602.
9. Perkins DN, Pappin JC, Creasy DM, Cottrell JS. Probability-based on protein identification by searching database using mass spectrometry data. Electrophoresis 1999;20:3551–67.
10. Halligan BD, Ruotti V, Twigger SN, Greene AS. DeNovoID: A web-based tool for identifying peptides from sequence and mass tags deduced from de novo peptide sequencing by mass spectrometry. Nucleic Acids Res 2005;33:376–81.
11. Frank A, Tanner S, Bafna V, Pevzner PA. Peptide sequence tags for fast database search in mass spectrometry. J Proteome Res 2005;4:1287–95.
12. Zhang Z. Prediction of low-energy collision induced dissociation spectra of peptides. Anal Chem 2004;76:3908–22.
13. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR. Similarity among tandem mass spectra from proteomic experiments: detection, significance and utility. Anal Chem. 2003;75:2470–7.
14. Liu J, Bell AW, Bergeron JJM, Yanofsky CM, Carrillo B, Beaudrie CEH, Kerney RE. Methods for peptide identification by spectral comparison. Proteome Sci. 2007;5:3.
15. Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. Proteomics 2004;4:950–60.
16. Tabb DL, Thompson MR, Khalsa-Moyers G, VerBermoes NC, McDonald WH. MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. J Am Soc Mass Spectrom 2005;16:1250–61.
17. Craig R, Corteins JC, Beavis RC. Using annotated peptide mass spectrum libraries for peptide identification. J Proteome Res 2006;5:1843–9.
18. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem 2006;78:5678–84.
19. Savitski MM, Nielse M, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post translational modifications, finding novel types of modifications, and fingerprint complex protein mixtures. Mol Cell Proteomics 2006;5:934–48.
20. Banderia N, Tsur D, Frank A, Pevzner PA. Protein identification by spectral networks analysis. PNAS 2007;104:6140–5.
21. Cox DM, Zhong F, Du M, Duchoslav E. Sakuma T, McDermott JC. Multiple reaction monitoring as a method for identifying protein posttranslational modifications. J Biomol Tech. 2005;16:83–90.
22. Hewel JA, Liu J, Onishi K, Fong V, Sandhu C, Talukder S, et al. High-resolution biomarker discovery: Targeted tandem mass spectrometry methods for quantitative validation of transcription factor candidates. Proceedings of the 56th ASMS Conference, Denver, CO, 2008.
23. Hewel JA, Liu J, Onish K, Fong V, Yue M, Sandhu C, et al. Targeted proteomics of transcription factors in breast cancer and embryonic stem cells. Proceedings of the HUPO 2008, 7th World Congress, Amsterdam, Netherlands, 2008.
24. Kulaingam V, Smith CR, Batruch I, Buckler A, Jeffrey DA, Diamandis EP. Product ion monitoring assay for prostate-specific antigen in serum using a linear ion-trap. J Proteome Res 2008;7:640–7.
25. Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, Schafer J, Rameseder J, Wong C-H, Deutsch EW, Brusniak M-Y, Buhlmann P, Bjorck L, Domon B, Aebersold R. Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. Mol Cell Proteomics. 2008;7:1489–500.
26. Fälth M, Svensson M, Nilsson A, Sköld K, Fenyö D, Andren PE. Validation of endogenous peptide identifications using a database of tandem mass spectra. J Proteome Res 2008;7:3049–53.
27. Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom. 2003;17:2310–6.
28. Xu H, Freitas MA. Monte Carlo simulation-based algorithms for analysis of shotgun proteomic data. J Proteome Res 2007;7:2605–15.
29. Krishnan V. Probability and random process. Wiley; 2006.
30. Wan Y, Yang A, Chen T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. Anal Chem 2006;78:432–7.
31. Zhang J, He S, Ling CX, Cao X, Zeng R, Gao W. PeakSelect: preprocessing tandem mass spectra for better peptide identification. Rapid Commun Mass Spectrom 2008;22:1203–12.
32. Sandhu C, Hewel JA, Badis G, Talukder S, Liu J, Hughes TR, Emili A. Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer. J Proteome Res 2008;7:1529–41.
33. Shinoda K, Sugimoto M, Yachie N, Sugiyama N, Masuda T, Robert M, Soga T, Tomita M. Prediction of liquid chromatographic retention time of peptides generated by protease digestion of the *Escherichia coli* proteome using artificial neural networks. J Proteome Res 2006;5:3312–7.
34. Klammer AA, Yi X, MacCoss MJ, Noble WS. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. Anal Chem. 2007;79:6111–8.
35. Krokhin OV, Ying S, Cortens JP, Ghosh D, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA. Use of peptide retention time prediction for protein identification by off-line reversed phase HPLC-MALDI MS/MS. Anal Chem 2006;78:6265–59.
36. Sun W, Zhang L, Yang R, Shao C, Zhang Z, Gao Y. Improving peptide identification using an empirical peptide retention time database. Rapid Commun Mass Spectrom. 2009;23:109–18.