DOI 10.1002/pmic.200800564

Research Article

Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures

Guo-Zhong Li[1], *Johannes P. C. Vissers*[2], *Jeffrey C. Silva*[1]*, *Dan Golick*[1], *Marc V. Gorenstein*[1] *and Scott J. Geromanos*[1]

[1] Waters Corporation, Milford, MA, USA
[2] Waters Corporation, Manchester, UK

A novel database search algorithm is presented for the qualitative identification of proteins over a wide dynamic range, both in simple and complex biological samples. The algorithm has been designed for the analysis of data originating from data independent acquisitions, whereby multiple precursor ions are fragmented simultaneously. Measurements used by the algorithm include retention time, ion intensities, charge state, and accurate masses on both precursor and product ions from LC-MS data. The search algorithm uses an iterative process whereby each iteration incrementally increases the selectivity, specificity, and sensitivity of the overall strategy. Increased specificity is obtained by utilizing a subset database search approach, whereby for each subsequent stage of the search, only those peptides from securely identified proteins are queried. Tentative peptide and protein identifications are ranked and scored by their relative correlation to a number of models of known and empirically derived physicochemical attributes of proteins and peptides. In addition, the algorithm utilizes decoy database techniques for automatically determining the false positive identification rates. The search algorithm has been tested by comparing the search results from a four-protein mixture, the same four-protein mixture spiked into a complex biological background, and a variety of other "system" type protein digest mixtures. The method was validated independently by data dependent methods, while concurrently relying on replication and selectivity. Comparisons were also performed with other commercially and publicly available peptide fragmentation search algorithms. The presented results demonstrate the ability to correctly identify peptides and proteins from data independent acquisition strategies with high sensitivity and specificity. They also illustrate a more comprehensive analysis of the samples studied; providing approximately 20% more protein identifications, compared to a more conventional data directed approach using the same identification criteria, with a concurrent increase in both sequence coverage and the number of modified peptides.

Received: July 3, 2008
Revised: October 16, 2008
Accepted: October 20, 2008

Keywords:
Database searching / Data independent LC-MS / Multiplexed LC-MS / Parallel LC-MS / Shotgun sequencing

Correspondence: Dr. Johannes P. C. Vissers, Waters Corporation, Atlas Park, Simonsway, Manchester M22 5PP, UK
E-mail: hans_vissers@waters.com
Fax: +44-161-435-4444

Abbreviations: DDA, data dependent analysis; **MRM,** multiple reaction monitoring; **ROC,** receiver operating characteristic

1 Introduction

Traditional mass spectrometric approaches for the identification of peptides from enzymatically digested proteins

* Current address: Cell Signaling Technology, Inc., 3 Trask Lane, Danvers, MA 01923, USA.

include MALDI-TOF [1–6] for single or simple protein mixtures and ESI MS/MS [7–9] and LC-MS/MS [10] for more complex mixtures. As the complexity of the samples increases in terms of absolute number, dynamic range, and molecular weight, the use of accurate mass measurements alone, as utilized by MALDI-TOF PMF, does not provide enough specificity to impart unambiguous identification. Allowing for missed cleavages and/or PTMs can compound this. In the instances where the accurate mass measurements of the proteolytic peptides do not provide adequate specificity for successful protein identification, samples are typically analyzed by ESI data dependent analysis (DDA) using MS/MS in conjunction with LC [10, 11].

The major advantage of LC-MS/MS-based DDA experiments is the generation of primary structure information. The added specificity afforded by fragment ions has proven to be effective for the identification of proteins [7]. Although LC-MS/MS is more efficient than MALDI-TOF PMF at identifying proteins in complex matrices, inherent limitations are associated with the technique. A DDA analysis cycle typically starts with an MS survey scan and is followed by the selection/isolation of a precursor ion, or by sequentially isolating a number of precursor ions, for subsequent MS/MS experiment(s) by either CID [12–14], electron transfer dissociation (ETD) [15], or electron capture dissociation ECD [16]. Hence, the experimental strategy is by design a serial process, which results in a competition to acquire as many MS/MS spectra to as many precursor ions as possible in a given period of time. To accomplish this, the time allotted for an MS/MS acquisition is usually compromised.

A mass spectrometer is typically configured to alter from MS to MS/MS acquisition mode for those peptides that exceed a minimum intensity threshold. However, the ionization efficiency of the detectable precursor ions from a constituent protein spans approximately three orders of magnitude, with the majority of the precursor ions present in the lowest intensity regime. The presence of medium and high abundance proteins distract the mass spectrometer from conducting MS/MS experiments on peptides of lower abundant proteins and on the lower abundant, poorly ionizing peptides of the medium, and high abundant proteins themselves. Since an MS/MS acquisition is typically conducted for a short period of time relative to the peak width of a peptide, it is unlikely that a product ion spectrum at the apex of the originating precursor is obtained. It is therefore not uncommon that insufficient spectral product ion quality is obtained for a valid peptide assignment from low signal intensity precursors in combination with fast MS/MS scanning rates. Technical approaches and their implications to improve the S/N of an MS/MS experiment have been discussed previously in great detail and length [17].

Different commercial, public, and open access search engines are available for protein database searches and identification. These include the crosscorrelative approaches of SEQUEST [18] and Global Proteome Machine [19] and the probability-based strategies used by MASCOT [20] and ProteinLynx Global*SERVER* [21]. There is however no consensus on what the initial search parameters should be. It has been suggested that the proteolytic digestion of proteins in complex matrices is not specific enough for enzyme specific databases searches [22, 23]. Other reports imply however that enzymes are specific and reproducible and as such use stringent enzyme specificity [24, 25]. Missed cleavages and variable modifications are other criteria that are used inconsistently. Furthermore, small and large mass window search tolerances are used interchangeably.

The selectivity and specificity of a database search are compromised with every missed cleavage and allowed variable modification. For instance, a database can exponentially increase in size from 1.2 million to over 30 million theoretical peptides with no enzyme specificity, one missed cleavage, and oxidized methionines, without the use of decoy databases [26]. Considering these type of statistics, if used inconsistently, the replication rate of identifications from an identical sample run on different instruments, in different laboratories, can be interpreted as being relatively low [27].

The mass accuracy of both precursor and fragment ion data can have a profound impact on the interpretation of the results and yet the identification algorithm used for the peptide/protein identification may not appropriately consider these relevant statistics. Product mass accuracy provides increased selectivity and specificity for peptide fragment ion spectra identification compared to precursor mass accuracy only [28]. The accuracy of the peptide assignments can be improved in a number of ways, including additional MS/MS data processing [29], improved charge-state determination [30], low quality MS/MS data removal [31], redundant MS/MS spectra clustering [32, 33], and the use of more sophisticated scoring schemes such as those implemented in OLAV [34], SALSA [35], and EPIR [36]. In addition, correlative algorithms have been described which integrate retention time by introducing hydrophobicity prediction models [37]. While the methods for protein identification from MS/MS data described above have been successfully demonstrated in a variety of applications, the quality of the output of these algorithms is directly related to the quality of the MS/MS fragmentation spectra. This is particularly so when a single MS/MS spectrum is used for the identification of a biologically relevant protein. Sometimes reversed or randomized databases are used in conjunction with a species-specific database for monitoring false positive rates [28, 38]. More often, however, searches are conducted with just a species-specific database.

Limited mass measurement accuracy further reduces the selectivity and specificity of a given method. As an example, a human database of 24 179 proteins contains over 800 000 tryptic peptides without considering missed cleavages. On average, a single precursor mass will match to approximately 600 peptides in the database if the query is based on nominal mass (±1 Da) and the analysis mass range is between 750 and 4750 amu. Similarly, with an average number of 250 product ions *per* MS/MS experiment and a ±1 Da fragment ion mass bin, close to 50% of a typical MS/MS fragment ion window from 50 to 1700 Da is populated. The selectivity and specificity of these types of database queries, using nominal mass and pattern matching, are arguably an apparent function of peptide length [28]. A probability of greater specificity should not be expected to be obtained with larger peptides to compensate for this bias since approximately 70% of the tryptic peptides in any given proteome are less than 1750 molecular weight.

Recent literature reports suggest that a significant number of proteins identified using fast scanning MS/MS instrument types are based on one or two peptide-based protein matches, despite the fact that a very large number of MS/MS spectra were acquired [39–41]. The one or two peptide-based identifications accounted for approximately 40% of the total number of reported proteins. Interestingly, however, approximately two-thirds of the DDA experiments were acquired on m/z values at the same retention times as data acquired from replicate injections on multiple instruments and instrument types. These observations cast doubt on the claimed serendipitous nature of DDA. This is, however, to be expected since the relative protein concentration of the sample and ionization rates are constant. Hence, the majority of the data should match to the same proteins if the majority of the components that led to an MS/MS event are the same. In cases where this is not observed, this is clearly indicative of either an alternative fractionation strategy or of incorrect identifications and challenges the belief that the analysis of the same sample run on similar or different instruments should provide complementary results [28].

A novel database search strategy is described that is designed for data independent accurate mass acquisitions. The strategy is based upon utilizing physicochemical attributes associated to peptides, separated by RP LC and gas phase mass analyzed by quadrupole TOF-MS. The algorithm employs a hierarchal database search strategy in which tryptic peptides are tentatively identified according to initial search parameters. These tentative peptide identifications are ranked and scored by how well they conform to a number of predetermined, physicochemical attribute models. If the number of identified peptides exceeds a minimum initial score, retention time, and fragmentation models are refined and used to further score the identified peptides. All tentative peptides are collapsed into their parent proteins utilizing only the highest scoring peptides that contribute to the total protein score. A depletion algorithm is subsequently used to prevent the peptide detections from the highest scoring protein, typically the most abundant one, from being considered in the subsequent identification of less abundant proteins. Once a protein has been securely identified, all top ranked precursor ions and their corresponding product ions are removed from all other tentatively identified proteins. The remaining unidentified peptides, and tentatively identified proteins, are then reranked and rescored, and the process is repeated until the user-defined false positive rate, based on identification of random or reverse proteins, is obtained. A subset database is created from the validated protein identifications and the next iteration initiated. The second iteration allows for the identification of in-source fragments, the neutral loss of H_2O and $NH_{3,}$ missed cleavages, oxidized methionines, deamidations, N-terminal alkylations, and any other user-defined modifications. The processing software sets automatically the initial search parameters, including the precursor and product ions mass tolerances as well as a product-to-precursor ion time-alignment window. Since selectivity and specificity increase with increasing product ion mass, only time-aligned fragment ions whose mass is greater than y_3 and b_3 are scored. In those instances where the experiment is conducted in replicate, the resulting database searches can be combined, which further increases the significance of identification. The decision to combine data from technical replicates is predicated on the identification of the same peptide sequence at the expected retention time and similarity of time-aligned product ions. The underlying premise guiding the decision is based on the notion that signal replicates, and noise does not.

Utilizing an absolute quantification strategy [42], data are presented to illustrate the algorithm's ability to correctly identify proteins across a dynamic range of nearly three orders of magnitude. The replication rate of proteins, peptides, and product ions exceeds 75% in data sets of multiple injections of the same or similar samples. Using the absolute quantification capabilities of the acquisition/analysis process, the data presented illustrate protein sequence coverage to be commensurate with its relative abundance in the sample matrix [43]. In addition, the results depict how coeluting precursors sharing similar product ions are deconvolved *via* time alignment to their parent peptide and protein.

2 Materials and methods

2.1 Sample preparation

Fifty microliters of 0.5% aqueous formic acid was added to 100 μg of cytosolic *Escherichia coli* digest standard (Waters Corporation, Milford, MA, USA). A tryptic digest stock solution containing four standard proteins, alcohol dehydrogenase, phosphorylase B, albumin, and enolase, was prepared in 0.1% aqueous formic acid and diluted to concentrations of 200, 200, 200, and 100 fmol/μL, respectively. Equal volumes of the *E. coli* digest and the standard proteins were combined to give a sample concentration of 0.5 μg/μL of *E. coli* digest and 100, 100, 100, and 50 fmol/μL of alcohol dehydrogenase, phosphorylase B, albumin, and enolase, respectively. The tryptic digests of the four proteins were also prepared in 0.1% aqueous formic acid without the presence of the *E. coli* digest standard at the same concentration level of 100, 100, 100, and 50 fmol/μL, respectively. Unless stated otherwise, these solutions were used as stocks for all the experiments described in this manuscript.

2.2 LC-MS configuration

Nanoscale LC separation of tryptic peptides was performed with a nano-ACQUITY system (Waters Corporation), equipped with a Symmetry C_{18} 5 μm, 5 mm × 300 μm precolumn and an Atlantis C18 3 μm, 15 cm × 75 μm analytical RP column (Waters Corporation). The samples, 1 μL full loop injection, were initially transferred with an aqueous 0.1% formic acid solution to the precolumn at a flow rate of 4 μL/min for 3 min. Mobile phase A was water with 0.1% formic acid while mobile phase B was 0.1% formic acid in ACN. After desalting and preconcentration, the peptides were eluted from the precolumn to the analytical column and separated with a gradient of 3–40% mobile phase B over 90 min at a flow rate of 300 nL/min, followed by a 10 min rinse with 90% of mobile phase B. The column was reequilibrated at initial conditions for 20 min. The column temperature was maintained at 35°C. The lock mass compound, [Glu^1]-fibrinopeptide B, was delivered by the auxiliary pump of the LC system at 250 nL/min at a concentration of 100 fmol/μL to the reference sprayer of the NanoLockSpray source of the mass spectrometer. All samples were analyzed in triplicate.

Mass spectrometric analysis of tryptic peptides was performed using a Q-TOF Premier mass spectrometer (Waters Corporation, Manchester, UK). For all measurements, the mass spectrometer was operated in v-mode with a typical resolution of at least 10 000 FWHM. All analyses were performed in positive mode ESI. The TOF analyzer of the mass spectrometer was externally calibrated with a NaI mixture from *m*/*z* 50 to 1990. The data were postacquisition lock mass corrected using the doubly charged monoisotopic ion of [Glu^1]-fibrinopeptide B. The reference sprayer was sampled with a frequency of 30 s. Accurate mass LC-MS data were collected in an alternating, low energy, and elevated-energy mode of acquisition [17, 44]. The spectral acquisition time in each mode was 1.5 s with a 0.1 s interscan delay. In low energy MS mode, data were collected at constant collision energy of 4 eV. In elevated-energy MS mode, the collision energy was ramped from 15 to 40 eV during each 1.5 s integration. One cycle of low and elevated-energy data was acquired every 3.2 s. The RF amplitude applied to the quadrupole mass analyzer was adjusted such that ions from *m*/*z* 300 to 2000 were efficiently transmitted, ensuring that any ions observed in the LC-MS data less than *m*/*z* 300 were known to arise from dissociations in the collision cell.

Quantification experiments were conducted with nanoscale LC-MS/MS using a tandem quadrupole system. The Quattro Premier XE mass spectrometer (Waters Corporation) was operated in the multiple reaction monitoring (MRM) mode of analysis and using the same chromatographic conditions described above. The transmission window of both mass analyzers was typically 1 Da, the dwell time 25 ms and the collision energy approximately 20 eV.

2.3 Ion accounting process

2.3.1 Ion detection

The raw data from the three data functions, low energy, elevated energy, and lock spray, were processed as previously described [45], with minor modifications. The result of this process is a time-aligned inventory of accurate mass-retention time components for both the low and elevated-energy data [17]. Included for each component is the monoisotopic accurate mass, a calculated mass deviation, the summed peak area of all isotopes of all charge states, a calculated area deviation, the apex retention time, the chromatographic peak start and end retention times, and average fractional charge state.

2.3.2 Time alignment

The low and elevated-energy accurate mass-retention time components are time aligned into precursor/product ion tables upon completion of the ion detection process. In the case of coeluting peptides, the same fragment ions will be initially assigned to multiple precursors. However, the origin of the fragment ions is later facilitated by the inherent mass accuracy present within the data, and the ability of the search engine to deplete product ions from securely identified peptides and proteins throughout the iterative search process, so as not to interfere with the identifications of less abundant peptides/proteins. The basis of the time-alignment algorithm is that elevated-energy ions whose calculated apex retention times equal the apex retention time of a low-energy accurate mass-retention time component, plus or minus one-tenth of the chromatographic peak width of the low energy component, are associated. Each elevated-energy accurate mass-retention time component can be associated with multiple low energy components. The third element in the process flow diagram shown in Fig. 1 represents the time alignment process.

2.3.3 Filtering

The time aligned precursor/product ion tables are filtered before submission to the search algorithm. The filtering process eliminates all low energy precursors under 750 Da and all elevated-energy product ions under 350 Da. These masses are excluded because the afforded selectivity and specificity is low. These ions are reclaimed in a later iteration of the search algorithm, after the constituent proteins have been determined. Typically, all proteins generate tryptic peptides under 750 Da; however, these peptides have high sequence similarity and are therefore not specific for a given protein sequence. Moreover, sequence specificity increases with peptide length. Numerous peptides start and end with the same three amino acids. As such, the sequence selectivity

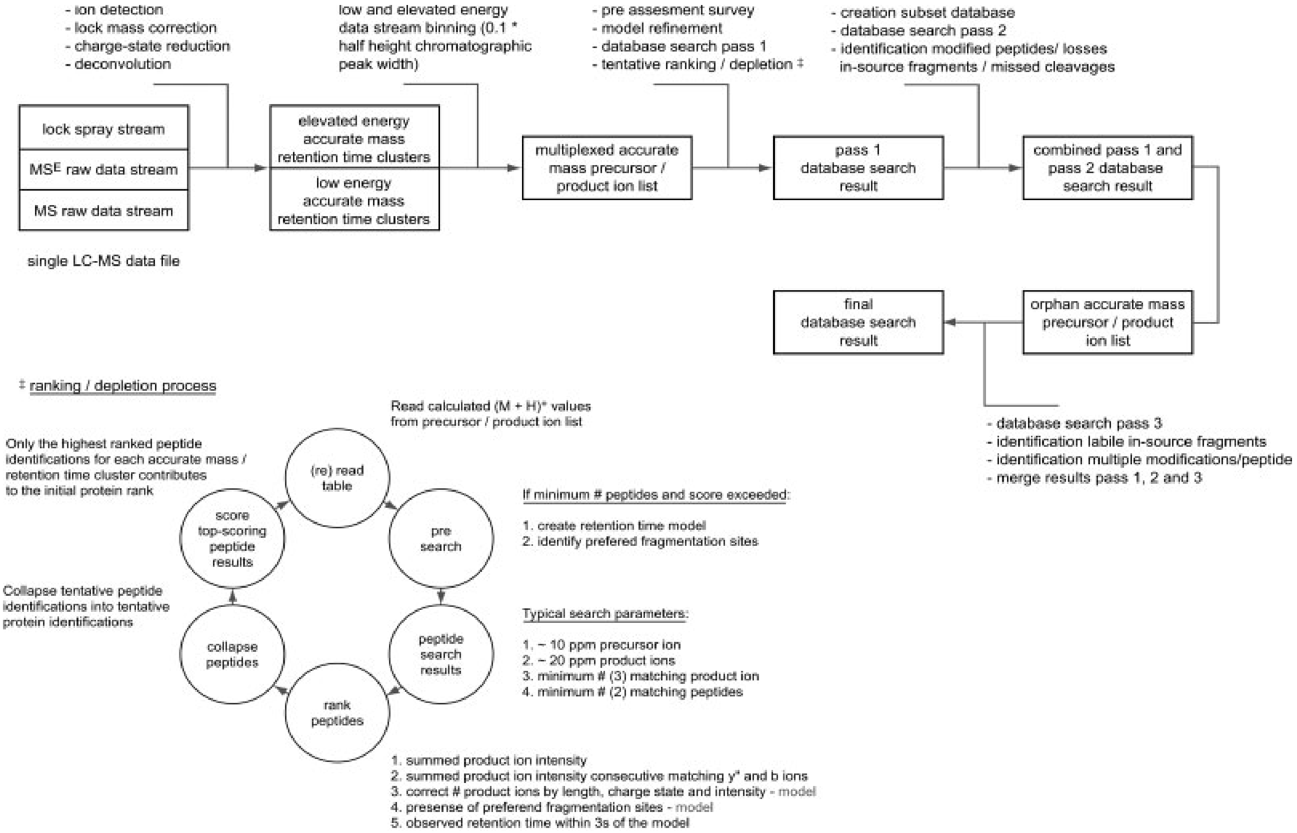
- ion detection
- lock mass correction
- charge-state reduction
- deconvolution
low and elevated energy
data stream binning (0.1 *
half height chromatographic
peak width)
- pre assesment survey
- model refinement
- database search pass 1
- tentative ranking / depletion ‡
- creation subset database
- database search pass 2
- identification modified peptides/ losses
in-source fragments / missed cleavages
lock spray stream
MSE raw data stream
MS raw data stream
single LC-MS data file
elevated energy
accurate mass
retention time clusters
low energy
accurate mass
retention time clusters
multiplexed accurate
mass precursor /
product ion list
pass 1
database search
result
combined pass 1 and
pass 2 database
search result
final
database search
result
orphan accurate mass
precursor / product
ion list
- database search pass 3
- identification labile in-source fragments
- identification multiple modifications/peptide
- merge results pass 1, 2 and 3
‡ ranking / depletion process
Read calculated (M + H)+ values
from precursor / product ion list
Only the highest ranked peptide
identifications for each accurate mass /
retention time cluster contributes
to the initial protein rank
(re) read
table
pre
search
peptide
search
results
rank
peptides
collapse
peptides
score
top-scoring
peptide
results
Collapse tentative peptide
identifications into tentative
protein identifications
If minimum # peptides and score exceeded:
1. create retention time model
2. identify prefered fragmentation sites
Typical search parameters:
1. ~ 10 ppm precursor ion
2. ~ 20 ppm product ions
3. minimum # (3) matching product ion
4. minimum # (2) matching peptides
1. summed product ion intensity
2. summed product ion intensity consecutive matching y* and b ions
3. correct # product ions by length, charge state and intensity - model
4. presense of preferend fragmentation sites - model
5. observed retention time within 3s of the model

Figure 1. Workflow ion accounting database search algorithm illustrating the data processing (first three elements), the iterative search process (following four elements) and the peptide ranking and the ion depletion iteration of the tentative pass 1 identifications (genuine tryptic peptides, including fixed modifications).

and specificity of y_1 through y_3 and b_1 through b_3 do not contribute considerably. In the absence of other molecular ions within the ion transmission window of the first mass analyzer, y_1 through y_3, b_1 through b_3, immonium ions and any other low molecular weight product ions have significance [46]. Further filtering includes the removal of product ions higher in intensity than the precursor and the removal of product ions higher in mass than the precursor. In the case of in-source fragmentation, however, a fragment ion can be higher in intensity than its precursor. If the precursor peptide is labile enough to fragment during the ionization/ion transfer process, in-source fragment ions will produce an ion series identical to the true parent in the collision cell. The sum of similar product ion intensities, these ions are entering the collision cell at the same moment of time, can exceed the intensity of the residual precursor. The identification of in-source fragments is conducted during the second pass of the database search strategy.

2.3.4 Selecting a searchable database

Prior to submission of the filtered precursor/product ion tables to the database search algorithm, a database is specified. Either a reverse or random database can be generated for monitoring the false positive rate of a particular search. A reverse decoy database is produced by reversing the sequence of each protein from the original database. A randomized version of the query database is generated by randomizing the protein sequences, while holding the peptide amino acid composition and the number of tryptic peptides constant. With a random decoy database, each protein can be *x*-fold randomized, where *x* is a user-defined integer. In either case, the decoy database is merged with the original database and subsequently used to conduct the database search with a user-defined “false positive” rate. If neither option is selected, the algorithm creates by default a one-fold randomized decoy database and appends it to the user-selected query database. The created combined database will be queried and the algorithm will automatically determine the minimum score for a protein to be reported.

2.3.5 Preassessment survey

Prior to the database query, a presearch of the time-aligned precursor/product ion lists is conducted with the same protein sequence database, search parameters, and scoring/

ranking system as will be employed during the actual query. During the presearch, the algorithm creates and/or adjusts some of the model parameters related to the gas and liquid phase physicochemical properties of peptides. To account for any experimental variation, a minimum number of peptides, approximately 250, exceeding a minimum score must be identified for the algorithm to develop new models. If these criteria/thresholds are not satisfied, the default model parameters will be applied.

A retention time model is generated in real time using the identifications from the preassessment survey, for all peptides residing in the database, using a least squares fit to the observed elution times of the identified peptides. Additionally, a monoisotopic product ion mass distribution is generated for all unambiguous peptide identifications. The algorithm subsequently produces an optimized fragmentation model by comparing the experimentally derived frequency distribution of the product ions with the number of times a specific bond cleavage was identified. Further fine-tuning of the algorithm is performed by comparing the sequence length, charge state, and precursor intensity to the summed number of identified product ions. The summed intensity of these product ions is related to the precursor ion intensity and the total number of continuous and complementary identified *y* and *b* ions. Furthermore, the algorithm compares the composition of peptides and the location of specific residues with the observed precursor charge state, or states, and the summed *y*/*b* ion intensity ratios.

2.3.6 Database search pass 1

Other than the selection of the protein sequence database and an acceptable false positive rate, the software can provide all other database search parameters. Parameters settings can be manually modified and customized. The database search is a hierarchal process and analogous to the iterative search process described for the mapping and accounting of spectral MALDI components in the case of relatively simple protein mixtures [47]. Here, an iterative search strategy is described for data-independently acquired LC-MS data. During the first iteration, each parent/product ion table is queried against the protein sequence database considering only completely cleaved tryptic peptides for identification. For a peptide to be putatively identified it has to fit within the precursor ion mass tolerance, typically $<$10 ppm, and contain at least three fragment ions within the product ion mass tolerance, typically $<$20 ppm. These tentative peptide identifications are scored, based upon how well they correlate to a number of different models, using the physicochemical attributes of peptides in both the gas and liquid phase.

The number of matched product ions is initially compared to a model, which predicts the number of product ions that a tryptic peptide of a given length, charge state, and intensity should produce. The initial score is adjusted based upon the correlation of the matched data to the model. Due to the nature of the acquired data, and the extremely large number of peptide identifications that are considered up to this point, the fragmentation model previously generated in the presearch is consulted to increase or decrease the initial score by the presence or absence of preferred fragmentation sites. In addition, a Markov Chain extension is employed whereby a continuous ion series increases the score. In the chain extension model, higher weighting (increased score) is given to continuous, higher molecular weight product ions (those higher in molecular weight then the multiply charged precursor). Other physicochemical attributes that contribute to the peptide score are (i) the presence of certain amino acids at or near the N-terminus, affecting the ratio of the total *y* ion intensity to the total *b* ion intensity, (ii) the presence of product ions corresponding to loss of H_2O and NH_3 from peptides containing specific amino acids, (iii) complementary C or N terminal product ions, (iv) how well the tentative amino acid sequence supports the observed peptide charge states, and (v) how well the experimental retention time matches the theoretical retention time, assigned to that peptide from the preassessment search. The highest scoring tentative peptide identification for a given peptide in the precursor/product ion table is given a peptide rank value of one. If other possible identifications for a given precursor have scores very near that of the highest score, all tentative identifications are also given the rank value of one. Other possible identifications for the precursor (multiplexed spectrum) are given the rank value of 0 at this time.

After ranking, the tentative peptide identifications are collated into tentative protein identifications. Initially, protein identifications are ranked by summing the intensity of product ions arising from the tentative peptide identifications with a rank value of one for that protein. The putatively identified protein with the highest total product ion intensity is given the rank of one with the other proteins ranked in descending order by this measure. At this point the proteins are scored. Initially, the protein score is derived from the matching tryptic peptides with a rank value of 1 and normalized to the length of the protein, and the total intensity of the three best ionizing peptides. This initial protein score is adjusted by consideration of various physicochemical attributes of proteins. The first adjustment to the score is made by comparing the total number of product ions associated to that protein with the expected number of product ions from a protein of that length and total precursor ion intensity. Secondly, the ionization efficiency distribution of identified precursors to that protein is compared to a model based on proteins of similar molecular weight and concentration.

Other contributing attributes include sequence coverage, the total number of continuous and complementary *y* and *b* ions, the ratio of total product ion intensity to the total precursor ion intensity, and the number of ions observed at preferred fragmentation sites. Upon completion of the protein scoring process, the highest scoring protein is considered identified and all precursor and product ions associated with the top ranked peptide sequence identifications for that protein are excluded from all subsequent protein identifications. The ranking and scoring process is repeated until either no protein identification exceeds the minimum score or the acceptable false positive rate has been exceeded. To determine the false positive rate, the number of proteins identified prior to the highest scoring reverse/random identification is counted. This number is multiplied by the chosen acceptable false positive rate to determine the maximum number of reverse/random identifications allowed. This calculation is repeated with each subsequent reverse/random identification encountered in the protein list until the number of actual reverse/random identifications equals the calculated allowable number, at which point no further protein identifications are allowed.

2.3.7 Database search pass 2

The second pass of the database algorithm is designed to identify peptide modifications, and nonspecific cleavage products to proteins positively identified in the first pass. The software realigns the low and elevated-energy accurate mass-retention time components that were not identified in pass one and constructs a temporary subset database of the proteins identified from pass one. During the creation of the subset database, the software allows for all peptides associated with the pass one proteins to exist in modified forms. These include in-source fragments, in-source loss of H_2O and NH_3, missed cleavages, oxidized methionines, deamidations, N-terminal alkylations, and other variable modifications, including PTMs. The first iteration of the second pass of the algorithm specifically looks for in-source fragments and their unique constituent fragment ions. Identification of in-source fragments is confirmed by the accurate mass of the fragments, and the fact that the apex retention time and chromatographic peak width of the fragments must be the same as that of the originating precursor. The addition of unique, elevated-energy ions, increases the validity of protein identifications and allows for further depletion of elevated-energy ions from other proteins and peptides prior to the next iteration. The second iteration of the second pass specifically looks for precursor ions that have lost H_2O or NH_3. Again, these identifications are confirmed by mass accuracy and the fact that their apex retention time must be the same as that of the precursor, plus/minus one-tenth of the chromatographic peak width of the precursor. The final iteration at this stage identifies missed cleavages, oxidized methionines, deamidations, and other variable modifications. Low-energy accurate mass-retention time components may be assigned to multiple variant peptides in these iterations. The results are filtered such that the variant containing the largest number of matched fragment ions, with additional weighting placed on the peptides that have fragments indicating the point of modification, is selected as the best match. Any of these variants not having matched elevated-energy accurate mass-retention time components are excluded.

2.3.8 Database search pass 3

Following completion of pass two, the algorithm realigns the remaining low and elevated-energy accurate mass-retention time components that were not identified from the first two passes. The precursor/product ion tables are again searched against the complete database, without any restriction on product ion intensity. Highly labile peptides producing in-source fragments are identified during this stage of the database strategy. In this case, the total product ion intensity can exceed the intensity of the precursor.

In addition, the assignment of all possible ions to the abundant proteins increases the sensitivity and selectivity of the identification of less abundant proteins based on the remaining ions. For example, additional multiple, variable modifications *per* peptide and single-point amino-acid modifications are identified at this stage. In a similar manner to pass one, all tentative peptide identifications are scored, ranked, and subsequently collapsed into proteins identifications. The proteins are scored and ranked and the depletion process repeated until a minimum protein score can no longer be maintained or the specified false positive rate of identification is breached.

2.3.9 Output

The final output is a table containing all the theoretical and experimental attributes associated with each product ion for each precursor ion of each protein identified, including those calculations made to determine the peptide sequence and protein scores. In addition, an output of the initial search can be provided. This table contains the information on every ion that could have been tentatively assigned to any peptide for any protein.

3 Results and discussion

3.1 Physicochemical properties of proteolytic peptides

Since all proteolytic peptides are constructed from the same 20 commonly occurring amino acids, they are expected to display similar physicochemical properties. This has been demonstrated in chromatography research by retention time prediction models that incorporate amino acid composition and sequence, overall hydrophobicity, chain length, chemical nature of termini, and p*I* [37, 48–52]. Similarly, prediction models for gas-phase fragmentation spectra have been suggested using either empirically or theoretically derived rules based upon the primary amino acid sequence [53–58]. It is rare, however, for both to be utilized simultaneously for the purpose of peptide identification from LC-MS data. A number of physicochemical properties will be cited that are considered by the algorithm described in this paper for the purpose of validating protein identifications.

3.1.1 Intact protein MW, intensity, and number of detected peptides (ionization efficiency)

A previous study demonstrated that the relationship between the number of tryptic peptides produced by digestion and detected by LC-MS and the intact protein molecular weight is nearly linear [42]. The number of peptides that can be identified increases with protein molecular weight and concentration. Also there is a relationship between the total intensity of the three most intense tryptic peptides identified to a protein and the molar amount of digested protein injected on-column [42]. Supporting Information Fig. S1 shows the identified peptides to 100 fmol of glycogen phosphorylase, sorted in descending intensity that replicated in three out of three injections. The reproducibility of peptides replicating across three injections is typically 85%, compared to the total number of peptides identified on an individual injection basis. By expressing the number of identified peptides to a protein as a linear function of the molecular weight of the protein, it is possible to calculate an expected number of peptides that may be identified, for any given protein of any molecular weight at the same concentration. Table 1 summarizes the results for four tryptic digested protein standards, in which one of the proteins was used as a reference for the prediction of the number of peptides that should be identified for other proteins at the same or similar concentration. The results presented indicate that a good correlation was obtained for the number of predicted *versus* the number of identified peptides. Based on this observation, the number of identified peptides for a protein of a given molecular weight, at a given concentration, can be predicted and used as an indicator or rule to support the correct identification of a protein [43].

The composition of a sample in terms of its constituent proteins and their corresponding concentrations in the sample is constant and independent of the instrument type used for analysis. Furthermore, the best ionizing peptides produced by ESI are very often the best ionizing peptides, independent on interface design, and type of mass analyzer. Label free or isotopically labeled quantification strategies could not be utilized for protein quantification if the latter two arguments did not hold. Moreover, the tryptic peptides produced from a protein do not ionize with equal efficiency, nor do they elute at the same moment in time. In practice, they typically elute across the chromatographic profile and exhibit a range of ionization efficiencies, with the majority of the peptide intensities lying in the lower third of the range. Hence, the most abundant proteins should be identified with the highest sequence coverage, as the number of peptides that can be identified to a protein is directly proportional to the molecular weight and the on-column concentration of the protein. It should also be noted that the number of peptides that can be identified to a protein, with increasing concentration, will reach an asymptotic maximum. This, simply because the maximum number of tryptic peptides that can be identified has been reached. The number of identified peptides to a protein is the basis for some LC-MS spectral counting-based absolute quantification algorithms [43, 59–62], which can bias the results obtained by this method, for the above reason. Moreover, the ionization intensity distribution shown in Supporting Information Fig. S1 illustrates that the difference in ionization efficiency between any two adjacent peptides is less than a factor of two. As such, there should always be more than one identifiable peptide to a protein; unless the applied method is capable of sampling and qualitatively identifying signals at the absolute LOD.

This behavior is not observed in a DDA experiment due to the inherent duty cycle limitations and partial peak sampling. This is clearly apparent in DDA experiments where there is a prevalence of single-peptide-based protein identifications. This is not because a DDA method is more efficiently sampling the lowest levels of signal intensity, but due to the compromising nature of the technique. That is, if the MS signals of all detectable precursor ions, not only the ones selected/isolated for a collision induced fragmentation experiment, were processed in a similar fashion to those acquired by parallel, alternate scanning [17, 44], in which the chromatographic peak area for every precursor ion was calculated, it would be observed that DDA acquisitions sample the lower level signals very inefficiently [17]. This observation is platform independent and will be described in detail in a subsequent paper. For this reason, the ability to use the number of peptides identified, or the number of times a

Table 1. MW *versus* number of identified peptides (at a fixed concentration)

Protein	MW (kDa)	No. of observed peptides	Response factor	No. of predicted peptides	No. of observed/ no. of predicted peptides
Glycogen phosphorylase[a)]	97	56	0.577	56	1[a)]
Serum albumin	70	40		40.4	1.01
Alcohol dehydrogenase	37	21		21.4	1.02
Enolase[b)]	47	16		(27·1/2) 14	1.14

a) Reference protein.
b) The Enolase concentration is half that of the other standard proteins throughout the manuscript.

peptide is identified, from a DDA experiment as a means for determining the concentration of a protein at the lower levels of sensitivity is also limited.

Further, it is well understood that processed or degraded proteins may exist in complex samples. As a result, these proteins may not conform to the theoretical number of peptides that should be identified based on the criteria described here. The presented algorithm however does not accept or reject a protein identification solely on conformity to any single physicochemical property. Moreover, the position for each identified peptide in the protein sequence is annotated in the output. In these instances, the location of the securely identified peptide sequences, relative to the protein, can be confirmed either to the N- or C-terminus, thereby validating processing or degradation.

3.1.2 Peptide chain length, mass, charge state, and total product ion intensity

Supporting Information Fig. S2 and Table S2 show combined chromatographic and peptide properties and ionization efficiency characteristics. Supporting Information Figs. S2a and b show the peptide precursor *m/z* and charge state as a function of retention time, respectively. Peptides typically elute from an RP column in order of increasing hydrophobicity, peptide chain length, and charge state [63]. Superimposing Supporting Information Figs. S2a and b possibly provide a tool to model the chromatographic behavior of peptides and be employed in identification scoring schemes. However, the search algorithm presented in this paper makes use of a more sophisticated retention time prediction model to rule out false positive identification, which will be discussed in a subsequent paragraph.

Figure 2a illustrates a reasonably linear relationship at the peptide level between the intensity of a precursor ion to the total product ion intensity of all product ions matched to that precursor. Figure 2b shows an even closer linear relationship at the protein level between the total precursor ion intensity of all precursor ions assigned to a protein and the total product ion intensity of all matching product ions from the matched precursors. The ability to generate such relationships is enabled by the ability to determine the chromatographic peak area of all ion types, both precursor and fragments. The data presented illustrate a predictable relationship between the intensity of a precursor ion and the expected total product ion intensity at both the peptide and protein level, which can be used to rank or score tentative peptide and/or protein identifications by how well these relationships conform to the established models.

The number of consecutive product ions is a property that is also typically used to score peptide assignments [64]. The expected number of consecutive identified *b* and *y* ions should ideally increase, with both peptide chain length and precursor and fragment intensity. Other qualitative database search engines have implemented the use of consecutive ion series identification as a spectrum quality validation tool [21, 64, 65]. The number of identified consecutive ions, as a function of peptide chain length and intensity, is not widely utilized by these search algorithms, whereas the results shown in Fig. 3 clearly show that there is a trend, which can be modeled, between the number of matching consecutive *b* and *y* ions and the intensity and length of the originating peptide.

In the case of lower resolution mass spectrometers, MS/MS data are typically acquired at nominal mass, with the consequence that product ions can easily be assigned to an incorrect amino acid sequence. This is particularly relevant in the analysis of highly complex tryptic mixtures, where tandem mass spectra containing fragment ions of multiple peptides are acquired throughout the course of the LC separation. Therefore, during an MS/MS experiment, the coincident presence and fragmentation of multiple precursors in the collision cell [66, 67], can give rise to a multitude of amino acid combinations with similar nominal mass. This effect can lead to higher false positive protein identification rates, based upon the incorrect assignment of fragment ions at low mass accuracy and resolution. The problem is compounded by the fact that the median peptide length from any given tryptic proteome is about 11 residues in length and that close to half of the 20 naturally occurring amino acids have a neighboring amino acid within 1 mass unit.

Tryptic peptides containing additional basic residues within the peptide backbone, in addition to the normal tryptic C-terminal lysine or arginine, tend to exist at multiple charge states. In addition, when these basic residues are at, or near, the N-terminus upon fragmentation these peptides often show a summed *b* ion intensity that can and often does exceed that of the *y* summed product ion intensity. For tryptic peptides, absent of any basic residues at or near the N-terminus the opposite is true; the summed *y* ion product intensity under low energy CID conditions is typically greater than that of the corresponding *b* ions. Hence, these additional physicochemical attributes can be used for tentative peptide ranking and scoring.

3.1.3 Preferred fragmentation sites

The majority of the detected product ions from the fragmentation spectra of tryptic peptide precursor ions from complex mixtures are below $(M + H)^+$ 1200, which is illustrated by the tryptic peptide product ion frequency distribution shown in Fig. 4a. As a result, all binary tentative amino acids bond identifications, between y_3 to y_{15} and b_3 to b_{15}, can be calculated and directly compared with the experimentally identified amino acids bond cleavages. The results from this analysis are shown in Fig. 4b. As expected, and reported by other groups [68], the XP bond was identified with the highest frequency. For example, the relative occurrence of three frequently observed bond cleavages was equal to 87, 66, and 70% for IP, LP, and AP, respectively. This observation suggests that the relative intensity distribution of the fragment ion spectra should be predictable, and that rules can be applied to the identification criteria. For example, it is un-

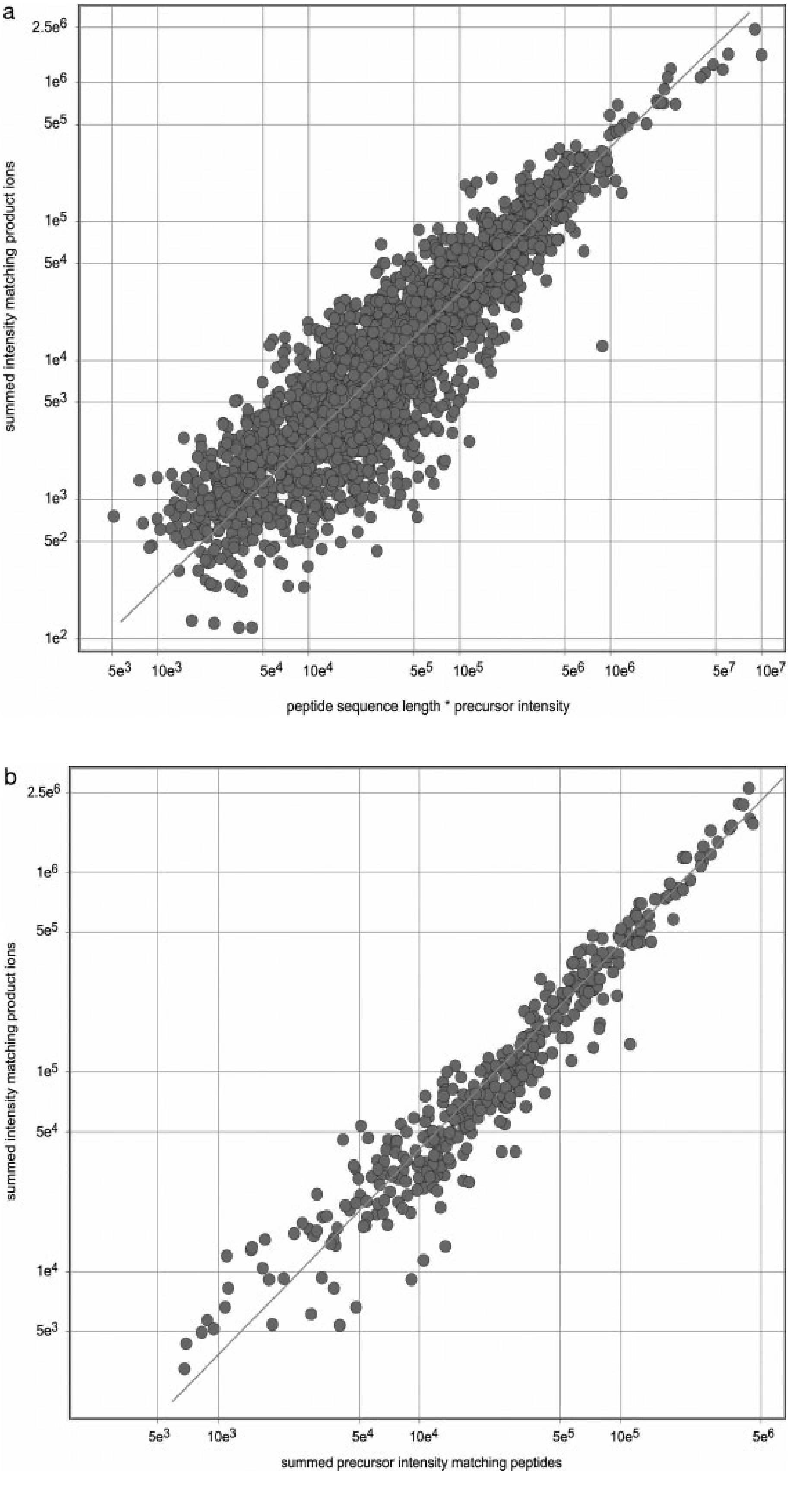

Figure 2. Summed product ion intensity of all matching fragment ions identified to a precursor as a function of the associated precursor ion intensity (a) and summed product ion intensity of all matching fragment ions identified to all precursors of a protein as a function of the summed precursor ion intensities assigned to the protein (b).

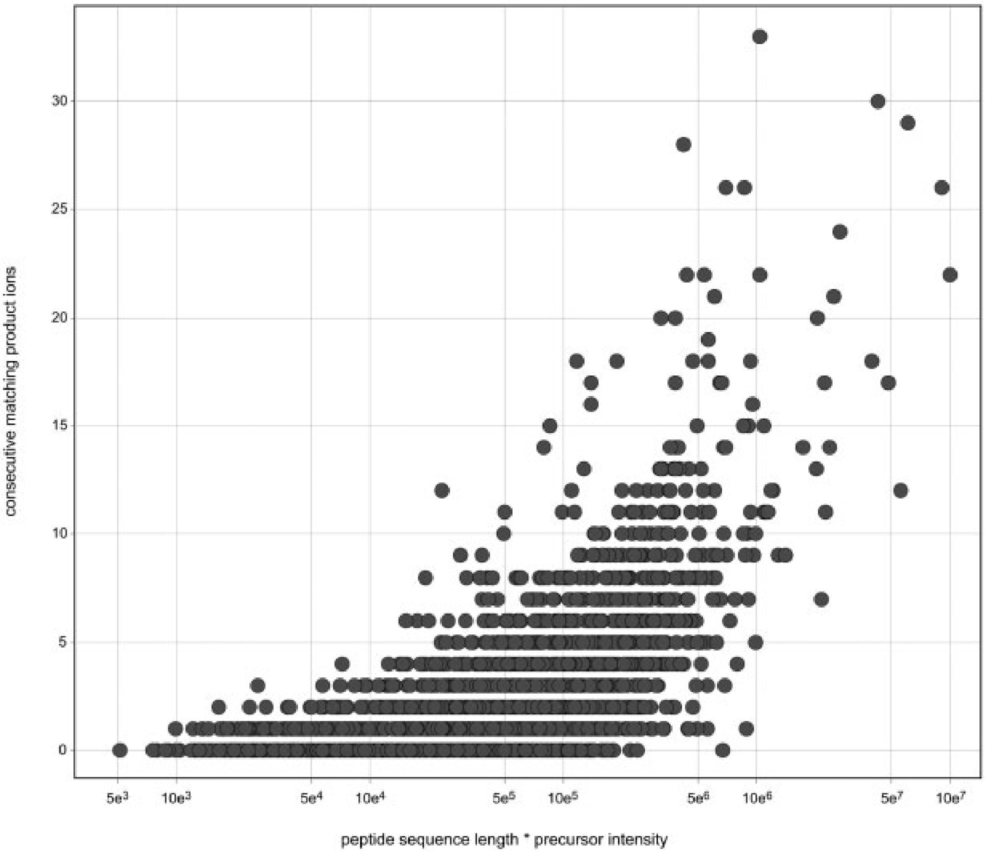

Figure 3. Number of matching consecutive *b* and *y* ions as a function of peptide chain length multiplied by the precursor intensity.

likely that an AA bond will generate a product ion of higher intensity than an IP bond. Hence, in those instances where both bonds are tentatively identified in the same fragmentation spectrum, the fragment ion originating from the IP bond should have at least the same/or higher intensity than the fragment ion originating from the amino acid bond cleavage. In other words, tentative peptide identification can either be accepted or rejected based on this rule set. The complete frequency distribution as a function of bond type cleavage is provided in Supporting Information Table S1.

The empirical derived preferred fragmentation model described here uses fragment ion type and molecular weight. A recent report [69] shows comparable results obtained by linear regression analysis of MS/MS fragmentation spectra. It was concluded that the most abundant *y* ion, preferred fragmentation sites excluded, is preferentially found at *m*/*z* approximately 60% of the precursor peptide mass, and the most abundant *b* ion typically at *m*/*z* 15–20% of the precursor mass.

3.1.4 Retention time modeling

The search algorithm employs a retention time model to predict the elution time of the peptides [70]. Retention depends on gradient slope, column length, stationary phase, and other typical RP peptide separation parameters. Hence, refinement of the model is required if any of these parameters are changed. This is addressed by the search algorithm on an injection-by-injection basis, thereby using the preassessment data set, comprising assigned sequence and experimental retention time, as input for the retention time prediction model. Supporting Information Fig. S3 shows the theoretical peptide retention time as a function of the observed retention time for a complex tryptic digest mixture. A relatively broad, but consistent correlation between the observed and predicted retention can be observed, which implies that retention time modeling can be utilized as a liquid phase physicochemical property for tentative ranking purposes and to remove gross outlier peptide identifications.

3.1.5 Conservation of mass/charge (matter) and LOD

The intensity of a precursor ion is expected to be related to the total intensity of the generated product ions. The summed intensity of the product ions should equal the intensity of the precursor if there are no subsequent losses in the collisional process, including detection by the second mass analyzer, and no neutralization of charge occurs during fragmentation. In

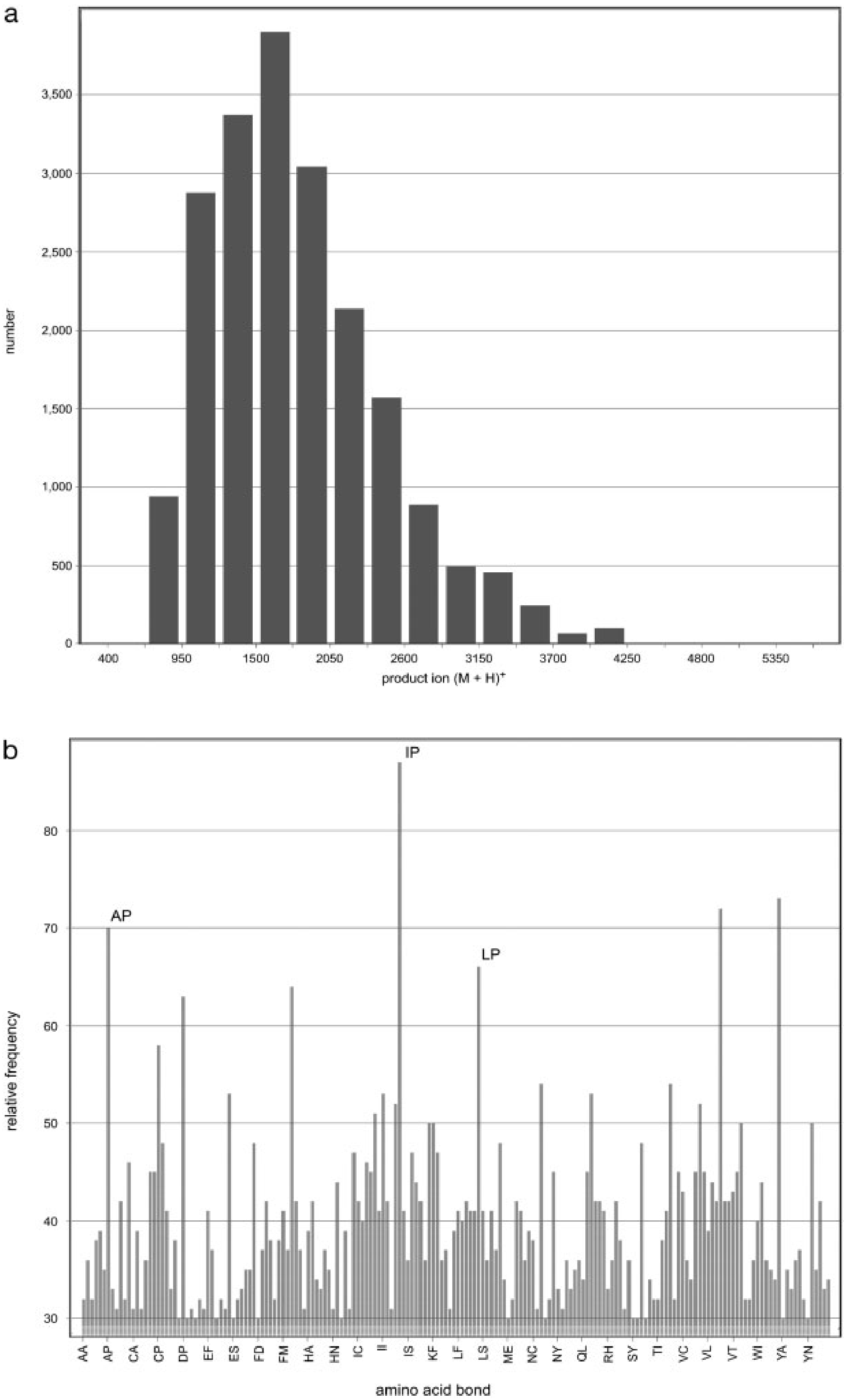

Figure 4. Precursor mass (deisotoped and charge-state reduced) frequency distribution (a) and relative fragmentation side frequency distribution (b) of the detected and identified peptides.

practice, this is not the case with collision-induced fragmentation of peptides [71]. The results shown in Fig. 5 suggest a near linear relationship between the total intensity of the fragment ions and the intensity of the precursor ion. Further examination of the total intensity of the product ions reveals that for the most intense peptides the total product ion intensity easily breaches 50% of the precursor ion intensity. Therefore, peptide identifications with a total product ion intensity exceeding that of the precursor will flag these identifications as a possible error, except in the instances where the precursor intensity was reduced due to in-source fragmentation.

As stated in the previous sections, none of the discussed physicochemical attributes are used as a hard filter to either confirm or reject a protein identification on their own. However, in combination they are extremely selective regarding the assignment of fragment ions and can be employed for the scoring of peptide and protein identifications [72]. To summarize, the algorithm utilizes the following parameters for the identification of both peptides and proteins: (i) accurate precursor mass, (ii) accurate product ion masses, (iii) total product ion intensity, (iv) number of consecutive *y* and *b* ions, (v) complementary *y* and *b* ions, (vi) experimental fragmentation at preferred cleavage sites, (vii) total *y* ion intensity/total *b* ion intensity, (viii) conformance with retention time model, (ix) total product ion intensity/precursor ion intensity, (x) neutral losses conforming to amino acid composition, (xi) multiplicity of charge states conforming to model, (xii) number of matched peptides conforming to the model, (xiii) number matched product ions conforming to model, and (xiv) total product ion intensity/total precursor ion intensity. To the knowledge of the authors, these parameters have rarely been used to support peptide/protein identifications, and could be of use in data dependent methods as well as the multiplexed data independent acquisition method presented here.

3.2 Four-protein mixture: Proof of concept example

The ability of the algorithm to provide a highly selective and sensitive method for protein identification was first assessed by thoroughly evaluating the search results for the tryptic digest of a four-protein mixture. This protein mixture has been well characterized and enables generation of unambiguous peak-list files [17]. The latter is possible because near baseline separation of the components present in the mixture is possible. In addition, this sample permitted the use of traditional MS/MS or fragment ion database search algorithms on the multiplexed data to allow the search results to be compared and contrasted. For more complex mixtures, data de-

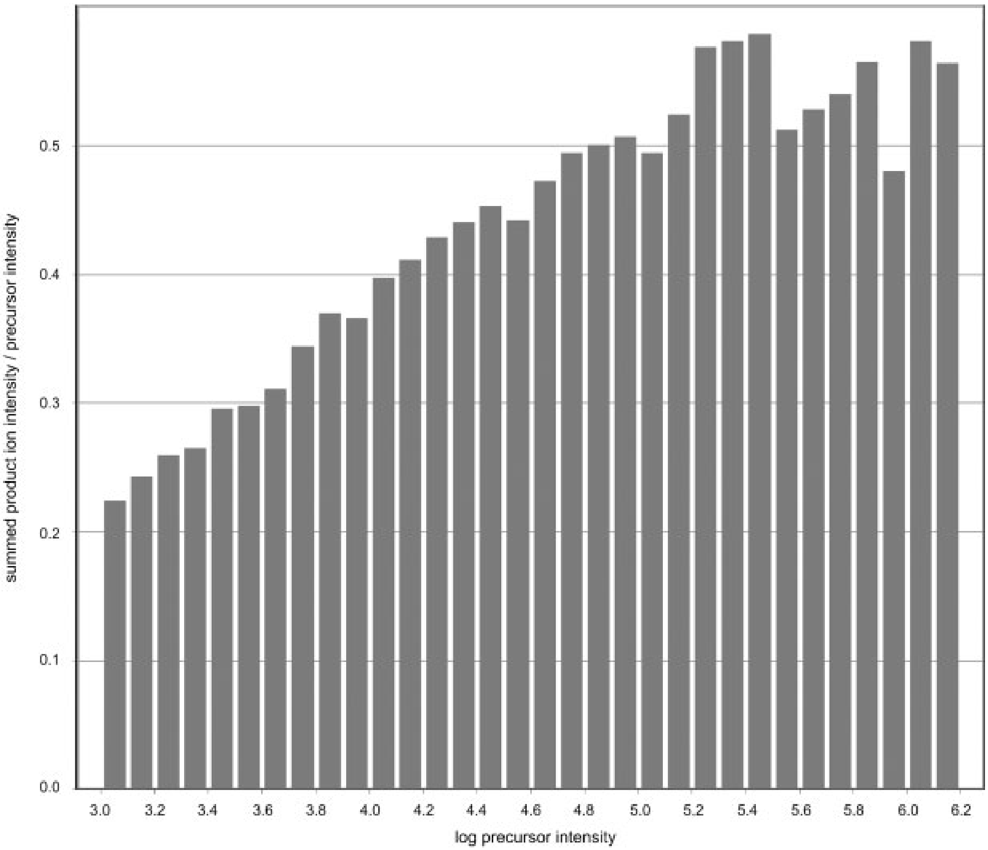

Figure 5. Ratio summed product ion intensity/precursor intensity as a function of the log value of the precursor intensity.

pendent search algorithms, especially probability based ones, can penalize the search results obtained from multiplexed fragmentation spectra because of the presence of unrelated or contaminating products ions from coeluting peptides.

Table 2 summarizes the search results for peptide and protein identifications from the algorithm, as well as the search results provided by other commercially and publicly available MS/MS database search programs. Interestingly, however, the crosscorrelation search algorithm appeared to be less affected by the presence of unrelated or contaminating products ions in the multiplexed fragmentation spectra. As noted above, these database search programs are not designed to search multiplexed fragmentation spectra. However, given a relatively simple mixture, the results should permit a valid comparison of the search results obtained. Furthermore,

Table 2. Evaluation of the ion accounting search algorithm with a standard four-protein mixture

	No. of identified unique peptides (coverage (%))			
	Glycogen phosphorylase	Serum albumin	Alcohol dehydrogenase	Enolase
Ion Accounting[a)]				
Inj. 1	68 (68)	50 (55)	31 (55)	26 (50)
Inj. 2	69 (72)	52 (57)	28 (58)	27 (52)
Inj. 3	64 (64)	54 (60)	30 (56)	28 (49)
average	67 (67)	52 (58)	30 (56)	27 (50)
Mascot[b)]				
Inj. 1	29 (37)/32 (44)	15 (27)/23 (43)	7 (21)/13 (40)	6 (17)/6 (15)
Inj. 2	23 (26)/29 (38)	13 (23)/23 (41)	10 (28)/10 (35)	9 (24)/6 (18)
Inj. 3	21 (23)/25 (31)	18 (30)/24 (44)	10 (29)/12 (34)	8 (20)/11 (26)
average	24 (29)/29 (38)	15 (27)/23 (43)	9 (26)/12 (36)	8 (20)/8 (20)
Sequest[c)]				
Inj. 1	28 (37)/31 (42)	22 (44)/29 (56)	10 (30)/14 (47)	6 (20)/8 (27)
Inj. 2	29 (35)/34 (43)	22 (38)/30 (51)	10 (29)/10 (37)	9 (25)/9 (23)
Inj. 3	26 (32)/30 (39)	21 (37)/25 (50)	10 (29)/13 (37)	7 (22)/9 (25)
average	28 (35)/32 (41)	22 (40)/28 (52)	10 (29)/12 (40)	7 (22)/9 (25)
Scaffold (X! Tandem)[d)]				
Inj. 1	21 (27)/25 (38)	11 (21)/23 (46)	5 (16)/11 (40)	5 (14)/6 (23)
Inj. 2	16 (19)/27 (37)	12 (23)/20 (39)	5 (18)/9 (34)	6 (20)/6 (20)
Inj. 3	14 (14)/24 (35)	11 (19)/21 (42)	6 (16)/10 (32)	6 (17)/8 (24)
average	17 (20)/25 (37)	11 (21)/21 (42)	5 (17)/10 (35)	6 (17)/7 (22)
ProteinLynx GlobalSERVER[e)]				
Inj. 1	24 (29)/31 (44)	10 (17)/20 (34)	8 (21)/11 (41)	7 (21)/8 (22)
Inj. 2	27 (31)/27 (34)	12 (21)/20 (32)	6 (19)/10 (37)	8 (20)/8 (20)
Inj. 3	27 (30)/28 (35)	12 (19)/22 (37)	8 (22)/10 (32)	9 (21)/9 (22)
average	26 (30)/29 (37)	11 (19)/21 (34)	7 (21)/10 (37)	8 (21)/8 (21)

Search parameters: carbamidomethylation (C) fixed modification; acetylation (N-term), deamidation (NQ), and oxidation (M) variable modifications; 1 missed cleavage.

Database: Swiss-Prot v53 appended with a peptide reversed version of the database.

The reported number of identified peptides and coverage for [b)], [c)], [d)], and [e)] preceding the solidus character represent the multiplexed spectra search results; the numbers following the solidus character represent the results from the complementary data directed analysis search results.

a) v2.3, peptide mass tolerance: automatic (approx. 9–10 ppm); fragment ion tolerance: automatic (approx. 20–23 ppm), ≥3 fragment ions/peptide; ≥7 fragment ions/protein; ≥2 peptides/protein.

b) v2.2, peptide mass tolerance: 10 ppm; fragment ion tolerance: 0.05 Da, $p<0.05$; ion score cut-off: 20 homologue proteins not reported (bold red identifications only); ≥2 peptides/protein.

c) v27.12, peptide mass tolerance: 10 ppm (postdatabase search applied); fragment ion tolerance: 0.05 Da (postdatabase search applied), manual spectrum identification validation (X_{corr} and ΔC_n scores used as guidelines); ≥2 peptides/protein.

d) v01_07_00, peptide mass tolerance: 10 ppm; fragment ion tolerance: 0.05 Da, minimum protein probability: 99%; minimum peptide probability: 95%; ≥2 peptides/protein.

e) v2.3, peptide mass tolerance: 10 ppm; fragment ion tolerance: 0.05 Da, minimum protein probability: 95%; automatic fragmentation spectrum validation (≥3 consecutive fragment ions) ; ≥2 peptides/protein.

complementary DDA LC-MS/MS experiments were conducted to confirm the authenticity of the identifications. The experimental and identification details of the complementary DDA LC-MS/MS experiments details are provided elsewhere [17]. The search parameters were kept as similar as possible for all algorithms, and a reversed database was used to estimate the rate of false positive identifications [26, 38]. None of the search algorithms reported false positively identified proteins with the use of replication as an identification filter. Note that this is not an evaluation of the capabilities of the employed search algorithms for protein identification from sequence (decoy) databases using MS data.

The peptide assignments obtained from the replicate injections indicate a high degree of reproducibility of the peptides detected. The average chromatographic reproducibility of the assigned peptides was better than 1.5% RSD, indicating the usefulness of the retention time as an efficient filter for modeling in the search algorithm, and possibly also as a reference marker for the purpose of identifying the same protein in future sample analyses. The average relative retention time difference between the ion accounting identifications and those obtained by the various applied search algorithms for the DDA experiments was well within the time associated to the parent precursors chromatographic peak width at half-height, affirming correct identification. The mass accuracy obtained for the assigned peptides was within 10 ppm for the precursors and 20 ppm for the associated fragment ions. The retention time difference between precursor and assigned product ions was, in all instances, better than 0.05 min.

A summed total of 455 redundant tryptic peptides were identified to the four proteins from the triplicate analysis of the sample. Of the 455 peptides, 103 (309 total) of the peptide assignments were found in all three injections, with 48 (96 total) found in at least two out of three injections. This represents 405 out of the 455 peptide identifications replicating in two out of three injections, equating to 89% identification reproducibility. In general, the ion accounting algorithm identified more peptides and provided higher sequence coverage than the other search algorithms tested. It may be concluded that the ion accounting algorithm, which is specifically designed to provide optimum results from multiplexed fragmentation spectra, succeeded in producing better results from this type of data. It should be noted that the majority of the peptide identifications were confirmed by independent DDA analysis under identical chromatographic conditions [17].

3.3 Four-protein mixture in a complex biological background

The same four proteins were spiked into a tryptic digest of the cytosolic proteins of *E. coli* to test the ability of the peak detection and search algorithms to extract correct information regarding alignment of precursor and fragment ions in a complex data set. A comprehensive overview of the LC-MS acquisition method used to obtain the precursor and fragment ion information of the constituent peptides is provided elsewhere [17].

In summary, the majority of the peptides (>90%) from the four proteins of interest added to the mixture were identified, despite the presence of the highly complex *E. coli* background. The peptide assignments obtained from the replicate injections to the four-protein spikes and also the *E. coli* proteins, indicates a high degree of reproducibility of peptide detection. In this experiment, the chromatographic reproducibility of the assigned peptides was under 2% RSD. In addition, the relative intensities of the peptides, to each of the four proteins, were consistent with that observed when analyzed without the *E. coli* background [17]. The consistent relative ionization profiles of tryptic peptides to a protein is not only a property that is utilized by the ion accounting algorithm, but it can also aid in guiding the design and implementation of a study for a particular set of marker proteins. With a given ionization profile to a protein, it is feasible to predict which peptides and fragment ions should be identified within a sample, which can be subsequently used as criteria to validate the presence of the protein(s) of interest. The latter will be illustrated in more detail in the next paragraph, where the same protein was identified at various concentrations in different type of samples.

As with the simple protein mixture, the mass accuracy obtained for the assigned peptides were within 10 ppm for the precursors and 20 ppm for the associated fragment ions. For this experiment, the retention time difference between the precursor and assigned product ions was in all instances better than 0.05 min. A summed total of 411 redundant tryptic peptides were identified to the four proteins from the triplicate analysis. Of the 411 peptides, 92 (276 total) of the peptide assignments were found in all three injections with 41 (82 total) found in at least two out of three injections, representing 358 out of the 411 peptide identifications. In this case, there was 87% identification reproducibility. Supporting Information Fig. S4 shows the fragmentation spectrum of one of the lower-to-medium scoring peptides identified to one of the four standard proteins, yeast alcohol dehydrogenase, in the four-protein mixture (top pane) and the four-protein mixture spiked into the *E. coli* digest (bottom pane). In both instances, the algorithm correctly identified the precursor and product ions at the same chromatographic elution time, in accordance with the retention time model, and with the same ion precursor/fragment ion intensity distribution, again in agreement with the physicochemical identification rules utilized by the search algorithm. The spiked proteins of interest were either identified with very low sequence coverage or not identified at all by the other tested database search algorithms (data not shown). Additional product ion MS/MS experiments were conducted to validate the peptide identifications that were not initially identified by the other search algorithms. A comparative overview in terms of the number of peptide identifications to the four proteins in the complex *E. coli* background and the coverage compared to the simple mixture analysis is provided in Table 3. In addition to the four spiked proteins, more than 400 *E. coli* proteins were identified with a single-dimension RP gradient separation in the complex mixture analysis. More detail on these identifications is provided elsewhere [17] and the last paragraph of this section.

3.4 Physicochemical properties of peptide and proteins to aid complex mixture analysis

In an LC-MS experiment, the peptide and product ion spectra obtained from the proteolytic peptides are independent of their environment. Once digested, a protein provides, depending on its molecular weight and concentration in a mixture, a predictable set of precursor, and product ions. In addition, the observed precursor and product ions should conform to the physicochemical property predictions described in the previous sections. Hence, the intensity distribution of the observed ions should also be predictable. This was illustrated in the context of the simple four-protein mixture in the presence and absence of the *E. coli* peptides. The relative intensity distribution of the identified peptides to each of the four spiked proteins was consistent between the two samples, exhibiting a characteristic fingerprint at both the precursor and product ion level. These results suggest that these properties can be leveraged in the analyses of other biological samples.

To illustrate this in more detail, a number of different biological sources were analyzed, and the ionization behavior of the peptides and product ions to a particular protein were examined. This is demonstrated by the results shown in Figs. 6a and b. Figure 6a shows the intensity distribution of securely identified peptides of human glucose-regulating protein GR78 in three different matrices, namely serum, glioma tissue, and pituitary tissue. It can be observed that the intrasample intensity distribution is consistent, although there are a few minor reversals of intensity order from sample to sample. It has been previously shown that the summed intensity of the top three most intense precursor ions to a given protein is proportional to its molar amount and that this relationship can be used to estimate the quantity of proteins in a complex sample [42]. The search algorithm implements this approach and provides an estimate of the absolute quantity of all identified proteins in a given sample. The samples were spiked with a known concentration of yeast alcohol dehydrogenase digest, allowing determination of the absolute concentration of the protein in each sample, 60, 12, and 15 fmol, respectively. These molar amounts agree well with the intensity ratios for the identified peptides in each sample, thereby providing additional evidence that the peptides have been properly assigned to the correct protein sequence in each sample.

A similar trend is seen for the identified fragment ions, illustrated in Fig. 6b. The left pane shows the product ion spectrum of the T6 tryptic peptide and the right pane the spectrum of the T5 tryptic fragment of glucose-regulating protein GR78. In both instances, the relative intensities of the product ions are consistent and the ratios of product ion intensities track those of the precursors. Note that the time-aligned background ions are different in each elevated-energy MS fragmentation spectrum. The only ions in common are the identified *y* and *b* peptide sequence ions. This is to be expected, since the data were obtained from three different samples, prepared at different times, and acquired on

Table 3. Evaluation of the ion accounting search algorithm with a standard four-protein mixture spike into a complex biological background

Ion Accounting[a)]	No. of identified unique peptides (coverage (%))			
	Glycogen phosphorylase	Serum albumin	Alcohol dehydrogenase	Enolase
Inj. 1	67 (64)	48 (54)	26 (54)	22 (44)
Inj. 2	62 (60)	45 (51)	25 (57)	24 (48)
Inj. 3	66 (63)	47 (52)	27 (54)	23 (47)
Average	65 (62)	47 (53)	24 (52)	23 (46)
Peptide and coverage ratio				
No. of peptides ratio[b)]	0.97	0.94	0.88	0.85
Coverage ratio[c)]	0.93	0.91	0.96	0.92

Search parameters: carbamidomethylation (C) fixed modification; acetylation (N-term), deamidation (NQ), and oxidation (M) variable modifications.

Database: *E. coli* K12 species-specific data to which the sequence of the four proteins of interest was added and appended with a peptide reversed version of the database.

a) Peptide mass tolerance: automatic (approx. 9–10 ppm); fragment ion tolerance: automatic (approx. 20–23 ppm), ≥3 fragment ions/peptide; ≥ 7 fragment ions/protein; ≥ 2 peptides/protein.

b) No. of peptides identified to the protein of interest in the complex biological background/# peptides identified to the protein of interest in the four-protein mixture.

c) Coverage of the protein of interest in the complex biological background/coverage of the protein of interest in the four-protein mixture.

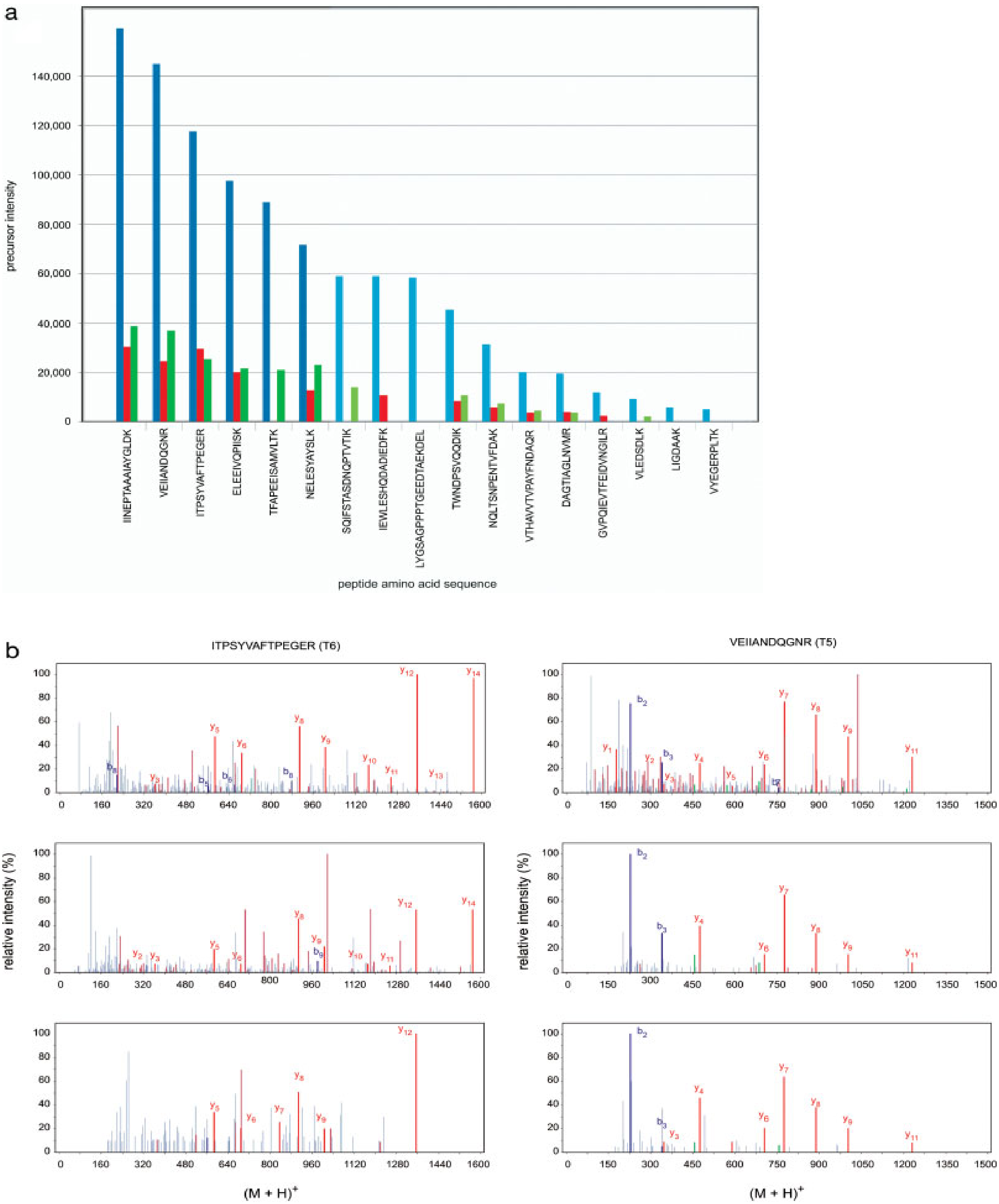

Figure 6. Peptide precursor intensity distribution (a) and fragmentation spectrum product ion intensity distribution of tryptic fragments T6 and T5 (b) of 78 kDa glucose-regulated protein for the digests of serum, glioma, and pituitary tissue, respectively. Fragment ion color legend: red, *y* ion; blue, *b* ion; green, immonium ion or neutral loss of NH_3 or H_2O; grey, not identified; magenta, fragment ion assigned to a coeluting peptide. Serum digestion was performed as described previously [73] and the digestion of glioma and pituitary tissue with minor modifications [74].

different instruments. While there can be hundreds of product ions in a single elevated-energy MS fragment ion spectrum, the algorithm clearly identified the same peptides to the same protein at similar retention times and with similar ionization distributions, at both the precursor and product ion levels. In addition, the product ions retention time apices were calculated to be within 600 ms of that of the parent precursors and their masses were within 20 ppm of those of the corresponding fragment ions in the database. The combination of time-resolved precursor and fragments and mass accuracy enhances the value of this type of data and database search results for populating bioinformatics pipelines, for the purposes of archiving the information into a database and using it to intelligently interrogate a biological sample.

3.5 Sensitivity of the search algorithm for the identification of multiplexed fragmentation data

The identification detection sensitivity of the algorithm was assessed by the analysis and quantification of a low-abundant marker protein in affinity-depleted human serum. The qualitative examples shown in the previous paragraphs were either from standard proteins alone or the same proteins spiked into a biological sample with a relatively small dynamic range compared to for instance human serum. The protein of interest in this example is a human form of chitinase, which is a marker for glucocerebrosidase deficiency that is in clinical use [75]. This protein has been studied in great detail in previous work by means of data independent LC-MS acquisitions [73] and consequently allows for the thorough evaluation of the specificity of the search algorithm in terms of time alignment of the precursor and fragment ions in the presence of a large number of other highly abundant contaminant background ions generated during multiplexed fragmentation.

Figure 7 shows the identification of the T25 tryptic fragment of chitotriosidase in its native form in post-treatment affinity-depleted human serum and that of a recombinant human form. Similar to the characteristics of the peptide and protein identification described in the previous paragraph,

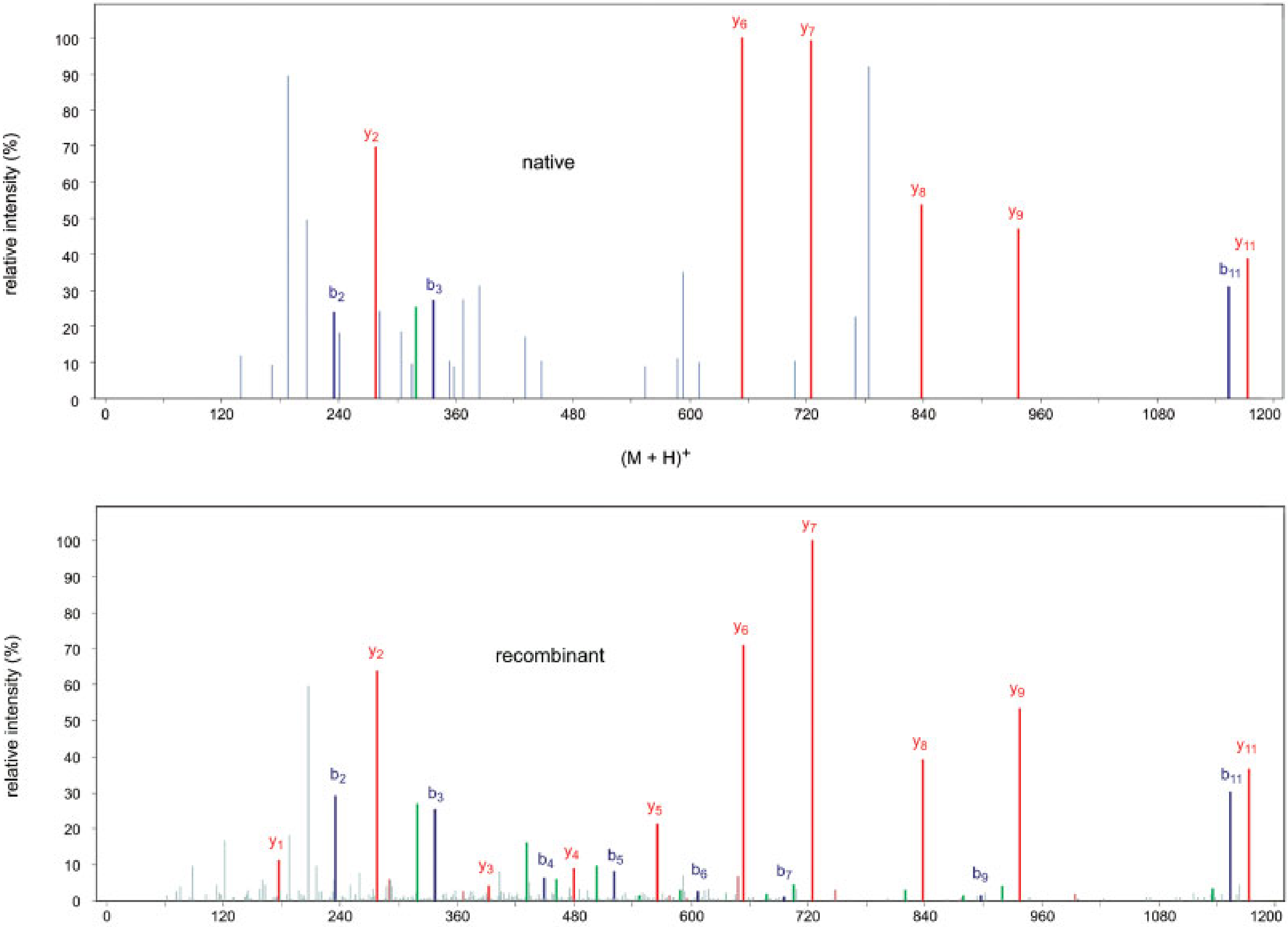

Figure 7. Product ion spectra of the T25 tryptic fragment of chitotriosidase in native depleted human pretreatment serum (top) and a recombinant human form (bottom). See reference [73] for sample preparation and analysis details. Fragment ion color legend: red, *y* ion; blue, *b* ion; green, immonium ion or neutral loss of NH_3 or H_2O; gray, not identified; magenta, fragment ion assigned to a coeluting peptide.

the only commonly identified ions between the native and the recombinant peptide are those that arise from sequence related ions. The estimated concentration of chitotriosidase using yeast alcohol enolase as the internal standard for absolute quantification was 0.99 ± 0.16 fmol/μL ($n = 3$) injected on column for this particular pretreatment sample. This measured protein marker concentration can be correlated to a serum enzyme activity of 15 600 ± 2 500 nmol/mL/h. The enzyme activity was also measured directly from undepleted serum by means of a biochemical substrate assay [75] and was found to be equal to 15 900 nmol/mL/h ± 5%. The chitotriosidase level measured with the two independent methods was found to be in the same order of magnitude.

The relative retention time difference of the T25 peptide of the native and the recombinant peptide was merely 0.06 min, which further validates the identification. The observed S/N of the identified product ions suggest that sub fmol identification is feasible with the proposed search algorithm in conjunction with data independent acquisition. This example also illustrates the correct alignment of fragment ions with the precursor. The average retention time difference between the precursor and the identified fragment ions was 0.0074 min, which is well within the time frame of a single scan. A summary of the identified proteins, peptides, and fragments ions and the relative and absolute proteins amounts of all of the identified serum proteins has been deposited in the Tranche pipeline and can be accessed with identification number 2178 (http://www.proteomecommons.org/dev/dfs/). The absolute amounts of all the identified serum proteins were also determined using yeast enolase as the internal standard and it was found that 78.6% of the detected protein mass could be accounted for.

The consistency of the fragmentation behavior of peptides generated by means of data independent acquisitions and analyzed by the ion accounting search algorithm allow for the prediction of MRM transitions for absolute quantification experiments with triple quadrupole mass spectrometers using isotopically labeled internal standards. T25 and T37 were found to be the most suitable candidate peptides of this particular protein for this purpose. N-terminal ^{13}C labeled versions of the two peptides were synthesized and used as internal standards. Both peptides, with two MRM transitions *per* peptide (T25: 586.2 → 652.3 and 586.2 → 723.3; T37: 502.2 → 630.4 and 502.2 → 652.3), were used to calculate the concentration of the protein in affinity-depleted serum, which was found to be 2.8 ± 0.2 μg/mL. This concentration measurement can also be correlated to a serum enzyme activity and equals 11 200 ± 800 nmol/mL/h.

3.6 Specificity

The specificity of the search algorithm was tested and evaluated by searching the processed multiplexed fragmentation data from the four-protein mixture added to the tryptic digest of the cytosolic proteins of *E. coli* against several microbial databases: *Bacillus subtilis*, *Mycobacterium bovis*, *Saccharomyces cerevisiae*, *Brugia malayi*, *Pseudomonas aeruginosa*, *Wolbachia* sp. and *E. coli*. The sequences of the four standard proteins and porcine trypsin were appended to each database, as well as a random version of the entire database. The data were processed and searched with default parameters and an initially allowed 4% false positive rate.

The results, and all of the associated search parameters and criteria, are summarized in Table 4. In all instances, the four internal standard proteins were identified with similar peptide complements, but only a search of the *E. coli* database returned a significant number of the 411 bacterial protein identifications. No more than 18 bacterial proteins, approximating the expected false positive rate, were returned from the searches against any other database. Several of these proteins are homologues of *E. coli* proteins, with identical tryptic peptides. As such, these identifications cannot be regarded as false, which demonstrates that ion accounting database searches are highly specific. An overview of the identified *E. coli* proteins is provided in Supporting Information Table S2. From the triplicate experiments, a summed total of 16 679 peptides were assigned to *E. coli* proteins of which 64% replicated in at least two out of the three injec-

Table 4. Identification specificity of closely related, species-specific proteomes (spiked with a four-protein digest and trypsin) in terms of the number of identified proteins and accounted on-column loading

Species specific database	No. of proteins identified	No. of internal standards proteins identified[a]	Trypsin identified	Accounted mass (%)
E. coli	411	4	Yes	91
B. subtilis	5	4	Yes	4.5
M. bovis	7	4	Yes	4.8
S. cerevisiae	1	4	Yes	4.0
B. malayi	18	4	Yes	6.1
P. aeruginosa	4	4	Yes	4.4
Wolbachia	6	4	Yes	4.7

a) Standard proteins constitute approximately for 4% of the total (accountable) mass.

tions. The peptides were matched to a summed total of 1347 proteins. Of the identified proteins, 329 (987 total) were identified in all three injections with 82 (164 total) found in at least two out of three injections, representing 1151 out of the 1347 identifications. In other words, the reproducibility of protein identification was 84%.

Since the applied scanning method generates an inventory of peptide detections with their mass, retention time, and corresponding intensities, an analysis can be performed to determine what fraction of the overall detected intensity has been identified. In the case of the *E. coli* proteins, 76% of the total detected intensity has been accounted for by the corresponding peptide assignments from the database search algorithm. Estimating the amount of all 411 *E. coli* using the recently published absolute quantification rules [42] accounted for 96% of the mass of the analyzed sample using yeast alcohol dehydrogenase as the internal standard for absolute quantification.

Receiver operating characteristic (ROC) plots were generated on the three replicate injections of the four standard proteins spiked into the digested *E. coli* cytosolic lysate to further illustrate the specificity of the search algorithm. ROC plots are graphical illustrations of the true positive rate *versus* the false positive rate. A primary purpose of an ROC plot is to depict a binary systems level of sensitivity and specificity. Figure 8 illustrates a signal plot with each injection plotted. The *y*-axis illustrates the true positive rate whereby the *x*-axis represents that of the false positive rate as a percentage of the total number of identifications. A false positive represents a random protein identification. As can be seen from the results presented in Fig. 8, the algorithm does not produce a false positive identification until approximately 80% of the proteins have been identified. In the case of the spiked *E. coli* samples, this correlates to approximately the 330th protein out of a total of 411 identifications. The 330th protein is marked by a red dot in Supporting Information Fig. S5.

Supporting Information Fig. S5 illustrates the dynamic range of the identified proteins. The *y*-axis represents the on column protein concentration for each of the 411 proteins. The *x*-axis depicts the identified proteins ranked by decreasing on column concentration. Moreover, the first 330 identified proteins account for 436 ng of the total reported 455 ng (approximately 96%). Hence, false positive identifications only occur in the lowest order of magnitude of detection. The concentrations of the identified proteins in the lower order of detection are typically in the low fmol/subfmol range. Identifications at low concentration are challenged by the fact that the number of peptides that can be identified to a protein is directly proportional to size and abundance. Moreover, the average protein molecular weight for the first 330 protein identification was found to be around 40 kDa and of the last 81 proteins around 42 kDa. This emphasizes that the intact protein molecular weight distribution of the lowest order of magnitude detected was not significantly different to that of the first two orders. This is anticipated since the identified proteins should, to a large extent, follow the intact protein molecular weight distribution of the examined proteome.

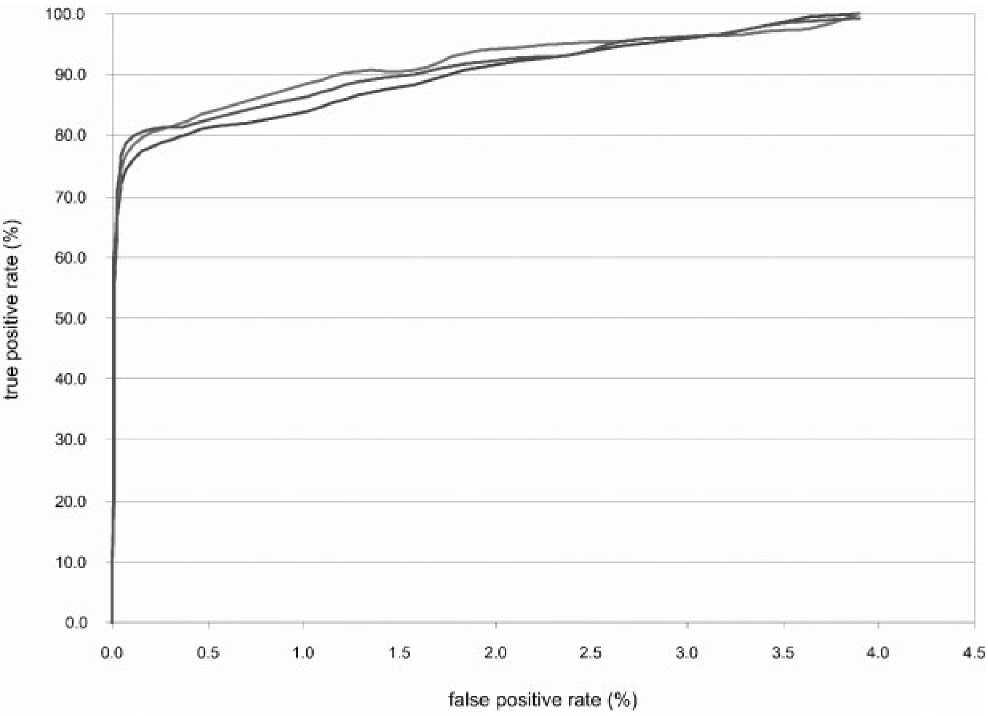

Figure 8. ROC curves for three replicate injections of a four-protein mixture spiked into the cytosolic content of *E. coli.*

It has also been shown that the number of higher ionizing peptides is directly proportional to the intact molecular weight of a protein. The number of detectable peptides at the detection limits is however limited not only by the number of peptides but also by the ability of the system to reproducibly detect the ions. The validity of the lower concentration level protein identifications was confirmed by replication (reproducibility). Each of the reported 411 proteins was identified in a minimum of 2 out of 3 replicate injections of the same sample. Moreover, the data presented in section 3.4 illustrates that for a protein that was identified at three different concentrations and in three different sample types that the majority of the peptides identified at the lower concentration regions were the better ionizing peptides for the most concentrated sample. In addition, the identified product ions as well as the fragmentation patterns were very reproducible, although the background ions were significantly different between the different sample matrices.

4 Concluding remarks

A search algorithm has been described for the analysis of data independent, multiplexed fragmentation data, to offer more comprehensive characterization of complex protein samples than that of traditional MS/MS-based algorithms. Novel aspects of the algorithm for identification and validation have been presented utilizing the physicochemical properties of peptides and proteins in both the liquid and gas phase. A number of these properties have been described in detail and indicated how they can be utilized in search strategies to improve the specificity and accuracy of peptide identifications. The combination of these properties provides a very selective mechanism to assign product ions from multiplexed fragmentation spectra to precursor ions. The presented iterative search process, utilizing a subset database search strategy and a number of precursor and product ion depletion iterations, further increases the sensitivity of the overall approach, without compromising specificity. This approach also holds significant promise as a facile means for identifying any number of different chemical and PTMs to peptides emanating from those securely identified proteins.

The search algorithm was successfully tested and validated by a comprehensive comparative analysis of the search results of well-characterized samples with a selection of alternative current search algorithms. The results from this comparison highlight the advantages of the search algorithm in conjunction with a data independent acquired LC-MS^E strategy over that of other MS/MS-based search algorithms combined with serially acquired LC-MS/MS data. The enhanced duty cycle is apparent from the overall increase in protein and peptide coverage across the entire protein molecular weight range. Of those peptides that were found in common to the other search algorithms, the same fragment ions were accounted for in both sets of results. The results described in this work, as well as the accompanying manuscript that details on the detection, correlation, and comparison of peptide precursor and products ions from data independent LC-MS with data dependent LC-MS/MS experiments [17], indicate that the additional peptides which are identified from the alternate scanning LC-MS^E analysis span a wider dynamic range than those obtained from an LC-MS/MS analysis of the same sample.

The qualitative search results have been presented, for both a complex samples and a (surrogate) clinical marker experiment, illustrate that the multiplexed fragmentation spectra, and their subsequent identification, are consistent and as such validate the qualitative identification capabilities of the presented search algorithm. The estimation of absolute quantification, from the search algorithm, was used to assess the concentration of certain proteins of interest. In the instance of the complex sample analysis, the intensity distribution of the identified peptides was used to estimate the amount and concentration of all proteins identified and a number of different proteins across different investigated sample matrices. The quantitative results of the clinical experiment were validated by two independent quantitative analytical techniques, which both confirmed the bioinformatically determined amount of the identified marker. The specificity and selectivity of the algorithm has been demonstrated by searching the data from a specific organism against species-specific databases closely related to the organism of interest; only the use of an appropriate species-specific database lead to correct identifications, proving selectivity. Lastly, the presented database search results also highlights the reproducible nature of the algorithm, which is afforded by the use of a multiplexed accurate mass data acquisition technique that presents very reproducible and consistent precursor and product ion maps, combined with a subtractive search algorithm designed specifically to deal with this data at high levels of sensitivity and specificity. This allows for accounting of the data, both in terms of observed intensity and mass.

Barry Dyson, Catalin Doneanu, James Langridge, Martha Stapels, Craig Dorschel, Joanne Connolly, and Timothy Riley are kindly acknowledged for their valuable contributions throughout the development of this work.

The authors have declared no conflict of interest.

5 References

[1] Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C. *et al.*, Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 1993, *90*, 5011–5015.

[2] James, P., Quadroni, M., Carafoli, E., Gonnet, G., Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 1993, *195*, 58–64.

[3] Mann, M., Hojrup, P., Roepstorff, P., Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 1993, *22*, 338–345.

[4] Pappin, D. J. C., Hojrup, P., Bleasby, A. J., Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 1993, *3*, 327–332.

[5] Yates, J. R. III, Speicher, S., Griffin, P. R., Hunkapiller, T., Peptide mass maps: A highly informative approach to protein identification. *Anal. Biochem.* 1993, *214*, 397–408.

[6] Strupat, K., Karas, M., Hillenkamp, F., Eckerskorn, C., Lottspeich, F., Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide Gel Electrophoresis. *Anal. Chem.* 1994, *66*, 464–470.

[7] Mortz, E., O'Connor, P. B., Roepstorff, P., Kelleher, N. L. *et al.*, Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA* 1996, *93*, 8264–8267.

[8] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 1994, *66*, 4390–4399.

[9] Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F. *et al.*, Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* 1996, *93*, 14440–14445.

[10] Washburn, M. P., Wolters, D., Yates, J. R., III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.

[11] Delahunty, C., Yates, J. R., III, Protein identification using 2D-LC-MS/MS. *Mass Spectrom. Proteomics* 2005, *35*, 209–314.

[12] Biemann, K., Papayannopoulos, I. A., Amino acid sequencing of proteins. *Acc. Chem. Res.* 1994, *27*, 370–378.

[13] Aebersold, R., Goodlett, D. R., Mass spectrometry in proteomics. *Chem. Rev.* 2001, *101*, 269–295.

[14] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198–207.

[15] Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J. *et al.*, The utility of ETD mass spectrometry in proteomic analysis. *Biochim. Biophys. Acta* 2006, *1764*, 1811–1822.

[16] Bakhtiar, R., Guan, Z., Electron capture dissociation mass spectrometry in characterization of peptides and proteins. *Biotechnol. Lett.* 2006, *28*, 1047–1059.

[17] Geromanos, S. J., Vissers, J. P. C., Silva, J. C., Dorschel, C. A. *et al.*, The detection, correlation and comparison of peptide precursor and products ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics*, *2009*, *9*, 1683–1695.

[18] Eng, J. K., McCormack, A. L., Yates, J. R., III, An approach to correlate tandem mass spectral data of peptides with amino acid Sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[19] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

[20] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[21] Skilling, J., Denny, R., Richardson, K., Young, P. *et al.*, ProbSeq – A fragmentation model for interpretation of electrospray tandem mass spectrometry data. *Comp. Funct. Genom.* 2004, *5*, 61–68.

[22] MacCoss, M. J., Wu, C. C., Yates, J. R., III, Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* 2002, *74*, 5593–5599.

[23] Wu, C. C., Yates, J. R., III, The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* 2003, *21*, 262–267.

[24] Beynon, R. J., Bond, J. S., *Proteolytic Enzymes: A Practical Approach*, 2nd Edn., Oxford University Press, Oxford, UK 2001, pp. 149–183.

[25] Olsen, J. V, Ong, S.-E, Mann, M., Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* 2004, *3*, 608–614.

[26] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[27] Tang, W. H., Shilov, I. V., Seymour, S. L., Nonlinear fitting method for determining local false discovery rates from decoy database searches. *J. Proteome Res.* 2008, *7*, 3661–3667.

[28] Fenyo, D., Ossipova, E., Eriksson, J., On the peptide fragment mass information required to identify peptides. HUPO 7th Annual World Congress, Amsterdam 2008, poster P-TUE-147.

[29] Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S. *et al.*, Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* 2002, *6*, 207–212.

[30] Sadygov, R. G., Eng, J., Durr, E., Saraf, A. *et al.*, Code development to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome Res.* 2002, *1*, 211–215.

[31] Moore, R. E., Young, M. K., Lee, T. D., Method for screening peptide fragment ion mass spectra prior to database searching. *J. Am. Soc. Mass Spectrom.* 2000, *11*, 422–426.

[32] Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., Yates, J. R., III, Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. *Anal. Chem.* 2003, *75*, 2470–2477.

[33] Beer, I., Barnea, E., Ziv, T., Admon, A., Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004, *4*, 950–960.

[34] Colinge, J., Masselot, A., Giron, M., Dessingy, T., Magnin, J., OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* 2003, *3*, 1454–1463.

[35] Liebler, D. C., Hansen, B. T., Davey, S. W., Tiscareno, L., Mason, D. E., Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal. Chem.* 2002, *74*, 203–210.

[36] Kristensen, D. B., Brond, J. C., Nielsen, P. A., Andersen, J. R. *et al.*, Experimental peptide identification repository EPIR): An integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* 2004, *3*, 1023–1038.

[37] Krokhin, O. V., Craig, R., Spicer, V., Ens, W. *et al.*, An improved model for prediction of retention time of tryptic peptides in ion pair reversed-phase HPLC: Its applications to protein peptide mapping by off-Line HPLC-MALDI MS. *Mol. Cell. Proteomics* 2004, *3*, 908–919.

[38] Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* 2005, *2*, 667–675.

[39] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D. *et al.*, Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* 2004, *76*, 3556–3568.

[40] Miyamoto., M., Yoshida, Y., Taguchi, I., Nagasaka, Y. *et al.*, In-depth proteomic profiling of the normal human kidney glomerulus using two-dimensional protein prefractionation in combination with liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* 2007, *6*, 3680–3690.

[41] Faca, V., Coram, M., Phanstiel, D., Glukhova, V. *et al.*, Quantitative analysis of acrylamide labeled serum proteins by LC-MS/MS. *J. Proteome Res.* 2006, *5*, 2009–2018.

[42] Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C., Geromanos, S. J., Absolute quantification of proteins by LCMS^E; A virtue of parallel MS acquisition. *Mol. Cell. Proteomics* 2006, *5*, 144–156.

[43] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* 2005, *4*, 1265–1271.

[44] Bateman, R. H., Carruthers, R., Hoyes, J. B., Jones, C. *et al.*, A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer for studying protein phosphorylation. *J. Am. Soc. Mass Spectrom.* 2002, *13*, 792–803.

[45] Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M. *et al.*, Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* 2005, *77*, 2187–2200.

[46] Nielsen, M. L., Bennett, K. L., Larsen, B., Moniatte, M., Mann, M., Peptide end sequencing by orthogonal MALDI tandem mass spectrometry. *J. Proteome Res.* 2002, *1*, 63–71.

[47] Jensen, O. N., Podtelejnikov, A. V., Mann, M., Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* 1997, *69*, 4741–4750.

[48] Petritis, K., Kangas, L. J., Yan, B., Strittmatter, E. F. *et al.*, Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* 2006, *78*, 5026–5039.

[49] Mant, C. T., Burke, T. W. L., Black, J. A., Hodges, R. S., Effect of peptide chain length on peptide retention behaviour in reversed-phase chromatography. *J. Chromatogr. A* 1988, *458*, 193–205.

[50] Kaliszan, R., Baczek, T., Cimochowska, A., Juszczyk, P. *et al.*, Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *Proteomics* 2005, *5*, 409–415.

[51] Meek, J. L., Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. USA* 1980, *77*, 1632–1636.

[52] Palmblad, M., Ramström, M., Markides, K. E., Håkansson, P., Bergquist, J., Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* 2002, *74*, 5826–5830.

[53] Biemann, K., Martin, S. A., Mass spectrometric determination of the amino acid sequence of peptides and proteins. *Mass Spectrom. Rev.* 1987, *6*, 1–76.

[54] McCormack, A. L., Somogyi, A., Dongre, A. R., Wysocki, V. H., Fragmentation of protonated peptides: Surface-induced dissociation in conjunction with a quantum mechanical approach. *Anal. Chem.* 1993, *65*, 2859–2872.

[55] Wysocki, V. H., Tsaprailis, G., Smith, L. L., Breci, L. A., Mobile and localized protons: A framework for understanding peptide dissociation. *J. Mass Spectrom.* 2000, *35*, 1399–1406.

[56] Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* 2004, *76*, 3908–3922.

[57] Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., Gygi, S. P., Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* 2004, *22*, 214–219.

[58] Zhang, Z., Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.* 2005, *77*, 6364–6373.

[59] Rappsilber, J., Ryder, U., Lamond, A. I., Mann, M., Large-scale proteomic analysis of the human spliceosome. *Genome Res.* 2002, *12*, 1231–1245.

[60] Liu, H., Sadygov, R. G., Yates, J. R., III, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[61] Colinge, J., Chiappe, D., Lagache, S., Moniatte, M., Bougueleret, L., Differential proteomics via probabilistic peptide identification scores. *Anal. Chem.* 2005, *77*, 596–606.

[62] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. *et al.*, Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[63] Li, X.-J., Yi, E. C., Kemp, C. J., Zhang, H., Aebersold, R., A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics* 2005, *4*, 1328–1340.

[64] Kristensen, D. B., Brønd, J. C., Nielsen, P. A., Andersen, J. R. *et al.*, Experimental peptide identification repository (EPIR): An integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* 2004, *3*, 1023–1038.

[65] Craig, R., Beavis, R. C., TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466–1467.

[66] Hoopmann, M. R., Finney, G. L., MacCoss, M. J., High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* 2007, *79*, 5620–5632.

[67] Luethy, R., Kessner, D. E., Katz, J. E., MacLean, B. *et al.*, Precursor-ion mass re-estimation improves peptide identification on hybrid instruments, *J. Proteome Res.* 2008, *7*, 4031–4039.

[68] Kapp, E. A., Schütz, F., Reid, G. E., Eddes, J. S. *et al.*, Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* 2003, *75*, 6251–6264.

[69] Barton, S. J., Richardson, S., Perkins, D. N., Bellahn, I. *et al.*, Using statistical models to identify factors that have a role in

defining the abundance of ions produced by tandem MS. *Anal. Chem.* 2007, *79*, 5601–5607.

[70] Gilar, M., Jaworski, A., Olivova, P., Gebler, J. C., Peptide retention prediction applied to proteomic data analysis. *Rapid Commun. Mass Spectrom.* 2007, *21*, 813–2821.

[71] Papayannopoulos, I. A., The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom. Rev.* 1995, *14*, 49–73.

[72] Li, G.-Z., Golick, D., Gorenstein, M. V., Vissers, J. P. C. *et al.*, A novel "ion accounting" algorithm for protein database searches. HUPO 5th Annual World Congress, Long Beach, CA 2006, Abstract 658.

[73] Vissers, J. P. C., Langridge, J. I., Aerts, J. M. F. G., Analysis and quantification of diagnostic serum markers and protein signatures for Gaucher disease. *Mol. Cell. Proteomics* 2007, *6*, 755–766.

[74] Vissers, J. P. C., Pons, S., Hulin, A., Tissier, R. *et al.*, The use of proteome similarity for the qualitative and quantitative profiling of reperfused myocardium. *J. Chromatogr. B* (in press), doi: 10.1016/j.jchromb.2008.10.024).

[75] Hollak, C. E. M., van Weely, S., van Oers, M. H. J., Aerts, J. M. F. G., Marked elevation of plasma chitotriosidase activity: A novel hallmark of Gaucher disease. *J. Clin. Invest.* 1994, *93*, 1288–1292.