



Cite this: DOI: 10.1039/c8mo00077h

Robust determination of differential abundance in shotgun proteomics using nonparametric statistics†

Patrick Slama,^{‡,ab} Michael R. Hoopmann,^{‡,c} Robert L. Moritz,^{‡,c} and Donald Geman^{*ad}

Label-free shotgun mass spectrometry enables the detection of significant changes in protein abundance between different conditions. Due to often limited cohort sizes or replication, large ratios of potential protein markers to number of samples, as well as multiple null measurements pose important technical challenges to conventional parametric models. From a statistical perspective, a scenario similar to that of unlabeled proteomics is encountered in genomics when looking for differentially expressed genes. Still, the difficulty of detecting a large fraction of the true positives without a high false discovery rate is arguably greater in proteomics due to even smaller sample sizes and peptide-to-peptide variability in detectability. These constraints argue for nonparametric (or distribution-free) tests on normalized peptide values, thus minimizing the number of free parameters, as well as for measuring significance with permutation testing. We propose such a procedure with a class-based statistic, no parametric assumptions, and no parameters to select other than a nominal false discovery rate. Our method was tested on a new dataset which is available *via* ProteomeXchange with identifier PXD006447. The dataset was prepared using a standard proteolytic digest of a human protein mixture at 1.5-fold to 3-fold protein concentration changes and diluted into a constant background of yeast proteins. We demonstrate its superiority relative to other approaches in terms of the realized sensitivity and realized false discovery rates determined by ground truth, and recommend it for detecting differentially abundant proteins from MS data.

Received 28th March 2018,
Accepted 10th September 2018

DOI: 10.1039/c8mo00077h

rsc.li/molomics

Introduction

A primary objective of quantitative proteomics-based analysis is to identify proteins and protein modifications whose abundances vary significantly between sets of samples. In proteomics experiments employing data-dependent analysis (DDA) (*i.e.*, shotgun or bottom-up analysis), protein samples are first subjected to enzymatic proteolysis. The resulting peptides are then fractionated by reversed-phase high-performance liquid chromatography (RP-HPLC) and analysed by tandem mass spectrometry (MS).^{1,2} In label-free quantitation methods, protein abundance is inferred by measuring constituent peptide signal intensity^{3,4} or by spectral counting,^{5–7} which measures the frequency of peptide selection.

Such inferred values are subsequently analyzed in order to identify those proteins that best differentiate between the classes of experimental interest, using some statistical methodology. Our focus here is on the elaboration and statistical analysis of quantitative values to identify possible protein markers, and on an exploration of fold changes detectable by proteomic analysis between phenotypic states.

Multiple statistical methods have been described for the efficient determination of differentially abundant proteins from label-free DDA proteomic experiments.^{6,8–10} A widespread approach is to directly construct surrogate values at the protein level from either spectral counts or peptide ion intensities, in a process coined ‘summarization’, and to compare these values between classes using a statistical procedure. Summarization-based methods do not usually account for peptide-to-peptide differences in detectability and may involve *ad hoc* parameter choices. Recent publications suggest superiority of peptide-based measures with respect to protein-based, summarized measures.^{10,11}

Most of these methods utilize either standard parametric models or variations developed in applications of high-dimensional statistics to fields outside biology. Peptide-based stochastic modeling, such as standard linear regression, is a common

^a Center for Imaging Science, Institute for Computational Medicine, Johns Hopkins University, USA. E-mail: geman@jhu.edu

^b Independent Researcher, Paris, France

^c Institute for Systems Biology, 401 Terry Avenue N, Seattle, WA, 98109, USA

^d Department of Applied Mathematics and Statistics, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD, 21218, USA

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8mo00077h

‡ These authors contributed equally to this work.

strategy to test for a significant class effect.¹² Still, stochastic modeling generally makes assumptions on the distribution of the observed data, such as normality, which are evidently inappropriate due to a high proportion of zero values, often as high as 50% or more, in proteomics data.^{10,12–14} Moreover, it usually requires estimating many parameters with a very small number of samples. It is therefore hardly surprising that many current methods are unable to control both the sensitivity (or recall) and the false discovery rate at the same time.^{12,15}

Statistical methods for feature selection and performance evaluation which are better suited to proteomics data have been developed in other fields, notably computational genomics. Both fields indeed have many objectives and challenges in common, as pointed out nearly ten years ago.¹⁶ For example, summarization-based protein quantitation methods employ the same type of statistical tests as those used in computational genomics to identify differentially expressed genes.¹⁷ Nonetheless, even recent work in computational proteomics rarely takes this literature into account, despite nearly twenty years of experience addressing similar goals (*e.g.*, biomarker discovery) and facing similar technical barriers (*e.g.*, sample size), with some exceptions (see ref. 16). More specifically, in both genomics and proteomics, phenotypes are identified from a small number of samples (n) with a large number of variables (d). In statistical inference and learning, this is referred to as the “small n , large d ” dilemma. As a result, model complexity must be tightly controlled in order to prevent “over-fitting” the data,¹⁸ a consequence of which is the lack of robustness of the FDR (False Discovery Rate) analysis; in such cases, the realized FDR may differ largely from the target (or nominal) FDR.¹⁵ With nonparametric methods, such as Wilcoxon and permutation-based tests, no assumptions are made about the data distribution, *e.g.* that of normality. Thus, distribution-free methods are more reliable, and hence very widespread in genomics.

With these considerations in mind, we propose to analyze shotgun proteomics data using statistical tests which have the following properties:

(i) There are no parameters to choose or estimate; in particular, there are no menus to select from (as in, *e.g.*, MaxQuant Perseus^{19,20}), or parameters to estimate from the data (*e.g.*, as in linear regression models with normal errors^{3,12}).

(ii) No surrogate value is assigned to proteins; instead, the analysis is peptide-centric.

(iii) No assumptions are made about the probability distribution of either the data or the test statistics; instead, all p -value calculations are permutation-based, meaning that all p -values are relative to a null distribution obtained by randomly permuting class labels.

The approach we propose satisfies criteria (i)–(iii). Each peptide value is first re-scaled across samples to a value comprised between 0 (min across samples) and 1 (max across samples). For each protein, we combine these normalized values over all replicates and peptides to produce a statistical descriptor for each condition. Then, given two conditions, for each protein, the two corresponding statistical descriptors are combined into one test statistic whose significance is

determined by permutation testing. Peptide values used in the initial step of the algorithm can be precursor ion intensity (PI) as well as spectral count (SC) values; however, the present analysis primarily focuses on PI, since these values enable one to better identify differential changes.²¹

We have applied our approach to finding differentially abundant proteins on a controlled dataset consisting of variable amounts of human proteins mixed into a yeast protein background. Our novel, enlarged dataset resembles one constructed in a CPTAC (Clinical Proteomic Tumor Analysis Consortium, cptac-data-portal.georgetown.edu) study,²² with efforts made to overcome some limitations of that dataset. Indeed, recent efforts have sought to extend the utility of the original CPTAC collection by expanding the quantitation standard using more dilutions levels and more modern instrumentation.⁹ Anticipating the permutation testing procedure and corresponding effect of sample size on statistical resolution, we collected data for twelve repetitive injections for each concentration, as opposed to only three or four in previous studies.^{9,23} In this manner, more meaningful comparisons can be made than in previous methods without relying on severely undersized “ground truth” datasets.^{9,10} Our method does obtain a better tradeoff between sensitivity and FDR than previously published, parametric methods; for instance, we reach higher sensitivity levels at FDR values equivalent to other studies. This method should thus prove useful for unbiased, robust proteomics-based discovery of differential protein abundance.

Experimental

Sample preparation

Yeast strain EDC3 (a gift from Prof. JD Aitchinson, ISB, Seattle) was grown to mid-log phase and harvested by centrifugation. The cells were lysed by flash freezing in liquid nitrogen prior to disruption using a Retsch ball mill grinder and resuspended in 100 mM ammonium bicarbonate buffer, as previously described.²⁴ Protein concentration was determined by micro BCA assay (Thermo-Fisher Scientific, San Jose, CA). A yeast protein stock of 100 ng μL^{-1} in 100 mM ammonium bicarbonate was reduced in 5 mM DTT for 30 minutes at 60 °C, alkylated with 7.5 mM iodoacetamide (IAM) for 30 minutes at room temperature in darkness, and digested with trypsin (Promega, Madison, WI) at a 1:200 ratio for 4 hours at 37 °C. Digestion was stopped by heat inactivation, samples were then acidified with 0.1% formic acid and stored at –80 °C. The universal proteomics standard set (UPS1, Sigma-Aldrich, St-Louis, MO) was prepared by resuspending one 6 μg vial with 100 μL of 50% TFE in 100 mM ammonium bicarbonate. The proteins were reduced with 5 mM DTT for 30 minutes at 60 °C and alkylated with 7.5 mM IAM for 30 minutes at room temperature and darkness. The TFE was diluted by addition of 500 μL of 100 mM ammonium bicarbonate, and the proteins were digested with trypsin (Promega) at a 1:100 ratio for 4 hours at 37 °C. Digestion was stopped by heat inactivation. The peptides were dried down and reconstituted in 100 μL of 0.1% formic acid solution to an approximate concentration of 50 fmol μL^{-1} .

Table 1 UPS1 dilution series. Conditions are numbered following increasing UPS1 standard concentrations

Condition	Standard conc. (fmol μL^{-1})	Yeast conc. (ng μL^{-1})
C0	0	60
C1	0.25	60
C2	0.74	60
C3	2.2	60
C4	3.3	60
C5	6.6	60
C6	20	60

A dilution series of UPS1 in yeast background was prepared by mixing varying concentrations of each solution in 100 mM ammonium bicarbonate, to approximate the dilution series of Paulovich *et al.*,²² with an additional dilution factor for increased granularity. The protein concentrations used in this study are listed in Table 1. For each spiked-in concentration, 12 identical injections were analyzed. This number was chosen as being sufficiently low to resemble clinical datasets as well as sufficiently high for providing some statistical power (datasets with three replicates are not reasonable for statistical analysis on samples containing hundreds of proteins). To minimize the batch effects experienced with chromatography column and nano-ESI detector signal loss, the samples were randomly analyzed in sets of three injections until twelve injections were collected for all UPS1 spiked-in concentrations, with blank injections performed between each set. Our acquisition was thus different from approaches that acquire all sample injections from lowest to highest concentration to minimize carryover. For example, UPS1 proteins were observed in the yeast control samples, despite the use of blank injections between sample sets.

Mass spectrometry analysis

LC-MS/MS analysis of each UPS1/yeast sample was performed using a 10.5 cm PicoChip (New Objective, USA) capillary (75 μm ID, ReproSil Pur C18 3 μm). Prior to loading onto the column, each sample was loaded onto a 2 cm Acclaim PepMap 100 trap (75 μm ID, C18 3 μm ; Thermo Fisher Scientific). For each sample injection, 2 μL of sample was loaded onto the trap using an Easy nLC-1000 system (Thermo Fisher Scientific). Mobile phase A consisted of 0.1% formic acid in Milli-Q water, and mobile phase B of 0.1% formic acid in acetonitrile. Each sample was separated using a binary mobile phase gradient to elute the peptides. The gradient program consisted of three steps at a flow rate of 0.3 $\mu\text{L min}^{-1}$: (1) a linear gradient from 2% to 40% mobile phase B over two hours, (2) a 10 minute column wash at 80% mobile phase B, and (3) column re-equilibration for 10 minutes at 2% mobile phase B. Mass spectra were acquired on a Q Exactive HF (Thermo Fisher Scientific) mass spectrometer operated by data dependent acquisition (DDA) using a top 30 selection count. The precursor ion scan range was 350–1400 m/z at 60 000 resolution. An isolation window of 1.2 m/z was used for selection, a normalized collision energy of 30 was set, and higher energy collision induced dissociation (HCD) MS/MS spectra were acquired at 15 000 resolution. An automatic gain control (AGC) target of 10^5 and maximum injection time of 50 ms was set.

Charge exclusion was set to 1 and greater than 5, with isotope exclusion. Dynamic exclusion time was set to 10 seconds.

MS/MS-data processing

Mass spectra (Thermo Fisher.raw files) were converted to mzML format using MSConvert (version 2.2.0)²⁵ and searched with Comet (version 2015.02 rev. 2).²⁶ Spectra were searched against the UniProt *S. cerevisiae* reviewed proteome (downloaded on December 4, 2015), supplemented with the 48 UPS1 protein sequences, and reversed decoy sequences (13 484 total protein sequences). Comet parameters included a fixed modification of +57.021464 Da on cysteine and a variable modification of +15.994915 Da on methionine. Precursor mass tolerance was set to 25 ppm and a fragment bin tolerance of 0.2 and fragment bin offset of 0 were used. Semi-tryptic enzymatic cleavage was set, allowing for up to 2 missed cleavages. Peptide-spectrum matches (PSM) were analyzed using the Trans-Proteomic Pipeline (TPP, version 4.8.0 PHILAE),²⁷ to assign peptide probabilities using PeptideProphet²⁸ and iProphet.²⁹ Spectral counts and precursor ion intensities were exported for each non-redundant PSM at a 1% false discovery rate (FDR). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium *via* the PRIDE partner repository³⁰ with dataset identifier PXD006447.

Experimental design and statistical rationale

The resulting data matrix containing intensity values for all the identified peptides and corresponding assigned protein names was used as an input for calculation, including all contaminant and decoy sequences. Peptides that were assigned to more than one protein were discarded from the analysis, in order not to overestimate their importance in the further data analysis with respect to the other peptides collected. Though methods have been developed that can incorporate these degenerate peptides,^{31,32} no consensus currently exists in the literature in this respect. Overall, the number of multiple assignments was low. These represented 1312 peptides over a total of 10 406 peptides (across all conditions). Then, for a comparison between two experimental conditions, all peptides which had at least one non-null value over the 2×12 samples analyzed were considered in our statistical pipeline. The present work is mainly intended for the discovery stage of a two-stage discovery-verification process, in which case the proteins selected here would be subsequently re-analyzed in a verification stage, presumably with more replicates and/or higher quality data.

The discovery of differentially abundant proteins is a two-class problem, with two experimental conditions (here, the UPS1 concentrations) and peptide-level measurements available for n_a samples from class a and n_b samples from class b. Let N_{pep} denote the number of identified peptides in the LC-MS/MS experiment and let N_{prot} be the corresponding number of proteins (identified for at least one sample) after assignment using PeptideProphet and iProphet (see above section). For each protein i , let $\text{pep}(i)$ be the set of detected peptides assigned to protein i , with $N_i = |\text{pep}(i)|$ the size of this set, for $i = 1, \dots, N_{\text{prot}}$. Recall that each peptide is uniquely assigned to a protein and consequently there is no overlap among the sets $\text{pep}(i)$.

We now describe our nonparametric test. We will write y_{jk} for the measured value for peptide j in sample k , expressed in original units (*i.e.*, not log-transformed); ‘peptide values’ later represent either PI or spectral counts. For any given comparison between two classes or conditions, the complete set of values y_{jk} is organized as a matrix where each row represents a peptide $j = 1, \dots, N_{\text{pep}}$ and each column represents a sample $k = 1, \dots, n_a + n_b$. The first n_a columns are counts for class a and the next n_b columns are the counts for class b. Note that in the current analysis, $n_a = n_b = 12$ for all comparisons performed. All calculations on the peptide values matrix were performed using MATLAB (The Mathworks, Inc., Natic, MA, USA). The MATLAB script is provided as ESI.†

Data properties

Our approach is motivated by some general properties of the peptide quantitation data. Suppose a protein i is fixed. The data available for designing an algorithm for detecting differential abundance are then two sets of measurements: the $n_a N_i$ observed peptide values $\{y_{jk}, j = 1, \dots, N_i; k = 1, \dots, n_a\}$ for class a and the $n_b N_i$ values $\{y_{jk}, j = 1, \dots, N_i; k = n_a + 1, \dots, n_a + n_b\}$ for class b. Whereas the two sets of values are statistically independent, the values within each set are generally neither independent nor identically distributed. They are not independent, even within a sample k , due to peptide-to-peptide correlations and they are not identically distributed, again even within samples, due to possible differences in both digestion and detection across peptides from a given protein. Finally, recall that if peptide j was not detected in a sample k , y_{jk} is set to 0. A notable property of these data is that there can be many zero values, especially for peptides associated with proteins at low concentrations. In particular, the assumption of normally-distributed y_{jk} values is rarely fulfilled (see Discussion).

Test statistics

In short, our algorithm performs a re-scaling of detected peptide values, then aggregates these values within each class, and evaluates the significance of the resulting difference using permutation testing (Fig. 1).

In a first step, for each peptide j , we re-scale the set of values y_{jk} , $k = 1, \dots, n_a + n_b$ (*i.e.*, one row of the matrix) to $[0,1]$; this normalizes for peptide-to-peptide variability, or the different responses observed for peptides of equal abundance originating from the same protein. Let m_j (respectively, M_j) be the minimum (resp., maximum) value observed over all samples $k = 1, \dots, n_a + n_b$ for peptide j , and define

$$y_{j,k}^{\text{norm}} = \frac{y_{jk} - m_j}{M_j - m_j} \quad (1)$$

In most of our experiments $m_j = 0$ due to the widespread presence of zero values. In that case (or if we use y_{jk}/M_j), this transformation has the property that, within peptides, the ratios of original counts from sample to sample are preserved; this is not the case when using, for example, a log-transformation. Note that the y^{norm} values are a surrogate for ranks, with ties allowed.

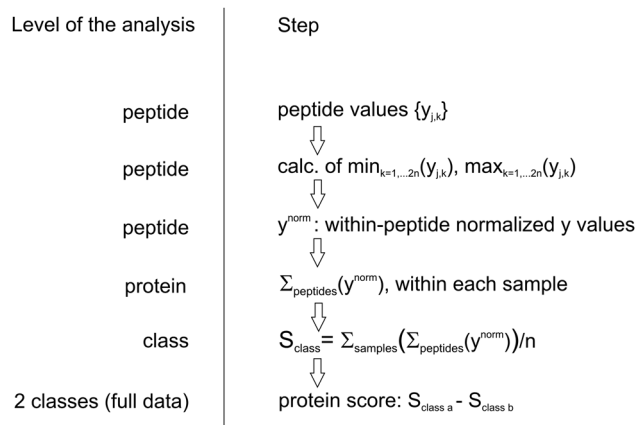


Fig. 1 Schematic description of our calculation pipeline. Note that class sizes are written as $n = n_a = n_b$ for readability.

The normalized peptide values are then averaged across peptides in order to assign an overall value to a protein within a sample, and next averaged across samples for classes a and b separately:

$$S_a = \frac{1}{n_a} \sum_{k=1}^{n_a} \frac{1}{N_i} \sum_{j=1}^{N_i} y_{jk}^{\text{norm}}, \quad S_b = \frac{1}{n_b} \sum_{k=n_a+1}^{n_a+n_b} \frac{1}{N_i} \sum_{j=1}^{N_i} y_{jk}^{\text{norm}}$$

The resulting test statistic or ‘score’ for a protein i is the difference of these two class-wise averages:

$$T_i^{\text{norm}} = |S_a - S_b| = \left| \frac{1}{n_a N_i} \sum_{k=1}^{n_a} \sum_{j=1}^{N_i} y_{jk}^{\text{norm}} - \frac{1}{n_b N_i} \sum_{k=n_a+1}^{n_a+n_b} \sum_{j=1}^{N_i} y_{jk}^{\text{norm}} \right| \quad (2)$$

We use the term ‘score’ to underline the fact that T_i^{norm} no longer represents a detected amount or other physically meaningful quantity, but a computed value. We refer to this method as CSNorm for Class-wise Sum of Normalized peptide values. Since this score can both be used on PI or SC as peptide values, we will further denote it as CSNorm-PI in the first case and CSNorm-SC in the second one. The performance of this score will be compared with that of related nonparametric scores, as well as with other approaches. It is worth noting that many alternative transformations were tried, including using ranks and more complex approaches, of which none uniformly outperformed the others. Our solution was to favor simple methods.

Significance

For each protein i we compute a p -value p_i for the test statistic T_i^{norm} (eqn (2)) by constructing a permutation-based null distribution. This avoids making any assumptions about the distribution of the test statistic under the null hypothesis that there is no difference between classes in the distribution of peptide signals. We follow the standard approach. We generate K ‘null values’ of the test statistic T_i^{norm} by randomly permuting the $n_a + n_b$ class labels (*i.e.*, selecting a permutation of $\{1, \dots, n_a + n_b\}$ at random) and computing the test statistic T_i^{norm} for the new, randomized

data. The estimated p -value (two-tailed) is the fraction of these K null values which exceed the value of T_i^{norm} on the original data. The choice of K (generally 2×10^4 to 1×10^5) determines the resolution of the p -values.

Alternative test statistics

There are several natural alternative test statistics to CSNorm that will be discussed in later sections. One is to use the plain $y_{j,k}$ values, which we denote as CSPep (-PI or -SC), for Class-wise Sum of Peptide values. Another is to replace y_{jk}^{norm} in eqn (2) by $\log(y_{jk})$; we denote the resulting score as CSLog and later use it as a reference method, as most similar to traditional fold-change based approaches.

The normalized spectral index (NSI), which was originally defined by Griffin *et al.* as a homogeneity measure,³³ was used to compare our different experimental conditions. Consistent with that publication, protein abundances were log-scaled prior to comparative analysis. An additional approach, the normalized spectral abundance factor (NSAF)⁷ was also compared to our methods, and computed using the protein quantification tool, StPeter (version 1.2.3).³² For both the NSI and NSAF analyses, protein values were computed in the two classes, and the difference of the class sums was used as a statistic, the significance of which was evaluated by permutation testing, as described above.

Our method was yet further compared to the popular MSstats v3.5.3 method,³ which was implemented using Tukey's median polish, and normalization set to "equalizeMedians". Precursor ion intensities were imported for all non-degenerate peptide sequences, missing measurements from any given sample being labeled as "NA". The MBImpute parameter was set to true.

Output and performance metrics

As an input, our algorithm requires a matrix for peptide values, each row being a peptide and each column a sample, the samples being grouped by condition or class. In addition, the algorithm requires the input of a two-column matrix: the first column corresponds to an index for the peptides present in the data matrix and the second column corresponds to an index for the protein to which each peptide was assigned.

The output of the algorithm is a list of p -values $\{p_i, i = 1, \dots, N_{\text{prot}}\}$, one for each protein, as well as the sign of the difference $S_b - S_a$, which indicates whether the corresponding protein concentration increases or decreases from one class to the other. For simplicity, the proteins are ordered by increasing p -value, so that $p_1 \leq p_2 \leq \dots \leq p_{N_{\text{prot}}}$. Detection is performed by specifying a threshold, and the resulting set of proteins are then those whose variations across the two conditions are declared significant at the chosen threshold.

In our approach we focus on two criteria for determining which proteins are selected as significantly differentially abundant.

The first criterion is to specify the number N_D of detected proteins (or discoveries). We thus propose distinct values for N_D , the smallest (30) for verification of detected candidates using for example affinity-based assays (*e.g.*, ELISA) and the

largest ones (120–150) for using, for example, selected-reaction monitoring (SRM) verification, for which it is feasible to survey much larger numbers of proteins through specific peptide surrogates.³⁴ Since the p -values are ordered, the corresponding set of proteins is simply $\{1, \dots, N_D\}$.

The second criterion, of broad use in computational biology, is based on the false discovery rate (FDR). (Note that this FDR applies to the detection of differential abundance across classes, and is hence different from that used to assign peptide identities to MS data.) A target or nominal FDR is selected, *e.g.*, 0.05, and this cut-off is transformed into a p -value threshold (and *vice versa*), for example using the Benjamini and Hochberg procedure.³⁵ In order to allow for proper comparisons among the results of different methods, it will be useful to make these terms more precise. Let TD be the number of true positives, or true detections, among the N_D detections determined by the p -value threshold. The false discovery rate for the given data is then $Q = \frac{N_D - \text{TD}}{N_D}$, which is the fraction of detected proteins which are false positives. We will refer to Q as the realized or empirical FDR. Some authors use an equivalent metric, precision, which corresponds to $1 - Q$. In statistics, the FDR itself is defined as $E(Q)$, where the expectation is over all possible datasets generated under the same conditions.

Another important performance metric is the sensitivity, or recall, of the method, namely the fraction TD/N_p , where N_p is the total number of true positives. In general cases, one does not know which protein detections are true positives and which are false positives. However, in the series of experiments presented here, these sets of proteins are known in advance: every human UPS1 protein is a true positive and every yeast protein is a false positive. This allows us to precisely quantify the performance of our method and of competing algorithms, for instance by comparing the target FDR with the realized FDR Q , which is the fraction of detected proteins which are yeast proteins. Similarly, for any given threshold, whether based on a target FDR or on the total number of detections, we can compute the (realized) sensitivity.

Although it is common to consider the realized FDR Q as an estimate of the true FDR $E(Q)$, the realized FDR may be quite different for various reasons (including high variance and violation of assumptions for the BH procedure to yield an unbiased estimate). From a practical perspective, a far more important difference is the one between the target FDR and the realized FDR. Only the latter is a valid performance metric.

Results

Data

MS/MS data were acquired in 12 separate replicate runs for each of the different concentrations of the UPS1 standard set of human proteins, which were mixed with a constant background of yeast (*S. cerevisiae*) proteins. Peptide identity and corresponding quantities were next derived from these data, as described in the Methods. Six different concentrations were used for the spiked-in protein set (see Table 1), additional runs being performed for

the sole yeast protein background (equivalent to a null concentration for the standard). Each sample replicate was acquired in batches of three injections to minimize batch effects, as described in the Methods. Our acquisition was thus different than previous approaches that acquire all sample injections from lowest to highest concentration to minimize carryover when running samples sequentially.^{9,22,23} For example, UPS1 proteins were detected in the yeast control samples run in between each batch as a result of sample carryover despite blank injections between sample sets. In the original CPTAC protocol, UPS1 concentrations followed a linear three-fold stepwise increase;²² we included an additional UPS1 concentration in the middle of the range (condition C4, see Table 1), so as to obtain a real life concentration and more difficult comparison, allowing for a finer resolution among methods, with one pair of conditions at a 1.5-fold concentration change (C3 vs. C4) and one at a 2-fold change (C4 vs. C5; see Table 1 and Methods).

There were between 2841 and 4826 peptides identified per sample, with the average number of observed peptides per condition and total number of UPS1 peptides observed in one condition described in Table S1 (ESI[†]). Peptides that were assigned to more than one protein were discarded from the analysis; these represented 1312 peptides from a total of 10 406 peptides, with *e.g.* 1151 peptides assigned to 2 proteins and 25 peptides assigned to more than five proteins. Overall, 9094 peptides were considered for analysis.

Performance of CSNorm-PI

We applied our score (CSNorm, eqn (2)) to PI data in order to detect proteins with differential abundance across samples. We also used a reference score, CSLog (see Methods), to represent direct class-based aggregation of the log-transformed data, which resembles the classical fold-change analysis. Summing peptide values within class, as in these two scores, produces a score with more resolution than fold-change based scores, since a difference between class is always defined, unlike a ratio.

The performance of our method was first evaluated on three 'hard' comparisons, C0 (no spike-in) vs. C2, C1 vs. C2 and C3 vs. C4 (1.5×-fold change; see Table 1), as well as an easier detection, C4 vs. C5 (higher concentrations). For each comparison, all identified proteins were ordered by increasing *p*-value according to the permutation-based procedure, and we considered the first N_D proteins for their UPS1 or yeast origin. When considering the 60 first discoveries ($N_D = 60$) produced by our methods, CSNorm-PI detected 31 UPS1 proteins when comparison condition C0 to condition C2, 30 when comparing C1 to C2, 40 when comparing C3 to C4 (1.5×-fold change) and 46 when comparing C4 to C5 (2×-fold change) (Fig. 2). The CSLog-PI score yielded 28, 29, 36 and 43 correct detections for the same four pairs of conditions, respectively (Fig. 2). In particular, when considering higher amounts of spiked-in UPS1 proteins, both methods detected nearly all proteins (46 out of 48) at $N_D = 100$ for comparison C3 vs. C4. For C5 vs. C6, CSNorm-PI detected all 48 UPS1 proteins at $N_D = 60$ (Table 2), whereas CSLog-PI missed 4 UPS1 proteins among its 60 most significant ones and only 1 among its 100 most significant ones.

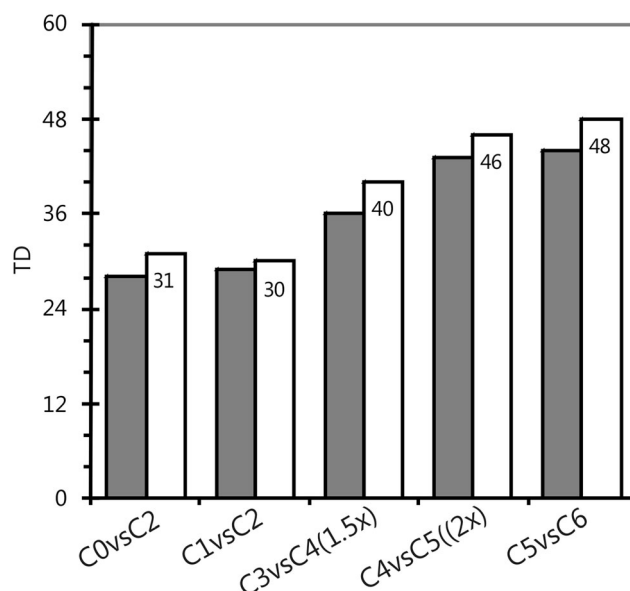


Fig. 2 Number of true discoveries for methods CSLog-PI (dark bars) and CSNorm-PI (white bars) for the hardest detections. The number of discoveries is fixed at $N_D = 60$ across all pairs of conditions.

Table 2 Summary of results at fixed numbers of discoveries $N_D = 60$ and $N_D = 100$ for two pairs of conditions with highest UPS1 standard concentrations. No differences in the detections were observed for $N_D > 100$ discoveries

Sample pair	Score	$N_D = 60$	$N_D = 100$
C4 vs. C5	CSLog-PI	43	46
	CSNorm-PI	46	46
C5 vs. C6	CSLog-PI	44	47
	CSNorm-PI	48	48

Consequently, CSNorm-PI reaches its peak performance with substantially fewer total discoveries, and therefore with a much smaller real FDR. In the remaining comparisons both scores produced exactly the same number of true discoveries when setting $N_D = 60$ (Fig. S1, ESI[†]).

When examining the protein list obtained using CSNorm-PI in more detail, it was observed that the missed proteins at $N_D = 60$ in comparison C4 vs. C5, UPS1 human proteins interleukin-8 (8.4 kDa) and thioredoxin (12.4 kDa), were both observed with only one peptide across these conditions. In comparison C5 vs. C6, these two proteins were detected as differentially abundant by CSNorm-PI at $N_D = 60$ (protein *p*-value $< 10^{-5}$), with 2 peptides observed for each of them. As expected, differential detection was thus easier when additional peptides were identified.

Checking for the direction of variation

In multiple recent biomarker discovery studies (see *e.g.* ref. 23), the relative direction of variation of a detected protein does not appear explicitly in the descriptions and it is not clear whether this direction is taken into account. Nevertheless, the knowledge of whether a protein concentration increases or decreases across two phenotypes is usually critical to understanding the underlying biological rationale for the observed variation.

Table 3 Number of UPS1 proteins detected as differentially abundant for two choices of the total number of discoveries ($N_D = 30, N_D = 120$) and for comparing conditions C3 to C4 and C4 to C3

Comparison	N_D	TD up	TD down
C3 vs. C4	30	27	0
C3 vs. C4	120	44	0
C4 vs. C3	30	0	27
C4 vs. C3	120	0	44

In the present dataset, all protein concentrations evolve in the same direction from one condition to another, *e.g.* $[\text{prot}]_{C1} < [\text{prot}]_{C2} < [\text{prot}]_{C3}$... due to our experimental design. In clinical experiments, the concentrations of the desired markers can be either increasing or decreasing across two phenotypes or clinical conditions. Hence there is no *a priori* knowledge about the direction of variation of a putative marker protein.

Here, the direction of variation was assumed unknown and taken into account when measuring performance. The lists of proteins detected as significantly changing across two conditions was split into two disjoint subsets, depending on whether the class score for the protein was going up or down from one condition to the other. Any detection of a UPS1 protein for which the direction of variation was opposite to the spiked-in amounts (Table 1) was declared a false positive. Overall, none of the UPS1 proteins detected were assigned the wrong sense of variation (Table 3). The symmetry of the equation defining our statistics (eqn (2)) ensures that the same be true when considering proteins whose concentrations decrease from the first to the second condition, which was *e.g.* verified on comparison C4 vs. C3 (Table 3, lower half).

Detection results as a function of target FDR

A common way of determining a list of differentially abundant proteins is to specify a target FDR.^{3,12,14,15,36} For example, in ref. 12, the “nominal” FDR is set at 0.05, and the BH method for correcting for multiple testing is used to determine a corresponding *p*-value threshold or, equivalently, the number of proteins detected as significantly differentially abundant at this target FDR.

Requiring a small realized FDR is useful in assigning peptide sequences to spectra, in which a large proportion of a spectra set are expected to have correct sequence assigned. However, such constraints are unrealistic in the context of finding differentially abundant proteins, in which the small *n*, large *d* scenario has pronounced effect, particularly because most candidates are true negatives. Sensitivity cannot be adequately evaluated under the constraint of a small realized FDR threshold.

To illustrate this point, we show the results of selecting a range of target FDR levels for score CSNorm-PI (Fig. 3). On the left panel is plotted the total number of discoveries N_D and the number of true discoveries TD as a function of target FDR values, over the interval $0 < \text{FDR}_{\text{target}} \leq 0.25$, for comparison C3 vs. C4. The realized FDR is $(N_D - \text{TD})/N_D$, which is plotted in the right panel for $0 < \text{FDR}_{\text{target}} \leq 0.65$. For target FDR values inferior to 2.3×10^{-2} , the number of discoveries remains low (< 3); this is most likely due to the difficulty of this comparison, with a 1.5-fold increase in UPS1 proteins concentrations from C3 to C4. The number of discoveries then jumps to 19, with 18 true discoveries, the number of false discoveries remaining low (realized FDR < 0.1) until a target FDR of about 0.09. The number of false discoveries then increases abruptly, with a second bump in the curve of the total number of discoveries at a target FDR of about 0.18.

Table 4 shows similar comparisons between target FDR and realized FDR for other comparisons, at the most commonly used target FDR thresholds. At a fixed target FDR of 5%, realized FDR values undergo a regular increase with the concentration of the samples that are being compared; it thus increases from 0.111 for comparison C0 vs. C1 to 0.455 for comparison C5 vs. C6. Note that for all 3-fold comparisons, discoveries were also obtained at a target FDR of 0.001, with realized FDR ranging from 0.083 for comparison C1 vs. C2 to 0.213 for comparison C5 vs. C6. In addition, for comparison C5 vs. C6, a 3-fold increase between the highest concentrations of UPS1, perfect recall was achieved at a target FDR of 0.001 (0.213 realized FDR); at increasing target FDR values, it becomes increasingly difficult to have both high sensitivity and low realized FDR (equivalently, high precision).

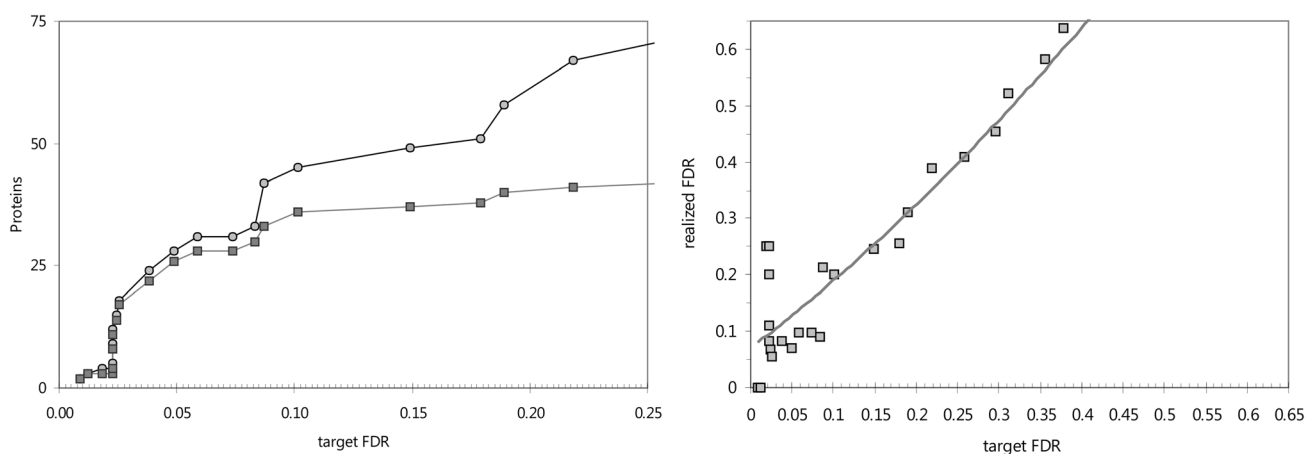


Fig. 3 Detection using CSNorm-PI on C3 vs. C4 as a function of target FDR. Left, discoveries (N_D , gray circles) and true discoveries (TD, black squares); right, realized FDR as a function of target FDR.

Table 4 Number of discoveries N_D and true discoveries TD at target FDR values of 0.001, 0.01 and 0.05 using score CSNorm-PI

Comparison	Target FDR	N_D	TD	Realized FDR	Recall
C0 vs. C1	10^{-3}	6	5	0.167	0.125
	0.01	8	7	0.125	0.146
	0.05	9	8	0.111	0.167
C0 vs. C2	10^{-3}	29	26	0.103	0.542
	0.01	31	26	0.161	0.542
	0.05	40	28	0.300	0.583
C1 vs. C2	10^{-3}	24	22	0.083	0.458
	0.01	30	27	0.100	0.563
	0.05	32	28	0.125	0.583
C2 vs. C3	10^{-3}	48	41	0.146	0.854
	0.01	50	41	0.220	0.854
	0.05	52	42	0.192	0.875
C3 vs. C4	10^{-3}	0	0	n.d.	0
	0.01	2	2	0	0.042
	0.05	28	26	0.071	0.583
C4 vs. C5	10^{-3}	49	44	0.102	0.917
	0.01	54	45	0.167	0.938
	0.05	68	46	0.324	0.958
C5 vs. C6	10^{-3}	61	48	0.213	1
	0.01	73	48	0.342	1
	0.05	88	48	0.455	1

Overall, the realized FDR values are thus controlled for all the comparisons considered (1.5 \times -fold change to 3 \times -fold change), with an increasing difference between target FDR and realized FDR at increasing concentrations of UPS1 proteins.

Comparison with other methods

The results above are consistent with previous findings, using similar ground-truth data sets, that showed realized FDR to be higher than target FDR, and that better recall was achieved by accepting a higher FDR, particularly among proteins at low concentrations.^{9,12}

More specifically, the difficulty in maintaining high sensitivity at high precision is demonstrated in a comprehensive series of simulated experiments in ref. 15, with target FDRs set at 0.01 and 0.05: none of the seven methods tested achieves high sensitivity together with high precision on most experiments, including two corresponding to C0 vs. C1 and C0 vs. C2, albeit with only three samples per class. Indeed, several methods return no protein with differential detection on these comparisons, and even the best-performing method, the SAM nonparametric test inherited from genomics, can only reach high sensitivity at the expense of a very high realized FDR; compare our Table 4 with Table 3 in ref. 15.

Methods based on peptide spectral counts (SC) have been proposed in multiple studies,^{37,38} and were used here for the sake of comparison. Since SC values are not on a log scale, the scores that were tested are CSPep and CSNorm. The application of these scores to SC values produced very similar results on the four pairs of conditions tested, hence only the results obtained using CSNorm-SC are shown (Fig. S2, ESI[†]). When compared to

CSNorm-PI, the SC-based scores returned between 1 (C1 vs. C2 and C4 vs. C5) and 5 (C3 vs. C4) fewer true detections than CSNorm-PI.

The performance of the CSNorm-PI score in detecting differential abundance was further compared with a protein-based method, NSI,³³ and the peptide-based linear regression model MSstats.³ Performance was assessed on comparisons C0 vs. C1, C0 vs. C2 and C3 vs. C4, which best discriminated among the methods; these are among the most difficult comparisons, the first two due to the low concentrations in UPS1 proteins, the third due to the low fold change (1.5 \times).

NSI was initially proposed as a plain quantification measure, which performed well across technical replicates.³³ Here, we implemented it as described in the Methods section. On comparison C0 vs. C1, method NSI detected 7 and 9 UPS1 proteins respectively at $N_D = 30$ and $N_D = 90$, while method CSNorm detected, respectively, 9 and 12 UPS1 proteins at these N_D values (Fig. 4). On comparison C0 vs. C2, NSI performed well at $N_D = 30$ discoveries, with 24 UPS1 proteins detected. However, between 60 and 150 discoveries, the number of discovered UPS1 proteins remained constant at 27. In comparison, CSNorm-PI was constant at 31 UPS1 discoveries over the same range. Method CSLog-PI also performed better than NSI over the whole range of number of discoveries Fig. 4.

By considering all the proteins to which a p -value was assigned by MSstats, this method could detect one UPS1 protein for comparison C0 vs. C1 at $N_D = 30$, three UPS1 proteins for comparison C0 vs. C2 at $N_D = 30$ and four among its 120 best discoveries for the latter comparison. This is much below the NSI-based detection as well as both our scores (Fig. 4, top and middle). On comparison C3 vs. C4, MSstats detected 27 UPS1 proteins for $N_D = 30$, sitting between our reference method (26) and CSNorm-PI (28 true discoveries), and it detected 45 UPS1 proteins for $N_D = 60$, while CSNorm-PI needed 120 discoveries (110, exactly) to detect 44 UPS1 proteins. MSstats also performed better than NSI in this setting (Fig. 4, bottom). It should be noted that MSstats only assigns a p -value to proteins with observed values in both conditions, hence its poorer performance on comparisons involving C0.

Score CSNorm was further compared to method NSAF.⁷ Again, CSNorm performed much better than this score on comparison C0 vs. C1, better than it on C1 vs. C2, and similarly on comparison C3 vs. C4, with NSAF largely outperformed at $N_D = 30$, but performing better by one to two true discoveries at higher N_D values (Fig. S3, ESI[†]).

Overall, CSNorm is the only score that performs well over all the range of concentrations used, thus making this method a good candidate algorithm for biomarker detection over the dynamic range observed in clinical samples.

Discussion

We propose a new, nonparametric method for detecting differentially abundant proteins in biological samples using unlabeled mass spectrometry. Our results show that this method

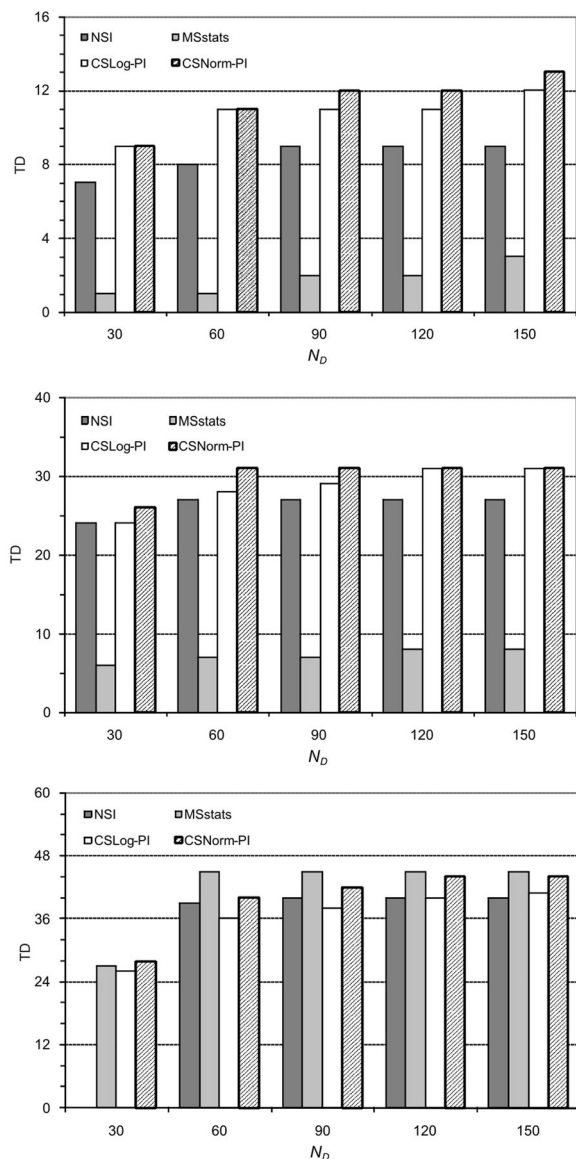


Fig. 4 Detection of UPS1 proteins when comparing condition C0 to C1 (top), C0 to C2 (middle) and C3 to C4 (bottom) at different numbers of discoveries N_D .

allows for the accurate detection and abundance determination of proteins based on peptide MS1 signals. This remained true even in the presence of multiple null values, making the current method superior to other statistical methods for the determination of differential abundance of proteins.

Specificity of the algorithm

Quantitative proteomics seeks to distinguish between biological phenotypes by finding differentially expressed variables, namely peptides or proteins, and using such “signatures” to identify cellular phenotypes from individual samples. Genomics, where variables such as mRNA concentrations are used to make phenotype inferences, has a similar goal. From the perspective of statistical inference and learning, both fields encounter the “small n , large d dilemma”, where n refers to the number of

samples (independent, biological replicates) and d refers to the number of measured variables. In genomics, d typically ranges from thousands to millions, and n typically ranges from tens to hundreds. In proteomics d is often in the thousands, while n is commonly below ten. Such an extremely unfavorable ratio is almost never seen in applications of statistical learning outside biology. The consequences of working under such constraints when analyzing differential abundance are profound, and severely limit what can be learned in a purely data-driven manner; see for example the discussion in ref. 39 and 40. As a result, model complexity must be tightly controlled in order to prevent “over-fitting” and improve consistency from study to study. One of the consequences of over-fitting is the lack of robustness in the analysis of FDR, in which there is a large gap between the target FDR and the realized FDR. Indeed, there are multiple studies in genomics concluding that “simpler is better” in this scenario.^{39,41}

In particular, the use of parametric models, such as parametric linear regression, can be dubious, at least with the sample sizes used in proteomics today. In addition, parameters present in parametric models are rarely known in advance, and must hence be estimated from the data. This leads to further instability in the small-sample regime encountered in most proteomics studies. Other approaches that require estimating multiple parameters, such as those based on least-squares, which do not require the Gaussian assumption, are also of limited relevance with very small sample sizes. Interestingly, despite these large bodies of research from the genomics field, nonparametric approaches are rarely (with few exceptions, *e.g.* ref. 4) applied to proteomics analyses.

Nonparametric tests are by definition distribution-free, and are therefore usually more robust than tests based on specific parametric distributions (*e.g.*, normality) with respect to departures from that distribution. In fact, even for data which are normally distributed, nonparametric tests are basically as efficient as t -tests when testing for a shift in values from one class to another,⁴² which is the standard characterization of “differential abundance”. Here each class has twelve samples, which is still quite small from a statistical perspective. However, this level of collected data must be taken into context when developing data for a differentially abundance measurement with respect to overall experimental design. Availability of sample amounts, time constraints, cost of mass spectrometry use, and data quality in terms of richness and variation are considerations for statistical analysis. Moreover, in proteomic measurements there is an extensive presence of null values in most studies, such as ours, which correspond either to peptides actually present but not detected due to signals below the threshold of noise, or to peptides actually not present in the samples. Overall, these data are clearly not normally distributed (see Fig. 5 and Fig. S4, ESI[†]). Such observations extend to the distributions of the peptides that we analyze in our peptide-centric method (see Fig. S5, ESI[†]). Nonetheless, the assumption of normality is made in approaches based on linear regression.^{3,12} Our approach enjoys the advantage of making no assumptions on the distribution of the data. Whereas we have focused on a

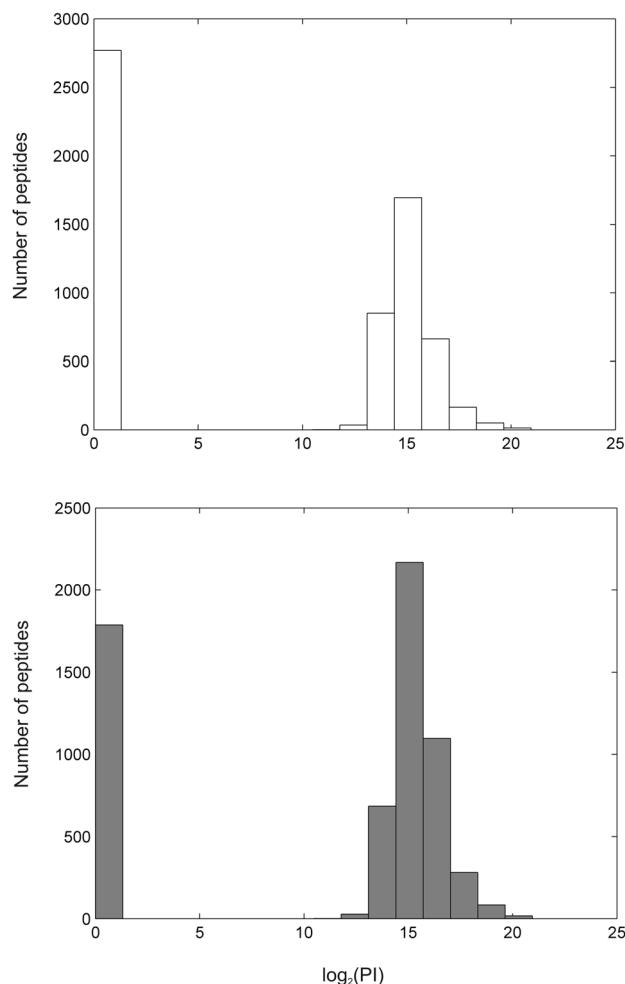


Fig. 5 Distribution of observed PI values in sample 1 for class C3 (top) and class C5 (bottom). Values are in log scale.

permutation-based method, other approaches may be appropriate for intensity data, such as the Wilcoxon test.

A further important feature of our approach is that robust estimates of all p -values are obtained by permutation analysis. This has become the default procedure in modern statistics, since it does not rely on any assumptions about the probability distribution of the score (test statistic). In permutation-based analyses, null data matrices are generated by randomly exchanging the class labels, thereby decoupling the class label and the observed peptide values within each sample. For each reshuffling of the labels, the test statistic is computed using the null data; the final p -value is the fraction of null matrices which result in a score at least as extreme as the score obtained on the real, unpermuted, data. In these null matrices, the overall sets of values obtained for each peptide across samples are preserved, as well as the peptide-to-peptide dependency structure. It is thus only the class label of a measured value which is modified, while existing peptide-to-peptide correlations are maintained. A permutation-based procedure such as ours thus overall provides more discriminating results than methods based on simulated data. The twelve repetitive injections we performed

for each concentration allows for a resolution in significance analysis of order 3.7×10^{-7} . In contrast, permutation testing is far too coarse with even six replicates (0.001 resolution), let alone three (0.05 resolution). The higher resolution obtainable with larger numbers of replicates is necessary to distinguish between conditions with small fold changes in the spiked-in concentrations.

Zero peptide values

As observed in multiple datasets, including ours, there may be many peptide values which are zero in the peptide data, even a majority within a class, depending on the concentration of the analyte. In the extreme case when all the peptide values are zero in one class, performance can be severely affected for some methods. This was the case, for example, with MSstats, which we used for comparison to our approach. Methods based on realistic proxies for relative protein abundance, such as MaxLFQ,⁴ where protein concentrations are reconstructed starting from ratios of peptide values from sample to sample (*i.e.*, ratios $y_{jk}/y_{jk'}$) are also compromised in the presence of many zero values. In contrast, the scores proposed here do not present such a limitation, since working with differences of summed (raw or transformed) peptide values across conditions (see eqn (2)). Detection on C0 vs. C1 and C0 vs. C2 using CSNorm-PI even yielded satisfactory realized FDR values when using a target FDR of 1%, as seen in Table 4.

Imputation of null values can sometimes be advantageous. Due to the high prevalence of zero peptide values in proteomics data, efforts have been made to evaluate the use of imputation in proteomic analyses.^{43,44} Still, due to limited success, we have chosen to avoid it, as recommended in ref. 43.

Some authors have proposed directly accounting for these zero values either by mixture modeling or by designing compound test statistics with separate terms for the zero and non-zero values; see ref. 13 and references therein. The CSNorm score implicitly accounts for zero values in the following sense. Recall that for comparing two classes a and b, the score for protein i is $T_i^{\text{norm}} = |S_a - S_b|$, where S_a (respectively, S_b) is the average of the normalized peptide values y_{jk}^{norm} over all peptides from protein i and all n_a samples from class a (respectively, n_b from class b). Let ν_a^+ (respectively, ν_b^+) denote the number of non-zero values in class a (resp., class b); note that normalizing does not change the set of zero values (see eqn (1)), so that the ν values are the same with (score CSNorm) and without (score CSLog) normalization. The fraction of non-zero values in class a is $\nu_a^+/n_a N_i$, and we have:

$$S_a = \left(\frac{\nu_a^+}{n_a N_i} \right) \left(\frac{1}{\nu_a^+} \sum_{k=1}^{n_a} \sum_{j=1}^{N_i} y_{jk}^{\text{norm}} \right)$$

and the same for class b with n_a replaced by n_b . Since class-wise summing over all values is obviously the same as class-wise summing over only the non-zero values, we have

$$S_a = (\text{fraction of non-zero values in class a}) \times (\text{average of non-zero values in class a}).$$

In comparing classes a and b, especially for low concentrations, the fraction of non-zero values may be as informative as their

average, whereas for high concentrations, the fractions may be near one and the decision is therefore mainly driven by the average of non-zero values.

Appropriate performance metrics

We have not used ROC (Receiver Operating Characteristic) curves to measure performance, as is commonly done in proteomics.⁹ The ROC curve plots the sensitivity (or recall) as a function of the false positive rate. However, whereas the ROC curve can be a discriminating metric for sample-based prediction, we must emphasize that it is usually not a useful performance measure for studies such as population-based biomarker discovery and the initial selection of candidates. From the perspective of statistical learning, searching for putative protein biomarkers from a large list of possible candidates is a prototypical feature selection problem: there are many candidates, with a majority of true negatives (non-discriminating features).

Moreover, feature selection, as studied here, is encountered throughout computational genetics (*e.g.*, finding DNA variants in GWAS), computational genomics (*e.g.*, finding differentially expressed genes) as well as, more recently, proteomics. In the former cases, the metrics of choice are the sensitivity (the fraction of true positives detected) and the false discovery rate, which expresses the ratio of the number of false detections to the total number of detections; in contrast, the false positive rate is a ratio to the total number of true negatives (non-markers), which is far larger. For example, in our current experimental settings, roughly 1000 proteins are involved in each pairwise comparison, while there are actually 48 true positives.

Suppose we have $TD = 24$ true detections among the $N_D = 30$ smallest p -values in a first case, whereas in the second case we detect the same number $TD = 24$ of true detections, but report $N_D = 60$ discoveries. In both cases the sensitivity is 0.5 ($24/48$).

The false positive rate will vary from $\frac{30 - 24}{1000} = 0.006$ (first case)

to $\frac{60 - 24}{1000} = 0.036$ (second case), both quite small. Since the sensitivity is the same, the difference in performance will barely be noticeable on an ROC curve. However, the realized FDR will

increase from $\frac{30 - 24}{30} = 0.2$ to $\frac{60 - 24}{60} = 0.6$, a huge difference, especially if follow-up experiments are costly. For these reasons, plotting sensitivity as a function of realized FDR (so-called “precision–recall” curves) is more relevant and informative than using ROC curves in feature selection problems such as those addressed in proteomics-based biomarker discovery.

Additionally, limitations to the utility of ROC curves were observed in the assessment of other protein quantification methods,¹⁰ in which some tests initially appeared to perform poorly until considering fewer discoveries.

Peptide-based detection

Since the first descriptions of abundance difference measurements in proteomics-based biomarker discovery, nearly all proposed algorithms were based on reconstructing protein-level measures, such as aggregated spectral count (SC) values^{6,38,45} or peptide

ion current ratios.⁴⁶ A major issue among proteomic label-free quantitation methods is how to aggregate peptide measurements associated with a given protein in order to most efficiently test the hypothesis that the underlying (but unobserved) concentration of this protein differs significantly from one class to the other. In particular, it is by no means evident, given the number of factors affecting peptide signal variability, how to combine the MS signals at the peptide level to obtain a physically meaningful value for each protein which is well correlated with the amount of protein originally present in the sample.³⁷ These factors include chromatographic separation,² the stochastic nature of DDA-MS, and the diverse range of physicochemical properties (*e.g.*, hydrophobicity, length, charge state *etc.*) for the different peptides from the same protein. Still, the actual outcomes of a proteomics experiment are observed values (mainly, PI or SC) associated with a peptide sequence within a sample. In recent years, a variety of peptide-centered methods have appeared in the literature; see *e.g.* ref. 4, 10, 11, 47 and 48.

In our approach we do not aggregate the raw peptide values to generate a score at the protein level. Instead, we first consider the distribution of the values across samples within each peptide separately. Our normalization step (eqn (1)) brings all peptides from a given protein into a common reference frame before summing these transformed values within each condition. This adjusts for possible heterogeneities in the generation of the samples or in the detection of peptides. This has been likewise shown in other methods to improve precision when quantifying proteins.¹⁰ In other existing methods,⁴ it is the ratio of two sample values within a fixed peptide which is assumed to best reflect the actual variations at the protein level. Our normalization step preserves those ratios whenever there is at least one zero value across the samples of the two conditions, which is the case in our experiments (hence $m_j = 0$ in eqn (1) and $y_{j,k}^{\text{norm}} / y_{j,k'}^{\text{norm}} = y_{j,k} / y_{j,k'}$). Indeed, there are usually many zero values, which may occur in actuality when comparing clinical phenotypes rather than being a limitation of the technology.

For clarity in our evaluation, we limited our analyses to peptide sequences assigned to a single protein. Peptides assigned to more than one protein produce a protein inference conundrum, in which the percent contribution of the peptide to each of the constituent proteins is unknown and likely to confound the quantification results for those particular proteins. Some quantification methods attempt to reconcile the contribution of degenerate peptide signals to each of the proteins using a variety of approaches.^{31,32} Alternatively, the effect of degeneracy is avoided simply by using the other peptides specific to each of those proteins for quantification, as each individual peptide is a distinct measurement of the protein abundance. For example, the removal of degenerate peptides from the analysis in favor of a ‘proteotypic’ approach was proposed more than a decade ago by several groups,^{49,50} and has gained further momentum over the past years with the growing popularity of SRM approaches. The effect of choosing to include or omit degenerate peptides on protein quantity

when nondegenerate peptides are available is difficult to assess, and may be influenced by the number of peptides observed for the protein, or simply the quality of those peptides for quantification, which is not assessed in shotgun MS. On a proteome-wide analysis typical in clinical settings, the proteins quantifiable solely based on shared peptides is typically minimal compared to the remainder of the observed proteins and can be evaluated as a group. These degenerate peptide analyses are rather extensive for the purposes of evaluating our approach and best suited for future developments of our method.

Dataset

The use of a two-species mixture with controlled quantities is not novel when testing analytical methods where ground truth should be known. Indeed, our dataset borrows heavily in design from a previously published method,^{20,22,23} with two distinctions. First, the additional dilution factor created smaller fold changes (as low as 1.5-fold) within the dilution series. These smaller fold changes allowed for more rigorous evaluation of method performance. Though large-interval fold changes are a necessity to adequately span the dynamic range of most mass spectrometers, they do not necessarily reflect true biological differences that may be observed in clinical samples. Even though our method performed as well as or better than alternative methods for all comparisons, we focused on the comparisons with the smallest fold changes. Second, we acquired twelve batch-developed replicates for each sample, as opposed to a minimal three replicates frequently found among testing datasets.^{9,20,23} As in other studies, such sampling can be seen as surrogates for biological replicates; with twelve replicates our method does obtain a better tradeoff between power and FDR. In a clinical context, three replicates are insufficient to provide the statistical power required for analysis of hundreds to thousands of proteins. Consequently, methods that perform well under minimal testing frameworks often struggle to perform in more complex, biologically relevant contexts.⁵¹ It is also worth mentioning that although we perform only two class comparisons, the data consist of six classes. Even though our two-classes approach corresponds to most clinical discovery experiments, this design provides the potential to extend this method, and others, to comparisons between more than two classes. Finally, we carefully inspected each chromatogram for consistency across runs to avoid the types of data acquisition artifacts, such as compressed chromatography and spray dropout, which are present in the dataset from Paulovich *et al.*²² As a result, the performance of the algorithms can be evaluated without the confounding effects arising from poor quality data.

Conclusions

The present study describes our nonparametric algorithm for the detection of potentially important protein and peptide markers from shotgun proteomics data. Rather than directly assigning a physical parameter to proteins by aggregating peptide spectral counts or precursor intensities, we first analyze the data using a

peptide-centric statistical approach. An overall score for the protein is then derived and its significance level is finally obtained using permutation testing. Our peptide-centric based statistical method was applied to a newly generated human UPS1 spike-in over a yeast background dataset, which is of common use for evaluating abundance detection performance. Across all comparisons used, even those involving the lowest concentration and even in the no-spike-in condition, our nonparametric algorithm showed sensitivity comparable to or better than existing parametric methods. The present, peptide-oriented, nonparametric method can have a high impact on the field of quantitative proteomics, where no consensus exists at present with respect to the statistical procedure, and where clinically-relevant markers are still scarce.

Authors' contributions

PS and DG designed the statistical methodology; PS created and ran the MATLAB codes for corresponding calculations; MRH, RLM and PS designed the experimental protocol; MRH performed the MS/MS acquisition and signal processing; DG and PS analyzed the performance of the statistics; all authors wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

PS and DG thank Professor Laurent Younès, JHU, for insightful comments. The work of DG and PS was partially supported by NIH NCRR Grant UL1 RR 025005. The work of MRH and RLM were supported by funds through the NIH National Institute of General Medical Sciences under Grants No. R01 GM087221 (to RLM) and 2P50 GM076547/Center for Systems Biology (to RLM), the National Heart, Lung and Blood Institute Grant No. R01 HL133135-01 (to RLM), and from the National Institute of Allergy and Infectious Diseases (NIAID) grant No. R21AI133335 (RLM).

References

- 1 S.-E. Ong and M. Mann, *Nat. Chem. Biol.*, 2005, **1**, 252–262.
- 2 M. Bantscheff, M. Schirle, G. Sweetman, J. Rick and B. Kuster, *Anal. Bioanal. Chem.*, 2007, **389**, 1017–1031.
- 3 T. Clough, S. Thaminy, S. Ragg, R. Aebersold and O. Vitek, *BMC Bioinf.*, 2012, **13**, S6.
- 4 J. Cox, M. Y. Hein, C. A. Lubner, I. Paron, N. Nagaraj and M. Mann, *Mol. Cell. Proteomics*, 2014, **13**, 2513–2526.
- 5 Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber and M. Mann, *Mol. Cell. Proteomics*, 2005, **4**, 1265–1272.
- 6 M. Li, W. Gray, H. Zhang, C. H. Chung, D. Billheimer, W. G. Yarbrough, D. C. Liebler, Y. Shyr and R. J. Slebos, *J. Proteome Res.*, 2010, **9**, 4295–4305.

- 7 B. Zybaylov, A. L. Mosley, M. E. Sardu, M. K. Coleman, L. Florens and M. P. Washburn, *J. Proteome Res.*, 2006, **5**, 2339–2347.
- 8 H. Choi, D. Fermin and A. I. Nesvizhskii, *Mol. Cell. Proteomics*, 2008, **7**, 2373–2385.
- 9 C. Ramus, A. Hovasse, M. Marcellin, A.-M. Hesse, E. Mouton-Barbosa, D. Bouyssié, S. Vaca, C. Carapito, K. Chaoui, C. Bruley, J. Garin, S. Cianférani, M. Ferro, A. Van Dorssaeler, O. Burlet-Schiltz, C. Schaeffer, Y. Couté and A. Gonzalez de Peredo, *Data Brief*, 2016, **6**, 286–294.
- 10 L. J. Goeminne, A. Argentini, L. Martens and L. Clement, *J. Proteome Res.*, 2015, **14**, 2457–2465.
- 11 Z. Ning, X. Zhang, J. Mayne and D. Figeys, *Anal. Chem.*, 2016, **88**, 1973–1978.
- 12 L. J. Goeminne, K. Gevaert and L. Clement, *Mol. Cell. Proteomics*, 2016, **15**, 657–668.
- 13 S. Taylor and K. Pollard, *Stat. Appl. Genet. Mol. Biol.*, 2009, **8**, 8.
- 14 M. Choi, C. Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean and O. Vitek, *Bioinformatics*, 2014, **30**, 2524–2526.
- 15 S. R. Langley and M. Mayr, *J. Proteomics*, 2015, **129**, 83–92.
- 16 N. Pavelka, M. L. Fournier, S. K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens and M. P. Washburn, *Mol. Cell. Proteomics*, 2008, **7**, 631–644.
- 17 R. A. Irizarry, C. Wang, Y. Zhou and T. P. Speed, *Stat. Methods Med. Res.*, 2009, **18**, 565–575.
- 18 B. Wu, *Bioinformatics*, 2006, **22**, 472–476.
- 19 J. Cox and M. Mann, *Nat. Biotechnol.*, 2008, **26**, 1367–1372.
- 20 J. Cox, I. Matic, M. Hilger, N. Nagaraj, M. Selbach, J. V. Olsen and M. Mann, *Nat. Protoc.*, 2009, **4**, 698–705.
- 21 W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing and N. G. Ahn, *Mol. Cell. Proteomics*, 2005, **4**, 1487–1502.
- 22 A. G. Paulovich, D. Billheimer, A. J. Ham, L. Vega-Montoto, P. A. Rudnick, D. L. Tabb, P. Wang, R. K. Blackman, D. M. Bunk, H. L. Cardasis, K. R. Clauser, C. R. Kinsinger, B. Schilling, T. J. Tegeler, A. M. Variyath, M. Wang, J. R. Whiteaker, L. J. Zimmerman, D. Fenyo, S. A. Carr, S. J. Fisher, B. W. Gibson, M. Mesri, T. A. Neubert, F. E. Regnier, H. Rodriguez, C. Spiegelman, S. E. Stein, P. Tempst and D. C. Liebler, *Mol. Cell. Proteomics*, 2010, **9**, 242–254.
- 23 O. Kannaste, T. Suomi, J. Salmi, E. Uusipaikka, O. Nevalainen and G. L. Corthals, *J. Proteome Res.*, 2014, **13**, 1957–1968.
- 24 K. E. Swearingen, M. R. Hoopmann, R. S. Johnson, R. A. Saleem, J. D. Aitchison and R. L. Moritz, *Mol. Cell. Proteomics*, 2012, **11**, M111 014985.
- 25 D. Kessner, M. Chambers, R. Burke, D. Agus and P. Mallick, *Bioinformatics*, 2008, **24**, 2534–2536.
- 26 J. K. Eng, T. A. Jahan and M. R. Hoopmann, *Proteomics*, 2013, **13**, 22–24.
- 27 A. Keller, J. Eng, N. Zhang, X.-j. Li and R. Aebersold, *Mol. Syst. Biol.*, 2005, **1**, 2005 0017.
- 28 A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold, *Anal. Chem.*, 2002, **74**, 5383–5392.
- 29 D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold and A. I. Nesvizhskii, *Mol. Cell. Proteomics*, 2011, **10**, M111 007690.
- 30 J. A. Vizcaíno, A. Csordas, N. del Toro, J. A. Dienes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q.-W. Xu, R. Wang and H. Hermjakob, *Nucleic Acids Res.*, 2016, **44**, D447.
- 31 Y. Zhang, Z. Wen, M. P. Washburn and L. Florens, *Anal. Chem.*, 2010, **82**, 2272–2281.
- 32 M. R. Hoopmann, J. M. Winget, L. Mendoza and R. L. Moritz, *J. Proteome Res.*, 2018, **17**, 1314–1320.
- 33 N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol and J. E. Schnitzer, *Nat. Biotechnol.*, 2010, **28**, 83–89.
- 34 F. L. Craciun, V. Bijol, A. K. Ajay, P. Rao, R. K. Kumar, J. Hutchinson, O. Hofmann, N. Joshi, J. P. Luyendyk, U. Kusebauch, C. L. Moss, A. Srivastava, J. Himmelfarb, S. S. Waikar, R. L. Moritz and V. S. Vaidya, *J. Am. Soc. Nephrol.*, 2016, **27**, 1702–1713.
- 35 Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, 1995, **57**, 289–300.
- 36 J. R. Whiteaker, H. Zhang, L. Zhao, P. Wang, K. S. Kelly-Spratt, R. G. Ivey, B. D. Piening, L. C. Feng, E. Kasarda, K. E. Gurley, J. K. Eng, L. A. Chodosh, C. J. Kemp, M. W. McIntosh and A. G. Paulovich, *J. Proteome Res.*, 2007, **6**, 3962–3975.
- 37 D. H. Lundgren, S.-I. Hwang, L. Wu and D. K. Han, *Expert Rev. Proteomics*, 2010, **7**, 39–53.
- 38 B. Cooper, J. Feng and W. M. Garrett, *J. Am. Soc. Mass Spectrom.*, 2010, **21**, 1534–1546.
- 39 D. Geman, C. d'Avignon, D. Q. Naiman and R. L. Winslow, *Stat. Appl. Genet. Mol. Biol.*, 2004, **3**, 19.
- 40 R. L. Winslow, N. Trayanova, D. Geman and M. I. Miller, *Sci. Transl. Med.*, 2012, **4**, 158rv11.
- 41 S. Dudoit, J. Fridlyand and T. P. Speed, *J. Am. Stat. Assoc.*, 2002, **97**, 77–87.
- 42 L. Breiman, *Statistics, with a view towards applications*, Houghton Mifflin, Boston, Massachusetts, 1973.
- 43 B.-J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds and K. M. Waters, *J. Proteome Res.*, 2015, **14**, 1993–2001.
- 44 C. Lazar, L. Gatto, M. Ferro, C. Bruley and T. Burger, *J. Proteome Res.*, 2016, **15**, 1116–1125.
- 45 X. Fu, S. A. Gharib, P. S. Green, M. L. Aitken, D. A. Frazer, D. R. Park, T. Vaisar and J. W. Heinecke, *J. Proteome Res.*, 2008, **7**, 845–854.
- 46 M. J. MacCoss, C. C. Wu, H. Liu, R. Sadygov and J. R. Yates, *Anal. Chem.*, 2003, **75**, 6912–6921.
- 47 Y. S. Ting, J. D. Egertson, S. H. Payne, S. Kim, B. MacLean, L. Käll, R. Aebersold, R. D. Smith, W. S. Noble and M. J. MacCoss, *Mol. Cell. Proteomics*, 2015, **14**, 2301–2307.
- 48 T. Suomi, G. L. Corthals, O. S. Nevalainen and L. L. Elo, *J. Proteome Res.*, 2015, **14**, 4564–4570.
- 49 B. Kuster, M. Schirle, P. Mallick and R. Aebersold, *Nat. Rev. Mol. Cell Biol.*, 2005, **6**, 577–583.
- 50 R. Craig, J. P. Cortens and R. C. Beavis, *Rapid Commun. Mass Spectrom.*, 2005, **19**, 1844–1850.
- 51 L. Gatto, K. D. Hansen, M. R. Hoopmann, H. Hermjakob, O. Kohlbacher and A. Beyer, *J. Proteome Res.*, 2016, **15**, 809–814.