BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 791–797
doi:10.1093/bioinformatics/btq036

Gene expression

Advance Access publication February 4, 2010

Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments

Marc Kirchner[1,2], Wiebke Timm[1,2], Peying Fong[3], Philine Wangemann[3] and Hanno Steen[1,2,*]

[1]Proteomics Center, Children's Hospital Boston, [2]Department of Pathology, Harvard Medical School and Children's Hospital Boston, Boston, MA and [3]Department of Anatomy and Physiology, College of Veterinary Medicine, Kansas State University, Manhattan, KS, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Mass spectrometry (MS) has become the method of choice for protein/peptide sequence and modification analysis. The technology employs a two-step approach: ionized peptide precursor masses are detected, selected for fragmentation, and the fragment mass spectra are collected for computational analysis. Current precursor selection schemes are based on data- or information-dependent acquisition (DDA/IDA), where fragmentation mass candidates are selected by intensity and are subsequently included in a dynamic exclusion list to avoid constant refragmentation of highly abundant species. DDA/IDA methods do not exploit valuable information that is contained in the fractional mass of high-accuracy precursor mass measurements delivered by current instrumentation.

Results: We extend previous contributions that suggest that fractional mass information allows targeted fragmentation of analytes of interest. We introduce a non-linear Random Forest classification and a discrete mapping approach, which can be trained to discriminate among arbitrary fractional mass patterns for an arbitrary number of classes of analytes. These methods can be used to increase fragmentation efficiency for specific subsets of analytes or to select suitable fragmentation technologies on-the-fly. We show that theoretical generalization error estimates transfer into practical application, and that their quality depends on the accuracy of prior distribution estimate of the analyte classes. The methods are applied to two real-world proteomics datasets.

Availability: All software used in this study is available from http://software.steenlab.org/fmf

Contact: hanno.steen@childrens.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 27, 2009; revised on January 22, 2010; accepted on January 25, 2010

1 INTRODUCTION

Current mass spectrometry (MS) technologies provide unpreceded mass accuracy and resolution, yielding detailed qualitative and quantitative insight into a sample under investigation. In fields including, but not limited to, proteomics, glycomics, lipidomics and metabolomics, MS is now an established workhorse methodology and high-resolution MS data acquisition have become the norm.

The accurate observed mass of a chemical compound carries more information than the mere molecular weight. It can be used to distinguish between peptides and non-peptides (Dodds *et al.*, 2006), cysteine-rich and/or highly acidic peptides, glycopeptides and non-polar peptides (Lehmann *et al.*, 2000); between peptide–oligonucleotide cross-links and DNA fragments (Pourshahian and Limbach, 2008; Steen *et al.*, 2001); between unmodified and post-translationally modified peptides (Bruce *et al.*, 2006); and between target precursors and interference peptides (Bateman *et. al.*, 2007; Steen and Mann, 2002; Tiller *et al.*, 2008; Zhang *et al.*, 2008), and it can pinpoint mass-defect calibrants in quantitative MS (Hall *et al.*, 2003). Accurate mass measurements enable on-the-fly analysis decisions that formerly required time-intensive off-line analytical procedures (Sweet *et al.*, 2006).

Mass defect, mass excess and fractional mass: the driving principle behind accurate mass-based decisions is the concept of *nuclear mass defect*, defined as the difference between the sum of the masses of the constituent nucleons and the measured exact mass of an atom (Inczedy, 1998). This difference is an instance of Einstein's special theory of relativity stating that mass and energy are interchangeable (Einstein, 1905): the nuclear mass defect accounts for the nuclear binding energy and is always non-negative. The concept of *mass excess* is a direct consequence of the nuclear mass defect: it is defined as the difference $d = m_{\mathrm{obs}} - m_{\mathrm{nom}}$ between the observed mass m_{obs} and the nominal mass m_{nom} of an element or compound. In particular, the mass excess of ^{12}C is defined as zero and mass excesses of other elements can either be positive (e.g. ^{1}H: 1.00783 and ^{14}N: 14.00307) or negative (e.g. ^{16}O: 15.99491, ^{32}S: 31.97207, ^{31}P: 30.97376 and ^{127}I: 126.90447). However, in practical experiments there is generally no prior knowledge about the chemical composition of an observed mass, and as a consequence, the nominal mass cannot directly be observed. Instead, high-resolution MS instruments provide the fractional mass $\phi(m) = m - \lfloor m \rfloor$, i.e. the fraction of mass after the decimal point of a mass measurement m. The fractional mass is confined to the interval $\phi(m) \in [0, 1)$ and, for mass excess larger than one, it wraps around to zero (Fig. 1).

[*]To whom correspondence should be addressed.

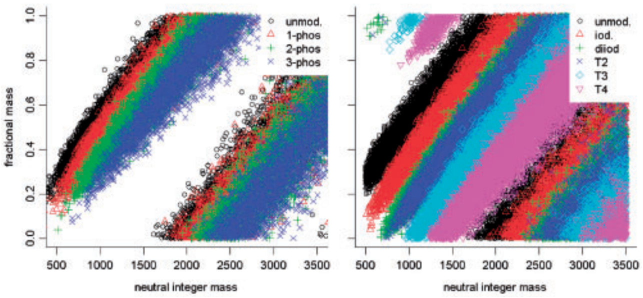

Fig. 1. Mass excess plots for the human protein phosphorylation (left) and the *Sus scrofa* protein iodination (right) sequence database training datasets with theoretical modifications. The plots show fractional mass over neutral integer mass for a subset of 10 000 randomly sampled peptides for each class. For phosphorylation the diagonal bands from left to right correspond to unphosphorylated, singly, doubly and triply phosphorylated peptides. For iodination the bands correspond to unmodified, iodinated, diiodinated, T_2 (a particular di-iodination), T_3 (tri-iodination) and T_4 (tetra-iodination) peptides. Because phosphorylation has a smaller negative mass excess than iodination, inter-class overlap for the phosphorylation data is significantly more pronounced than for the iodination data.

Mass excess in peptide analysis: the mass excess of a single hydrogen atom is ~0.0078 u and is the single major contributor to the average increase in peptide mass excess per atomic mass unit of 0.000488997. In conjunction with the well-defined, distinct (with the exception of leucine/isoleucine) stoichiometries of the 20 naturally occurring amino acids, the expected masses and mass excesses define a deterministic grid on which peptide/protein observations can be expected (Frahm *et al.*, 2006; Mann, 1995). Measurements that deviate from this grid stem either from non-peptide compounds or indicate peptide modifications. For example, (i) phosphorylation (+HPO_3) of S, T or Y residues causes a mass shift of 79.966331 u and bears a mass excess of −0.033669 u; (ii) oxidation (+O) of methionine (M) residues entails a mass shift of 15.994914 u at a mass excess of −0.005196 u; (iii) Palmitoylation (+$H_{30}C_{16}O$) of C, K, S and T residues or N-terminals causes a mass shift of 238.229666 u at a mass excess of 0.229666 u; and (iv) for different iodination modification types of histidine (H) or tyrosine (Y) residues, mass shifts and mass excesses range from 125.8966 u to 595.6128 u and from −0.1033546 u to −0.3872059 u, respectively.

Fractional mass/mass excess filtering: using fractional mass information as an input for a filter can be desirable in a number of different contexts: first, mass spectra are prone to noise effects and despite elaborate chemical protocols, incomplete separation and sample contamination with interfering compounds is a common problem in complex samples. Here, fractional mass filtering provides a straightforward approach to separating the peptide signal from the background and yields improved signal-to-noise ratios. This has been used to gain increased confidence in peptide mass fingerprinting results (Dodds *et al.*, 2006) and, known as forbidden region filtering or quantized peptide mass distribution filtering (Du and Angeletti, 2006), is a post-processing procedure in MS and liquid chromatography/mass spectrometry (LC/MS) feature extraction packages (Jaitly *et al.*, 2009; Renard *et al.*, 2008).

Second, following a similar rationale, mass excess filtering can be used for intelligent pre-filtering of high-resolution and high-accuracy tandem mass spectra in order to remove non-peptide fragments (Wolski *et al.*, 2006), to increase identification scores and to reduce database search times.

Third, common tandem MS acquisition schemes are based on data- or information-dependent acquisition (DDA/IDA), where fragmentation candidate masses are selected by intensity and are subsequently included in a dynamic exclusion list to avoid constant refragmentation of highly abundant species. With fractional mass filtering/classification, it is possible to bias the fragmentation process towards a set of analytes of interest (Bateman *et. al.*, 2007; Bruce *et al.*, 2006; LeBlanc *et al.*, 2009; Pourshahian and Limbach, 2008; Tiller *et al.*, 2008; Zhang *et al.*, 2008) and to automatically select suitable fragmentation techniques on-the-fly (Sweet *et al.*, 2006). Automated mass excess discrimination routines, thus, yield tailored preferential fragmentation of specifically modified peptides. Depending on the magnitude of the mass excess and the discrimination ability of the procedure, such an approach can greatly increase the fragmentation efficiency for a specific subset of molecules of interest.

Previous work: existing fractional mass filtering/classification methods are characterized by the following properties: (i) the input space directly corresponds to the 2D representation used in mass excess diagrams (Fig. 1). Measurements are described by nominal and fractional mass (Bruce *et al.*, 2006; Dodds *et al.*, 2006); (ii) to overcome the problem that the fractional mass is not a bijective measurement over m/z (the fractional mass returns to zero each time the mass excess switches to the next integer), the problem is linearized. Therefore, the integer mass difference for a specific class of compounds is estimated and used as an additive constant for all fractional masses above an optimized threshold (Bruce *et al.*, 2006); (iii) the accurate mass filtering/classification problem is formulated in terms of a linear discrimination problem (Bruce *et al.*, 2006; Dodds *et al.*, 2006; Pourshahian and Limbach, 2008; Zhang *et al.*, 2008). Depending on the analysis setup, classification thresholds are determined for lines parallel to the mass excess mean trend or symmetrically around it with widening boundaries for increasing masses; (iv) given the discrimination boundaries, crisp and/or probabilistic class assignments are reported; and (v) the theoretical classification performance is evaluated (Bruce *et al.*, 2006).

Shortcomings of existing approaches: The drawbacks associated with such a workflow are: (i) The linearization of the fractional mass plot is prone to error. There exist a considerable number of compounds that deviate from the mass excess mean trend (Bruce *et al.*, 2006) and, depending on the amount of deviation, the respective mass excess reconstructions may be erroneous. (ii) Linear discrimination approaches can only deliver limited partitioning of the mass domain and may fail to distinguish different overlapping classes of compounds (Pourshahian and Limbach, 2008). Mass excess mean trend-parallel classification boundaries (Bruce *et al.*, 2006) fail to account for the heteroscedastic character of the underlying data. (iii) For proteomics applications, previous publications provide training error estimates (Bruce *et al.*, 2006) derived from protein sequence databases. The proposed procedures lack cross-validation (CV) and consequently underestimate their true generalization error. It is therefore not clear how well the calculated theoretical errors transfer into practical application. Also, theoretical error estimates cannot provide insight into how variation in chemical enrichment protocols affects the performance of a fractional mass filtering routine.

This contribution introduces a 1D classification approach that can be trained to discriminate between arbitrarily distributed molecular classes of interest based on the information encoded in their accurate mass. The method makes use of a non-linear classifier and delivers class-wise posterior probabilities. In addition, we introduce a discrete derivate of the classifier that is amenable to fast on-the-fly filtering of precursor masses, which is of particular interest for hardware implementations. The viability of the proposed procedures is illustrated on real-world experimental datasets and classification performance is assessed using cross-validated quality measures and real-world peptide identification and modification information. We show that the theoretical classification errors hold in practical application but that transferability is dependent on prior knowledge about the modification state distribution. We discuss the practical consequences of this limitation and suggest a potential remedy. Additionally, we briefly investigate the potential of the procedure for multi-class fractional mass classification and provide the corresponding results. Although the presentation focuses on proteomics applications, the transfer of the underlying principles to other fields of research is straightforward.

The remainder of this contribution is organized as follows: we present the underlying methodology and statistical learning approach in Section 2 and describe the theoretical and experimental datasets and the experimental setup in Section 3. In sections 4 and 5 we list the results and discuss implications, respectively. Section 6 offers conclusions and an outlook.

2 METHODS

Mass excess-based classification is a 1D problem: the 2D representation shown in the mass excess plots in Figure 1 is suitable for human perception, every observation is characterized by its estimated nominal (Bruce *et al.*, 2006) or observed integer mass (Dodds *et al.*, 2006) on the x-axis and its fractional mass on the y-axis. This way, each measurement is split into its integer and real part, illustrating that the 2D display is merely a folded up representation of the 1D mass domain. Consequently, we may restate the mass excess/fractional mass filtering problem the following way: given a continuous mass range $\mathcal{M}$, we would like to find a function $f:\mathcal{M}\rightarrow\mathcal{S}$ that maps an observed mass $m\in\mathcal{M}$ to an element of the simplex $\mathcal{S}=\{(f_1,\dots,f_G)^T\,|\,\sum_g f_g=1\}$; here, G corresponds to the number of classes in the classification problem, $f(m)=(f_1(m),\dots,f_G(m))^T$ is a probabilistic assignment to these classes, and a crisp classification for a mass m is given by $\mathrm{argmax}_g\{f_g(m)\}$. This reduces the problem to finding the function f. See Supplementary Figure 1 for an illustration of $f(m)$ in a two-class classification scenario.

Random Forest training: we make use of a statistical learning approach to obtain the function f: (i) using protein sequence database information (see data section), we construct a representative set of ground truth data consisting of mass/label pairs $\boldsymbol{x}_i=(m_i,l_i)^T$, with $l_i\in\{1,\dots,G\}$; and (ii) based on this ground truth, we train a suitable classification method to deliver probabilistic predictions for new observations. The second step implicitly constructs f. Among an abundance of classifiers that could be used to obtain f, we choose the Random Forest (Breiman, 2001) due to its favorable training properties and the availability of a robust implementation in the form of the *randomForest* package (Liaw and Wiener, 2002) for the statistical programming language R (R Development Core Team, 2008). We predict posterior class probabilities; this allows the straightforward generation of receiver operating characteristic (ROC) plots. See the Supplementary Information for detailed parameters.

Discrete mapping classifier training: for online application of classification procedures, the evaluation time necessary to characterize a set of parent masses is crucial. With a trained classifier (i.e. an implicit construction of f) available, an obvious procedure to increase speed is the discretization of the mass domain and pre-calculation of the corresponding predictions. Because the resolution of any MS analyzer is finite we observe M distinct masses, and given a mass binning scheme $\boldsymbol{b}=(b_1,\dots,b_M)$, the class posteriors can be represented in an $M\times G$ matrix $\boldsymbol{P}$ with $p_{jg}=f_g(b_j)$. For prediction of an observed precursor mass m, the procedure determines the index $\hat{j}=\mathrm{argmin}_j\{|b_j-m|\}$ of the mass bin center closest to the observation and returns the corresponding class posteriors $f(b_{\hat{j}})$. If $\boldsymbol{b}$ equals the instrument binning, then $|b_j-m|=0,\forall j$. For each observed mass, class prediction then has $O(\log M)$ time complexity if the implementation uses a lookup technique or $O(1)$ if e.g. a hash table is used.

Neutral mass analyses are sufficient for high-resolution MS: current MS^1 experiments generally exhibit sufficient resolution to distinguish between isotopes of a compound and to determine its charge state z_u. With charge state information available for each observed monoisotopic mass m_u, the neutral mass $\bar{m}_u$ of a compound is given by $\bar{m}_u=z_u(m_u-m_{H+})$, where m_{H+} corresponds to the mass of a proton ($m_{H+}=1.007276\,\mathrm{u}$). This enables us to train the prediction function f in the neutral mass domain and allows common treatment of singly and multiply charged ions by transformation into their neutral mass representation.

Training set generation: to ensure that the training and test sets include a representative subset of the complete mass domain, a stratified sampling strategy was employed: we split the mass domain into R disjunct regions and for each region $r\in\{1,\dots,R\}$ and each class $g\in\{1,\dots,G\}$ we determine the set $\mathcal{R}_r$ of masses that falls into that region. For balanced training of the Random Forest classifier, the maximum number N_r of K-fold cross-validation samples that can be drawn from $\mathcal{R}_r$ is then given by the size N_r of the smallest class in $\mathcal{R}_r$. We uniformly draw K sets of samples $\mathcal{S}_{rk}$ without replacement from the $\mathcal{R}_r$, each of size $\lfloor N_r/K\rfloor$. The $(\sum_{r=1}^{R}N_r)\times K$ cross-validation data matrix $\boldsymbol{C}$ is then given by stacking the stratified sets of masses $\mathcal{S}_k=\bigcup_r\mathcal{S}_{rk}$ as column vectors. The respective region-wise label vectors $\boldsymbol{l}_r$ hold $\lfloor N_r/K\rfloor$-blocks of labels that are identical over all cross-validation folds and that can be stacked to form a $(\sum_{r=1}^{R}N_r)\times 1$ label vector $\boldsymbol{l}$. See Supplementary Table 4 for all combinations of K, N and R.

ROC and area under the ROC curve: to report classifier performance, we show ROC plots for the trained classifier in each of the application cases (see Section 4). The ROC plots visualize detailed tradeoffs between two different statistical quality features and allow comparison between experiments. In particular, we show ROC plots of the true positive rate (TPR) versus false positive rate (FPR) and compare results for the different classifiers using the area under the ROC curve (AUC).

We provide two displays for each dataset: (i) the cross-validated ($K=10$) training ROC/AUC (shown as solid lines in Fig. 2). These results are purely based on computational procedures; and (ii) the ROC for the application of the trained classifier to the corresponding iodination and phosphorylation datasets (dashed lines in Fig. 2). These curves show true classification performance on the experimental data. These ROCs/AUCs can only be calculated if peptides undergo identification. In contrast to a common application setup, we acquired this information for the datasets used in this study to allow for validation. ROC curves were generated using the ROCR package (Sing *et al.*, 2007).

3 DATA

3.1 Human protein phosphorylation dataset

3.1.1 Experimental data The phosphorylation dataset was taken from an ongoing SILAC-based quantitative phosphoproteomic study investigating the effect of a particular kinase inhibitor on HeLa S3 cells. Phosphopeptides were enriched with TiO_2 columns and the samples were analyzed by LC/MS on an LTQ Orbitrap (Thermo Scientific, San Jose, CA, USA) equipped with a micro-autosampler and a nanoflow HPLC system (both: Eksigent, Dublin, CA, USA). MS^2 spectra were acquired in a DDA mode. The 200 most intense fragment ions of each product spectrum were searched against a concatenated target and reverse decoy International Protein Index (IPI) human protein

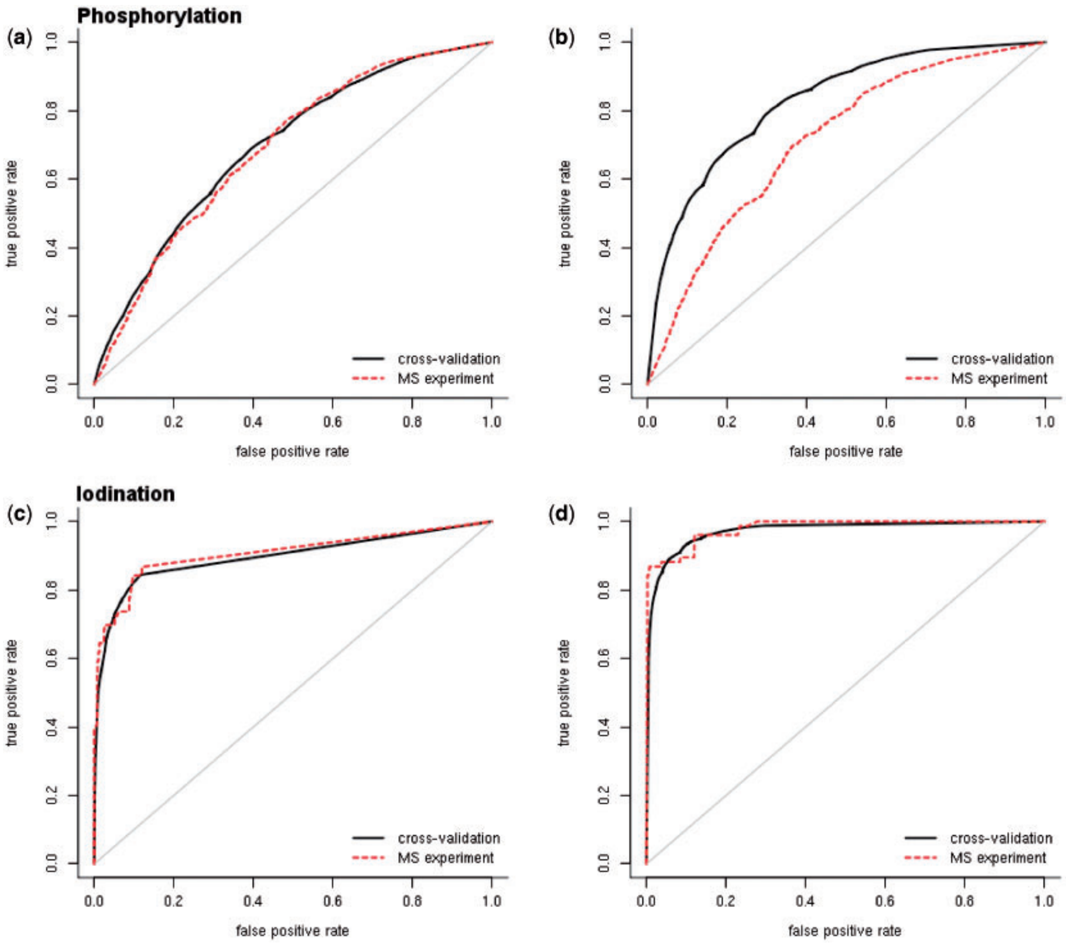

Fig. 2. ROC plots illustrating the performance of the discrete mapping classifier on theoretical and real-world data with correct (left column) and incorrect (right column) modification state prior distributions. Theoretical, cross-validation-based ROCs are shown as solid, validated real-world ROCs as dashed lines ($N = 128\,000$, $K = 10$). Panels a and c illustrate that the theoretical ROC provides an accurate estimate of the real-world performance if the modification state prior distribution is known. In panels b and d, the modification state prior distributions have not been adjusted to the distributions present in the data. Thus, the theoretical classification performance may misestimate the true error. The magnitude of this effect depends on the complexity of the underlying classification problem.

database (v3.36) using the Mascot search engine (Matrix Science, v2.2.04). One missed tryptic cleavage was allowed. See Supplementary Table 2 for detailed search parameters. Peptide identifications were accepted at a 1% global false discovery rate (FDR) (Choi *et al.*, 2008), corresponding to a Mascot peptide score cutoff of 30.5. The associated modification information was extracted. The resulting two label groups (unmodified and phosphorylated) served as a ground truth for determining the generalization ability of fractional mass classifiers derived from theoretical IPI digests (Experiments 1 and 2).

3.1.2 Computational data To obtain theoretical ground truth data, we extracted 69 013 protein sequences from the IPI human database (version 3.36). Tryptic *in silico* digestion (allowing one missed cleavage site) and exclusion of peptides with unknown or ambiguous residues yielded 2 164 634 unique peptide sequences. For mass calculations, cysteine residues were assumed to be carbamidomethylated. Peptides were allowed to be variably phosphorylated at one, two or three phosphorylation sites (serine, threonine, and tyrosine residues). We then generated four sets of peptide masses: (i) masses of all unmodified peptides; (ii) masses of all peptides with one occupied phosphorylation site; (iii) masses of all peptides with two occupied phosphorylation sites; and (iv) masses of all peptides with three occupied phosphorylation sites. Two training sets $\mathcal{P}_1$ and $\mathcal{P}_2$ ($N = 128\,000$ each) were constructed by: (i) drawing according to the modification ratios found in the MS^2 search results (2731:1779:17:2); and (ii) by drawing according to the modification distribution present in the peptide data (5.8:4.7:3.2:2) derived from the IPI database. Training set $\mathcal{P}_1$ serves as ground truth for Experiments 1 and 2, $\mathcal{P}_2$ for Experiment 2.

3.2 *Sus Scrofa* thyroglobulin iodination dataset

3.2.1 Experimental data Thyroglobulin protein samples were acquired from *Sus scrofa* thyroid glands. Cells were homogenized, centrifuged and the supernatant was collected for downstream analysis. The protein was denatured before gel separation and staining. The 600 kDa band was isolated for MS analysis. See the Supplementary Material for a detailed description of the sample preparation protocol. After tryptic digestion, peptide samples were analyzed by online C18 nanoflow reversed-phase HPLC (Eksigent nanoLC 2D) hyphenated to an LTQ Orbitrap mass spectrometer (Thermo Scientific). MS^2 spectra were acquired in DDA mode and searched with Mascot against a concatenated target and reverse decoy NCBI RefSeq *Sus scrofa* database

(Pruitt *et al.*, 2007, April 2009). Search parameters were tailored to iodination state detection (see Supplementary Table 3 for details). Peptide identifications were accepted at an FDR of 1%, corresponding to a Mascot score cutoff of 32.01. Extracted sequences, modification information and MS/MS precursor mass and charge were used as real-world ground truth in Experiments 1 and 2.

3.2.2 Computational data *Sus scrofa* protein sequence information was extracted from the NCBInr database, yielding 21 364 distinct proteins. Tryptic *in silico* digestion (1 missed cleavage) and exclusion of peptides with unknown residues resulted in 573 544 peptides. For sequence mass calculations, cysteine residues were assumed to be carbamidomethylated, histidine and tyrosine residues were subject to variable single and double iodination and tyrosines were additionally considered in their 3,3′-T_2, T_3 and T_4 forms (cf. Supplementary Table 1). Hence, the set of theoretical masses consisted of peptides that had no or exactly one of the variable iodination modifications. From this we generated two different collections of theoretical masses: (i) all iodinated masses were merged into a single collection for two-class classification (iodination versus unmodified); and (ii) diiodinated and 3,3-T_2-iodinated peptide masses (both carrying two iodines) were pooled, yielding a collection of masses representing a five-class problem (cf. Supplementary Table 1). From these collections, we generated three training sets $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ ($N = 128\,000$ each): (i) a two-class training set sampled according to the modification state distribution in the MS^2 search results (818:41:37:1:1:1); (ii) a two-class classification training set sampled according to the modification state distribution in the protein sequence database (4.0:1.9:1.8:1.1:1:1); and (iii) a five-class training set sampled according to the modification state distribution in the protein sequence database. For the last set, the mass range was cropped to 1000-3500 m/z to ensure applicability of the stratified sampling scheme. Training set $\mathcal{I}_1$ was used as ground truth in Experiments 1 and 2, $\mathcal{I}_2$ in Experiment 2 and $\mathcal{I}_3$ forms the basis of Experiment 3.

4 EXPERIMENTS AND RESULTS

Experiment 1: determining the practical value of theoretical generalization error estimates: we trained Random Forest and discrete mapping classifiers for the phosphorylation and iodination experiments. In order to derive results independent of modification state prior distribution influences (see Experiment 2), we trained on the sets $\mathcal{P}_1$ and $\mathcal{I}_1$ whose modification state distributions were derived from the MS^2 search results. Subsequent 10-fold cross-validation delivered a direct estimate of the theoretical generalization error. We then applied the trained classifiers to the experimental data sets and determined the real-world generalization error for comparison.

The results for Experiment 1 are shown in the rows of the ⊕ group in Table 1. For the discrete mapping approach the theoretical and practical error estimates are very similar: the observed real-world performances are within two SDs of the theoretical values. This is also illustrated in Figure 2a and c. These bounds also hold for the Random Forest applied to the *Sus scrofa* phosphorylation dataset; for the iodination dataset the estimated generalization error is conservative and underestimates the true performance attained on the experimental data (see Supplementary Fig. 2).

Experiment 2: judging the influence of the modification state prior distribution: we investigated the influence of the modification state prior distribution on the performance of the classifiers. Therefore, Random Forest and discrete mapping classifiers were trained on the $\mathcal{P}_2$ and $\mathcal{I}_2$ datasets that follow the distribution of modification states inherent to the peptide sequences given in the protein database digest. Again, cross-validated and real-world performance measures were obtained.

Table 1. Performance summary of the areas under curve (AUC) for all datasets and the Random Forest and the derived discrete mapping

		Random Forest		Discrete mapping	
	Data	10-fold CV	real	10-fold CV	real
⊕	phos.	0.697±0.005	0.694	0.696±0.004	0.689
	iod.	0.894±0.007	0.940	0.900±0.006	0.908
⊖	phos.	0.830±0.003	0.696	0.827±0.005	0.713
	iod.	0.973±0.001	0.987	0.973±0.001	0.979

AUCs in the *in silico* columns illustrate mean and empirical SD over 10 cross-validation runs. AUCs in the *real* columns were obtained by comparison against peptide identification results. Observations have been acquired with (⊕) and without (⊖) adjusting the modification state prior distribution in the training phase. The major observations are: (i) classifier performance is worse for the highly overlapping phosphorylation problem (see Figure 1); (ii) the accuracy of the predicted classification performance is highly dependent on the modification state prior distribution and problem complexity; and (iii) *in silico* performances of Random Forest and the faster discrete mapping are virtually identical.

For the protein phosphorylation dataset (Table 1, ⊖ group, top row and Fig. 2b), the discrete mapping as well as the Random Forest grossly overestimate the classification performance if the modification prior distribution, which was used during training, suggests a simpler classification task. Compared to the AUC acquired on the real-world data, we observe a misestimation of 13.4% for the Random Forest and of 11.4% for the discrete mapping classifier. On the protein iodination dataset (Table 1, ⊖ group, bottom row and Fig. 2d), the real-world AUC is well approximated despite the imperfect modification state prior distribution used in the training phase: both classifiers provide a conservative theoretical estimate of the true AUC.

Experiment 3: feasibility of multi-class fractional mass filtering: inspired by multi-class discrimination problems (Pourshahian and Limbach, 2008), we investigated if multi-class discrimination in fractional mass-dependent analyses is feasible. After training Random Forest and discrete mapping classifiers on the five-class $\mathcal{I}_3$ dataset, we calculated a cross-validated confusion matrix and estimated Cohen's κ and multi-class classification accuracy. Table 2 shows the Random Forest confusion matrix for the five-class *Sus scrofa* dataset. For equal class prevalences, a perfect classifier would yield $(0.2, 0.2, 0.2, 0.2, 0.2)^T$ on the diagonal. Off-diagonal elements are present for overlapping adjacent classes. The unmodified and T4 classes are adjacent due to the $[0,1)$ domain constraint of the fractional mass. Cohen's κ yields $\kappa = 0.691$ and the overall classification accuracy is 75.56%. The full set of summary statistics is available in Supplementary Table 5.

5 DISCUSSION

Theoretical fractional mass filtering performance measures hold in practical application: the results from Experiment 1 (cf. Fig. 2a and c) show that theoretical and practical performance estimates are close if the training and real-world modification state distributions are equal. In this case, the theoretical error estimates obtained for fractional mass classifiers trained on *in silico* datasets can be transferred to practical application and are valid estimates for the unknown and unobservable experimental errors. The required degree of modification state distribution similarity depends on the difficulty

Table 2. Confusion matrix for the five-class iodination state classification problem

		Predicted iodination state				
		∅	mi	di/T2	T3	T4
iod. state	∅	0.16 (20282)	0.03 (4020)	0.00 (289)	0.00 (46)	0.01 (738)
	mi	0.03 (4141)	0.14 (17713)	0.03 (3635)	0.00 (196)	0.00 (47)
	di/T2	0.00 (264)	0.03 (3634)	0.14 (18286)	0.02 (3056)	0.00 (174)
	T3	0.00 (48)	0.00 (188)	0.02 (3218)	0.15 (18810)	0.03 (3394)
	T4	0.01 (865)	0.00 (45)	0.00 (172)	0.03 (3492)	0.17 (21247)

Di-iodination and 3,3-T_2 both incorporate two iodine atoms and have been merged into a single class. The 10-fold cross-validated classification accuracy on the $N = 128\,000$ training dataset is 75.56%, Cohen's $\kappa = 0.691$.

of the classification problem (see detailed discussion below) and scales with the mass excess and the associated discrimination abilities of the targeted modification of interest.

Fractional mass classification is effective for targeted fragmentation: in Figure 2a and c as well as in Table 1 all real-world AUCs are larger than 0.68 and 0.90 for the phosphorylation and iodination problem, respectively. A random assignment to the associated classes according to the class priors would correspond to an AUC of 0.5, illustrating that fractional mass-based classification yields effective precursor selection. Consequently, there is a benefit in the combination of: (i) learning discrimination functions based on theoretical ground truth; and (ii) applying the resulting classifiers to characterize observations in practical MS experiments. More detailed efficiency analyses depend on the experiment at hand. For the phosphorylation data, we also calculated the positive and negative predictive values (PPV and NPV, see Supplementary Fig. 3). In the context of targeted fragmentation, the PPV can be understood as a measure of efficiency (the ratio of true positives among those reported as fragmentation candidates) and $1 - \mathrm{NPV}$ describes the miss rate (the ratio of false negatives among those reported as non-candidates). For the modification distribution-adjusted phosphorylation experiments, optimal values of PPV and NPV are between 60% and 70%, providing further evidence that the fractional mass classification is beneficial although the underlying classification problem is non-trivial.

On-the-fly classification is feasible and equally accurate: the numbers in Table 1 indicate that the impact of discretizing the neutral mass domain is marginal. For the two datasets used in this study, theoretical AUC values for the Random Forest and the discrete mapping classifier are within a single SD of each other. This observation also holds for all real-world classification experiments with the exception of the modification state distribution-adjusted *Sus scrofa* thyroglobin classification. Here, the Random Forest outperforms the discrete mapping approach quite significantly. Evidence from other thyroglobulin iodination datasets (data not shown) suggests that this is related to the comparatively small number of iodinated peptides present in the real-world dataset, which causes significant changes in the AUC with only very few contradicting classifications between the methods.

Classification performance depends on modification state prior distribution accuracy: real-world classification performance is influenced by two factors: (i) the complexity/difficulty of the underlying classification problem; and (ii) the discrepancy between the modification state prior distributions used for training and present in the real-world dataset. The former can easily be illustrated by comparing the theoretical ROCs and AUCs for the phosphorylation and iodination problems shown in Figure 2. Because the fractional mass shift for phosphorylation is smaller than for iodination, the classes representing the different phosphorylation states exhibit a larger overlap (cf. Fig. 1). Consequently, their discrimination is more difficult and the area under the theoretical ROC curve for the phosphorylation problem (Fig. 2a) is smaller than for the iodination problem (Fig. 2c). If the correct modification state distribution is used in the training step, the theoretical and real-world ROCs and AUCs are very similar (Fig. 2a and c), if not, the cross-validated performance estimate reported for the theoretical data may overestimate the true error rates attained in real-world application (Fig. 2b). It is important to realize that this is an observation that focuses on the theoretical error estimate with respect to its validity in real-world application: although more detailed sampling of masses of less prevalent modification states may yield improved performance (Fig. 2d), there is no guarantee that the error estimates are sufficiently conservative under such circumstances. In the phosphorylation dataset, the influence of imperfect prior distributions of the modification states had a huge influence on the accuracy of the theoretical AUC estimates but only little impact on the classification performance on real-world data (Supplementary Fig. 4). In cases where the classification problem is comparatively simple due to clear-cut, large mass defects, the performance predictions are relatively stable with respect to the class prior distributions, i.e. ROC curves and AUCs change only marginally for different prevalence settings. This is illustrated by the iodination ROCs in panels c and d in Figure 2. The reason for this behavior is that false positives and false negatives are costly in terms of MS^2 sampling efficiency and it is necessary to jointly optimize true positives and true negatives. For the optimization of the joint criterion, the (unknown) ratio between true positives and true negatives needs to be considered, causing a dependency on the modification state prior distribution (Hastie *et al.*, 2001).

However, in many proteomics experiments the peptide modification state prior distribution is unknown or in itself a subject of ongoing investigation. In such cases, available MS^2 datasets can be used to estimate a lower bound of PTM rates, but the applicability of such ground truth is heavily influenced by the bias introduced through sample preparation, the specific chemical enrichment protocols used, and DDA. Manually curated PTM databases (Bruce *et al.*, 2006) suffer from the same drawbacks and are not available for all organisms of interest. A potential approach to circumventing this problem is to include a data-dependent retraining scheme, where modifications identified by Mascot search as well as protein sequence databases serve as ground truth and the classifier undergoes constant, iterative re-training. Consequently, with a growing knowledge base, the classifier performance can iteratively adapt to the true modification prevalence and converge towards optimality. Such a procedure is computationally expensive and should thus only be applied to standardized, stable protocols.

Multi-class, on-the-fly fractional mass filtering is feasible: Table 2 indicates that the proposed procedure is amenable to multi-class discrimination problems. For the five-class iodination problem, classifying the iodination state by mere chance would yield an accuracy of 20% and making use of the fractional mass filtering approach thus yields a substantial improvement of >55%. Obviously, multi-class tasks are subject to the aforementioned constraints as well: depending on the difficulty of the discrimination problem, the applicability and usefulness of multi-class approaches can vary greatly. Nonetheless, given suitable sample preparation protocols, classification provides convenient multiplexing possibilities for the selection of precursors and precursor fragmentation technologies.

6 CONCLUSION

Based on the results presented in the previous sections we conclude that the theoretical fractional mass filtering/classification concept is applicable in practice. The proposed 1D representation and its combination with a non-linear classification approach provides a generic approach to obtain a discrete (and potentially multi-class) partitioning of the neutral mass domain. With a discrete mapping derived from the original Random Forest classifier, the method is sufficiently fast for real-time application. This allows on-the-fly fractional mass filtering for *arbitrary* mass excess structures.

ACKNOWLEDGEMENTS

We are indebted to Dominic Winter, Nurhan Ozlu and Judith Steen for the acquisition of the *Sus scrofa* thyroglobulin iodination and human protein phosphorylation mass spectrometry datasets.

Funding: Alexander von Humboldt Foundation (3.1-DEU/1134241, to M.K.).

Conflict of Interest: none declared.

REFERENCES

Bateman,K.P. *et al.* (2007) MSE with mass defect filtering for in vitro and in vivo metabolite identification. *Rapid Commun. Mass Spectrom.*, **21**, 1485–1496.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Bruce,C. *et al.* (2006) Probabilistic enrichment of phosphopeptides by their mass defect. *Anal. Chem.*, **78**, 4374–4382.

Choi,H. *et al.* (2008) Statistical validation of peptide identifications in large- scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.*, **7**, 286–292.

Dodds,E.D. *et al.* (2006) Enhanced peptide mass fingerprinting through high mass accuracy: exclusion of non-peptide signals based on residual mass. *J. Proteome Res.*, **5**, 1195–1203.

Du,P. and Angeletti,R.H. (2006) Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Anal. Chem.*, **78**, 3385–3392.

Einstein,A. (1905) Does the inertia of a body depend upon its energy content? *Ann. Phys.*, **18**, 639–641.

Frahm,J.L. *et al.* (2006) Accessible proteomics space and its implications for peak capacity for zero-, one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. *J. Mass Spectrom.*, **41**, 281–288.

Hall,M.P. *et al.* (2003) Mass defect tags for biomolecular mass spectrometry. *J. Mass Spectrom.*, **38**, 809–816.

Hastie,T. *et al.* (2001) *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, Vol. 22, Springer, New York.

Inczedy,J. (1998) *Compendium of Analytical Nomenclature: Definitive Rules 1997 (IUPAC Chemical Nomenclature)*. Blackwell Science.

Jaitly,N. *et al.* (2009) Decon2LS: an open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinform.*, **10**, 87.

LeBlanc,A. *et al.* (2009) Improved detection of reactive drug metabolites with bromine-containing glutathione analog using mass defect and isotope pattern matching. In *American Society for Mass Spectrometry, Annual Conference*. Philadelphia, PY.

Lehmann,W.D. *et al.* (2000) The information encrypted in accurate peptide masses-improved protein identification and assistance in glycopeptide identification and characterization. *J. Mass Spectrom.*, **35**, 1335–1341.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomforest. *R News*, **2**, 18–22.

Mann,M. (1995) Useful tables of possible and probable peptide masses. In *43rd Conference on Mass Spectrometry and Allied Topics*.

Pourshahian,S. and Limbach,P.A. (2008) Application of fractional mass for the identification of peptide-oligonucleotide cross-links by mass spectrometry. *J. Mass Spectrom.*, **43**, 1081–1088.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

R Development Core Team (2008) *R: Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renard,B. *et al.* (2008) NITPICK: peak identifcation for mass spectrometry data. *BMC Bioinform.*, **9**, 355.

Sing,T. *et al.* (2007) *ROCR: visualizing the performance of scoring classifiers.* R package version 1.0-2. Available at http://rocr.bioinf.mpi-sb.mpg.de/

Steen,H. and Mann,M. (2002) Analysis of bromotryptophan and hydroxyproline modifications by high-resolution, high-accuracy precursor ion scanning utilizing fragment ions with mass-deficient mass tags. *Anal. Chem.*, **74**, 6230–6236.

Steen,H. *et al.* (2001) Mass spectrometric analysis of a UV-cross-linked protein-DNA complex: tryptophans 54 and 88 of E. coli SSB cross-link to DNA. *Protein Sci.*, **10**, 1989–2001.

Sweet,S.M.M. *et al.* (2006) Strategy for the identification of sites of phosphorylation in proteins: neutral loss triggered electron capture dissociation. *Anal. Chem.*, **78**, 7563–7569.

Tiller,P.R. *et al.* (2008) Fractional mass filtering as a means to assess circulating metabolites in early human clinical studies. *Rapid Commun. Mass Spectrom.*, **22**, 3510–3516.

Wolski,W.E. *et al.* (2006) Analytical model of peptide mass cluster centres with applications. *Proteome Sci.*, **4**, 18.

Zhang,H. *et al.* (2008) Mass defect profiles of biological matrices and the general applicability of mass defect filtering for metabolite detection. *Rapid Commun. Mass Spectrom.*, **22**, 2082–2088.