Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data

Ermir Qeli[a,1,2], *Ulrich Omasits*[a,b,*,1], *Sandra Goetze*[a,b], *Daniel J. Stekhoven*[a], *Juerg E. Frey*[c], *Konrad Basler*[a], *Bernd Wollscheid*[b], *Erich Brunner*[a], *Christian H. Ahrens*[a,c,**]

[a]*Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

[b]*Institute of Molecular Systems Biology, ETH Zurich, Auguste-Piccard Hof 1, 8093 Zurich, Switzerland*

[c]*Agroscope, Institute for Plant Production Sciences, Research Group Molecular Diagnostics, Genomics & Bioinformatics, Schloss 1, Postfach, 8820 Wädenswil, Switzerland*

ARTICLE INFO

Article history:
Received 8 February 2014
Accepted 17 May 2014
Available online 27 May 2014

Keywords:
Targeted proteomics
Peptide detectability
Machine learning
Rank prediction algorithms
Proteotypic peptides
SRM

ABSTRACT

The in silico prediction of the best-observable “proteotypic” peptides in mass spectrometry-based workflows is a challenging problem. Being able to accurately predict such peptides would enable the informed selection of proteotypic peptides for targeted quantification of previously observed and non-observed proteins for any organism, with a significant impact for clinical proteomics and systems biology studies. Current prediction algorithms rely on physicochemical parameters in combination with positive and negative training sets to identify those peptide properties that most profoundly affect their general detectability.

Here we present PeptideRank, an approach that uses learning to rank algorithm for peptide detectability prediction from shotgun proteomics data, and that eliminates the need to select a negative dataset for the training step. A large number of different peptide properties are used to train ranking models in order to predict a ranking of the best-observable peptides within a protein. Empirical evaluation with rank accuracy metrics showed that PeptideRank complements existing prediction algorithms. Our results indicate that the best performance is achieved when it is trained on organism-specific shotgun proteomics data, and that PeptideRank is most accurate for short to medium-sized and abundant proteins, without any loss in prediction accuracy for the important class of membrane proteins.

Biological significance

Targeted proteomics approaches have been gaining a lot of momentum and hold immense potential for systems biology studies and clinical proteomics. However, since only very few complete proteomes have been reported to date, for a considerable fraction of a proteome

[*] *Correspondence to:* U. Omasits, Institute of Molecular Systems Biology, ETH Zurich, Auguste-Piccard-Hof 1, CH-8093 Zurich, Switzerland. Tel.: +41 44 633 6856.

[**] *Correspondence to:* C.H. Ahrens, Agroscope, Institute for Plant Production Sciences, Research Group Molecular Diagnostics, Genomics & Bioinformatics, Schloss 1, Postfach, CH-8820 Wädenswil, Switzerland. Tel.: +41 58 416 6114.

E-mail addresses: omasitsu@ethz.ch (U. Omasits), christian.ahrens@agroscope.admin.ch (C.H. Ahrens).

[1] Equal contribution.

[2] Current address: Swiss Reinsurance Company Ltd, Mythenquai 50/60, 8022 Zurich, Switzerland.

http://dx.doi.org/10.1016/j.jprot.2014.05.011

there is no experimental proteomics evidence that would allow to guide the selection of the best-suited proteotypic peptides (PTPs), i.e. peptides that are specific to a given proteoform and that are repeatedly observed in a mass spectrometer. We describe a novel, rank-based approach for the prediction of the best-suited PTPs for targeted proteomics applications. By building on methods developed in the field of information retrieval (e.g. web search engines like Google's PageRank), we circumvent the delicate step of selecting positive and negative training sets and at the same time also more closely reflect the experimentalist´s need for selecting e.g. the 5 most promising peptides for targeting a protein of interest. This approach allows to predict PTPs for not yet observed proteins or for organisms without prior experimental proteomics data such as many non-model organisms.

1. Introduction

Targeted quantitative proteomics approaches have been gaining momentum mainly due to their ability to generate accurate, reproducible, and complete quantitative data series [1–6], which are particularly relevant for systems biology studies and for clinical use. These approaches hold large potential for the quantitative evaluation of potential protein biomarkers for diagnostic and therapeutic uses e.g. from plasma samples [7,8]. Targeted quantitative proteomic approaches rely on selected reaction monitoring (SRM) [9], also referred to as multiple reaction monitoring (MRM) [10,11], which instructs a mass spectrometer to selectively focus on a combination of a peptide precursor ion mass and several fragment ion masses that are unique and specific for a target protein of interest. These so called transitions can be measured with about one hundred fold higher sensitivity [12], and allow to overcome the issue of low reproducibility in replicate measurements that is typical of shotgun proteomics [13]. To quantify selected proteins of interest, specific transitions of proteotypic peptides (PTPs), i.e. signature peptides that unambiguously identify one protein and which are repeatedly observed by mass spectrometry [1,14], have to be measured. Such a targeted SRM approach led to a proteomic map of the yeast proteome [15] and enabled the sensitive, reproducible quantification of induced perturbations over time within such a model system.

However, incomplete experimental proteomics data and large differences in the detectability of peptides from the same protein in a mass spectrometer complicate the selection of the best-suited peptides for targeted proteomics.

The identification of completely expressed proteomes by discovery shotgun proteomics has only recently been reported for a few organisms [16–18]. For the vast majority of organisms, however, public experimental protein data repositories like PeptideAtlas [19], PRIDE [20] or MassIVE (UCSD, San Diego) already contain a wealth of data, but the available datasets are far from being complete.

Secondly, while a varying percentage of peptides can uniquely identify one protein [21], only a subset thereof have additional – and mostly unknown – features that cause such peptides to be observed at very high frequency for a given protein. As a consequence, spectral counts for different peptides from the same protein can vary by orders of magnitude. Due to the difficulty in accurately predicting the best observable peptides purely in silico, most often the selection of target peptides is based on previous shotgun experimental data when available [2,3]. For so far unidentified proteins for a well-studied organism or for organisms without any prior experimental proteomics evidence, an accurate prediction of readily detectable and best-suited PTPs is desirable to fill this void. Accurate prediction of the best-suited PTPs is expected to significantly cut the costs of – and the time involved in – assay development, in particular for proteome-wide probes [5].

Since the first formulation of the peptide detectability problem in 2006 [22], several algorithmic approaches have been proposed to tackle this issue; these have used peptide detectability as a correction factor for spectral counts when estimating protein abundance [22,23], to improve protein inference [24–26], or to predict the best-suited PTPs for targeted quantitative proteomics either relying on MS/MS workflows [14,27–29] or for accurate mass and elution time (AMT) proteomics [30], which relies on high resolution MS [31]. Although these approaches use different machine learning concepts, they follow a generic schema in which they (i) extract a positive and negative training set from available experimental data; (ii) extract numerical features that characterize the peptide properties; (iii) apply a machine learning method on the training dataset to derive a model for prediction; and finally (iv) predict the detectability of peptides in different test sets.

The definition of a suitable training set is critical and different approaches have been explored. PeptideSieve [14] and CONSeQuence [29] focus on peptides that have been observed in 50% of all identifications of the corresponding protein in a set of experiments, while STEPP [30] relies on seen or not seen candidate peptides from proteins identified with at least one peptide in previous AMT studies [30]. Other tools consider peptides to be “best observable” if they have high signal peaks at the precursor ion (MS1) level (ESPPredictor [28]), or if they are observed from those proteins that have the highest total spectral counts (APEX [23]). Furthermore, the tools vary greatly with respect to feature selection and classification strategy. PeptideSieve [14] selects a few peptide properties to discriminate observed from unobserved peptides and uses a Gaussian mixture likelihood scoring function for prediction, STEPP considers 35 peptide features using a support vector machine (SVM) approach [30], while CONSeQuence [29] assesses more features including the predicted secondary structure of peptides and uses a combination of machine learning approaches. ESPPredictor [28] uses Random Forests classification [32], while Tang et al. use neural networks to classify peptides into detectable and undetectable based on different peptide features and, uniquely, information from their flanking regions [22].

In this study we consider the peptide detectability prediction problem as a ranking problem, similar to ranking problems in information retrieval and web searches. This approach has the great advantage that we can circumvent the delicate task of selecting a negative training set. In addition, the ranking approach ideally fits the problem of considering the varying frequency of peptide identifications within a given protein. The proteomics workflow is treated as a special ranking machine with a hidden (i.e. not fully understood) process that, for shotgun proteomics data as input, creates for each protein a list of peptides ranked according to their spectral counts (Fig. 1). We select a large training set to increase the number of proteins for which roughly a similar number of peptides were experimentally observed (i.e. around 50% of all theoretically observable peptides). These should represent good examples to optimally capture the difference in peptide detectability. Based on the ranking results for the training set, the most discriminative features among close to 600 different peptide features are identified and used to train a ranking classifier that is specialized to predict the ranking of peptides within a protein with respect to their detectability in LC–MS/MS experiments.

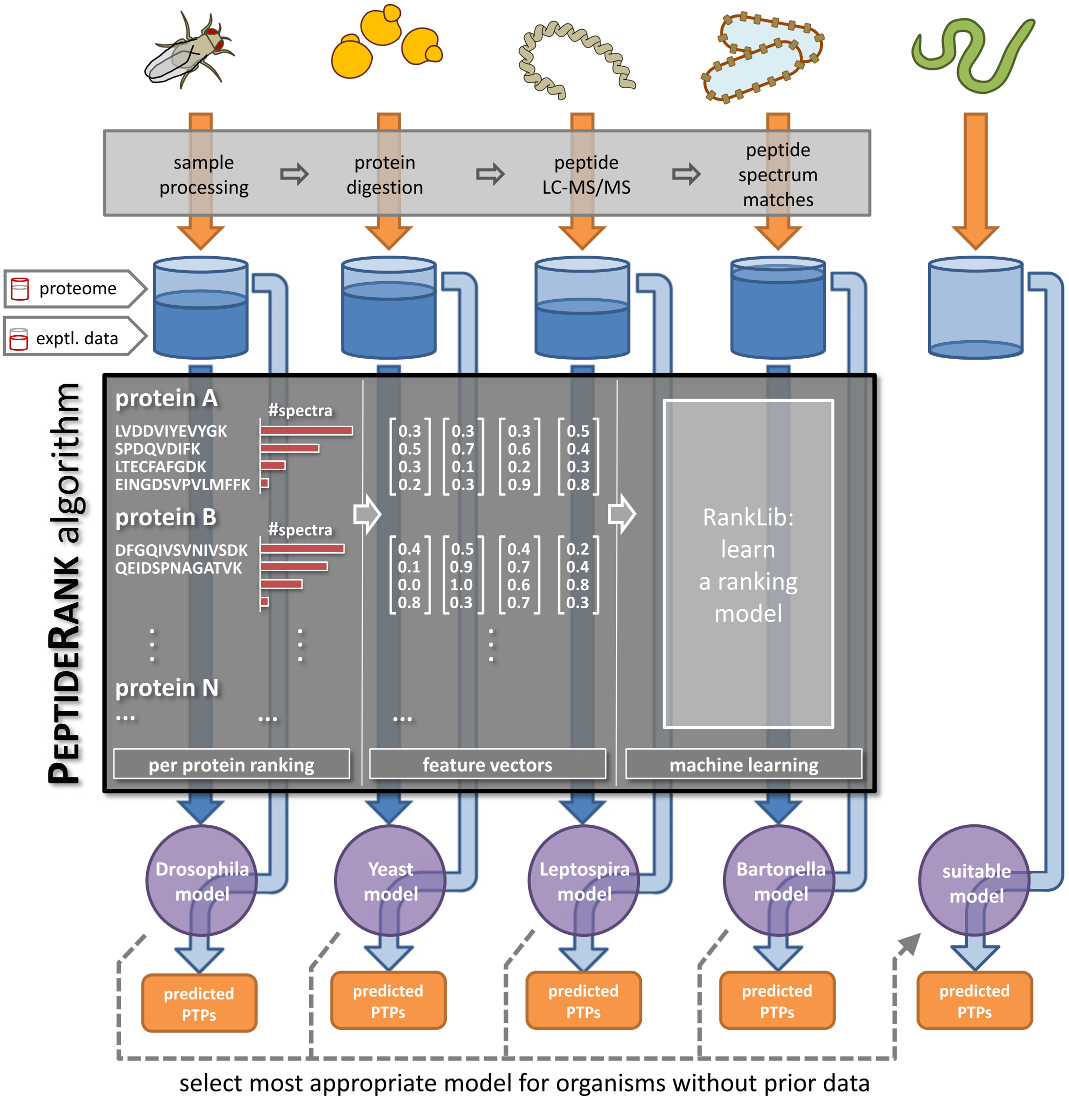

Fig. 1 – Overview of the PeptideRank approach to predict PTPs based on organism-specific data. Ranking algorithms for machine learning can use the information collected from a broad range of queries to learn the behavior of the ranking process and to subsequently apply it to rank documents in the context of a new, unrelated query. We apply this principle to a shotgun proteomics workflow (schematically represented by the light grey box). For each experimentally observed protein (protein A to protein N) all of its peptides within the visible range (see Methods) get ranked according to the respective observed sum of spectral counts (dark grey box; see also Fig. S1). By computing feature vectors, the information contained in the ranked peptide list can be used to model the behavior for peptides in the shotgun proteomics workflow with machine learning algorithms and to create organism-specific models (schematically shown for *Drosophila*, yeast, *Leptospira* and *Bartonella*). A ranked list of predicted PTPs (orange box) can then be obtained for not yet experimentally identified proteins (light blue arrows), or for organisms without prior experimental proteomics data (schematically shown by the green nematode).

The in silico ranking approach represents one additional solution for identifying and prioritizing the best suitable peptides for targeted quantitative mass spectrometry experiments, even in the absence of prior experimental shotgun proteomics data. We evaluate our approach using cross-validation on shotgun proteomics data from two Gram-negative prokaryotes, as well as from yeast and the fruit fly as examples of a simple and more complex eukaryote. Comparison of the PeptideRank results with those from existing software tools indicates that the rank-based approach is a very valuable addition: it complements the best existing predictors, and can provide better predictions, in particular when the training step can draw on previous experimental shotgun proteomics data from the organism for which the best-suited PTPs are to be predicted.

2. Methods

2.1. Datasets and data processing

To validate our methodology, we selected large shotgun proteomics datasets from two prokaryotic model organisms (*Leptospira interrogans*, *Bartonella henselae*), baker's yeast (*Saccharomyces cerevisiae*) as simple eukaryotic and fruit fly (*Drosophila melanogaster*) as a more complex eukaryotic model organism. All datasets had been generated using extensive sample fractionation, liquid chromatography (LC) and electrospray-ionization based tandem mass spectrometry (ESI–MS/MS). The *Leptospira* (Table S1) and fruit fly datasets (Table S3) were downloaded from PeptideAtlas [19] or acquired for this study. Data have been deposited to the ProteomeXchange Consortium (http://www.proteomexchange.org) via the PRIDE partner repository [20] with the dataset identifiers PXD000726 and PXD000730. The yeast dataset was obtained from B. Balgley [33] (Table S2) and the *Bartonella* dataset (Table S4) was published recently [18]. Peptide-spectrum matches (PSMs) were filtered for a PeptideProphet [34] probability of at least 0.9 (PeptideAtlas datasets), or for an estimated false-discovery rate of 0.01% (*Bartonella* dataset). Only fully tryptic peptides (no missed cleavage sites) longer than 6 amino acids but with a mass below 6000 Da were considered. This restriction is typical in bottom-up proteomics studies because shorter peptides have a rather low probability of being unique while heavier tryptic peptides are rare [35]. All peptides were classified with PeptideClassifier [21] and shared or ambiguous peptides implying proteins encoded by different gene models or different protein isoforms encoded by the same gene model were excluded from the training step (Table S5). Ranking of identified peptides for each protein was performed according to their experimentally observed spectral counts. A relevance was assigned to each rank position (see Table 1). To derive this ranking relevance $r(p)$ for a peptide p, we chose a logarithmic transformation of its spectral counts c_p:

$$r(p) = \log_2(c_p + 1).$$

Unique tryptic peptides that were not experimentally identified were added with zero spectral counts and zero relevance.

2.2. Calculation of peptide features

574 different numerical peptide features were calculated for each peptide (Table S7) and were evaluated for their influence on peptide detectability. These features included the 20 relative frequencies of each amino acid, 10 general peptide properties (length, mass, estimated isoelectric point, etc.), and 544 averaged physicochemical properties that were extracted from AAindex1 [36,37]. Values for each amino acid were downloaded from http://www.genome.ad.jp/aaindex and were scaled to the interval [0,1]. Next, the respective values were multiplied with the vector of amino acid frequencies for each peptide and divided by its length, resulting in a scaled physicochemical property average.

To analyze the redundancy among the 544 physicochemical properties we computed the scaled properties for 128,642 yeast peptides in the detectable range, and generated a matrix of pairwise correlations between the properties (Fig. S3A): several clusters of highly correlated and hence very similar physicochemical parameters are aligned along the main diagonal. As the 544 physicochemical features are linear combinations of the amino acid frequencies, the intrinsic dimension of their space cannot be more than twenty. Based on a hierarchical cluster and minimum spanning tree analysis (Fig. S3), a subset of 12 parameters that are most relevant for

Table 1 – Data format for spectral ranking.

Peptide	# spectra	Rel.	Rank
LEGANASAEEADEGTDITSESGVDVVLNHR	217	7.8	1
LVDDVIYEVYGK	33	5.1	2
SPDQVDIFK	21	4.5	3
LTECFAFGDK	18	4.2	4
EINGDSVPVLMFFK	10	3.5	5
ELQFFTGESMDCDGMVALVEYR	4	2.3	6
HGLEEEK	2	1.6	7
DIITGDEMFADTYK	1	1	8
QGDDIK	0	0	9.5
SYTLYLK	0	0	9.5

List of peptides for *D. melanogaster* protein CG4800-PA (translationally-controlled tumor protein homologue) along with their respective spectral counts, log-transformed relevance score (Rel., see Methods), and rank. For proteins in large experimental repositories like PeptideAtlas, the spectral count differences can span more than three orders of magnitude (Fig. S1).

rank-based prediction while containing as little redundancy as possible were selected as follows: a relevance for each physicochemical parameter was computed by ranking the peptides using the respective parameter and computing the average Discounted Cumulative Gain for the 5 top peptides (DCG@5) for all proteins of the yeast Peptide Atlas dataset. For each of the first 12 clusters, the parameter with the highest average DCG@5 was selected as the representative parameter for that cluster resulting in the 12 parameters that are most discriminative for rank-based prediction while containing as little redundancy as possible (Table S7B, Code Listing 1). These 12 parameters together with the 20 amino acid frequencies and the 10 general peptide properties yield a total of 42 peptide features (shown in Table S7) that are used for learning and predicting peptide-detectability rankings for each protein.

2.3. Learning to rank algorithms

Algorithms for learning and predicting rankings have been developed only in the past decade [38,39], in particular for application in the field of web search engines [40,41]. We tested several ranking algorithms for training and

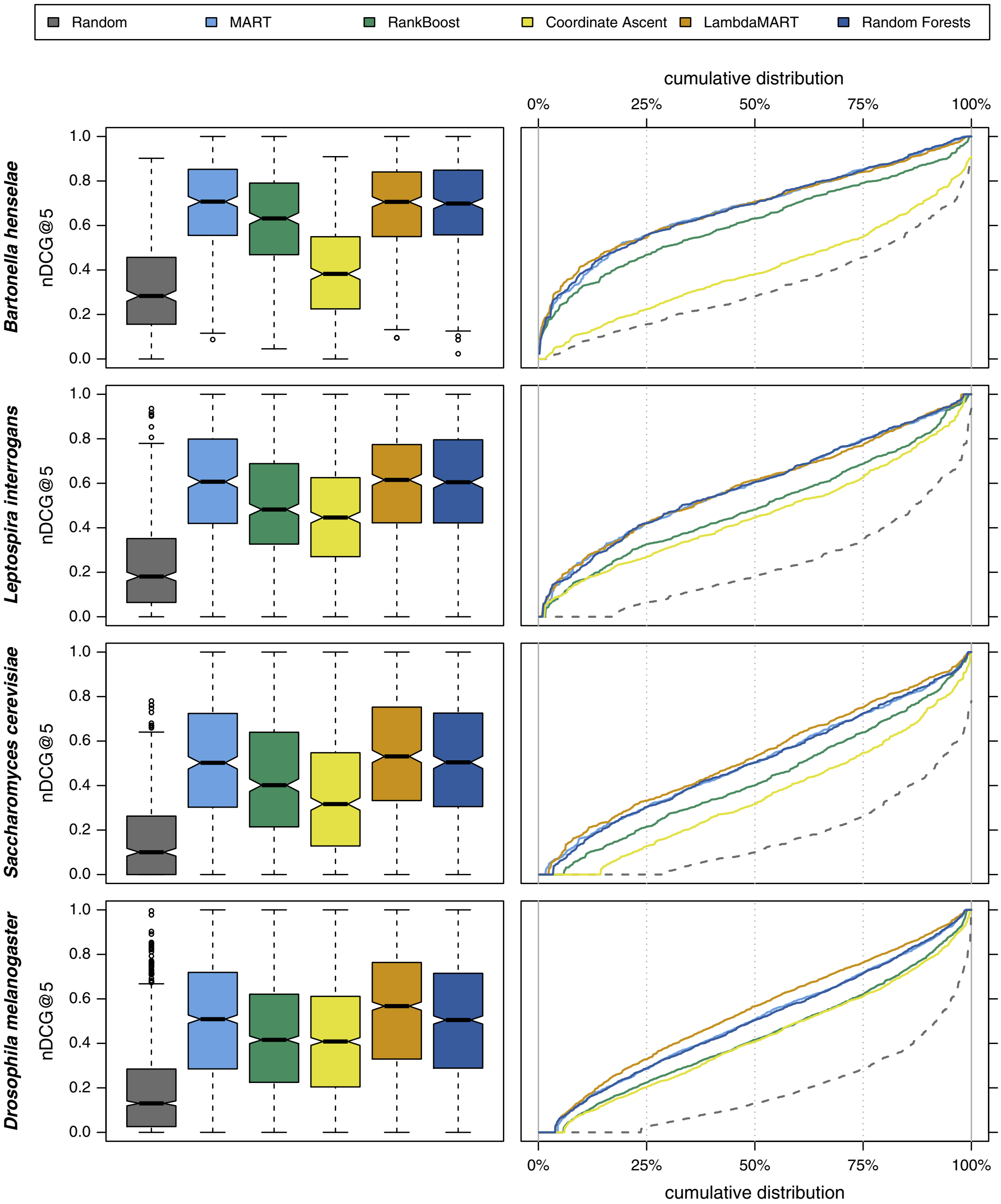

Fig. 2 – Comparison of different rank learning algorithms. We tested all rank learning algorithms provided by the RankLib library (v. 2.1) for their prediction accuracy of the top 5 peptides (nDCG@5). Five algorithms were consistently better than a random ranking (grey boxplot and dashed line in the cumulative distribution plot) on datasets from all four model organisms. Overall, LambdaMART (orange boxplot and line) showed the best performance.

classification including SVMLight [40] (http://svmlight.joachims.org) and learning to rank algorithms from the RankLib library (v. 2.1) (http://sourceforge.net/p/lemur/wiki/RankLib/). These included MART (Multiple Additive Regression Trees, a.k.a. Gradient boosted regression tree) [42], RankNet [43], RankBoost [44], AdaRank [45], Coordinate Ascent [46], LambdaMART [47], ListNet, [48] and Random Forests [32]. The best rank-based algorithm (LambdaMART, see Fig. 2) was chosen for the machine learning part of PeptideRank.

To train a classifier, the 500 proteins with the most spectral counts were considered and the 42 features that resulted from the feature selection process (see Results, Table S7) were calculated for their peptides. The peptide feature vectors are ranked by their relevance and input to the classifier, which learns a ranking model. For testing the learned classifiers on a specific organism, all proteins identified with at least 5 peptides were used, excluding the 500 proteins that were used for training.

The performance of PeptideRank was compared to previously published PTP predictors with a focus on models for LC–ESI–MS/MS. These included ESPPredictor [28], PeptideSieve [14] with both the MUDPIT–ESI and the PAGE–ESI predictor, and CONSeQuence, from which we chose the artificial neural network predictor (ANN), which had shown the best performance [29]. For the datasets obtained from PeptideAtlas, we also compare the results to the PeptideAtlas PSS, which provides a consensus score based on several predictors (Fig. S5). As PSS is only available for previously detected proteins of organisms listed in PeptideAtlas, this approach is not suitable as generic peptide detectability predictor and therefore not further discussed here.

2.4. Comparison of predicted ranking results

For a comparison of the ranking results of different classifiers we relied on the normalized Discounted Cumulative Gain (nDCG) method, a derivative of DCG [49], a metric that is commonly used in information retrieval to evaluate the performance of web search engines. nDCG measures the performance of the rank learning algorithm for the prediction of the top-k peptides; it is ideally suited to compare the ranking approach as it considers the relevance of different peptides observed within a protein based on their spectral count, and as such is more informative than e.g. Spearman rank correlation [50] or Kendall's τ rank correlation [51], which do not weight the ranked objects. nDCG weights more the prediction accuracy on the top results, and conversely penalizes less for the wrong predictions at the bottom of the list. The metric can be computed for the full list of peptides or for the top-k predictions, e.g. nDCG@5 for the top five peptides. It is defined as

$$nDCG@k = \sum_{i=1}^{k} \frac{2^{r(q_i)}-1}{\log_2(1+i)} \Bigg/ \sum_{i=1}^{k} \frac{2^{r(p_i)}-1}{\log_2(1+i)},$$

where r is the relevance of a peptide as defined above, q_i is the *i-th* peptide in the predicted ranking, and p_i is *i-th* peptide in the empirical peptide ranking.

2.5. Comparison of aggregated results across model organisms

We investigated how much of a benefit could be achieved when PeptideRank was trained and tested on shotgun proteomics data originating from the same model organism. This is relevant since all three major software solutions for PTP prediction (PeptideSieve, ESPPredictor, CONSeQuence) have been trained on yeast data. PeptideSieve (version 0.51, http://tools.proteomecenter.org) was run using the "PAGE_ESI" and "MUDPIT_ESI" model, without a score threshold. ESPPredictor was run online at the GenePattern platform (http://genepattern.broadinstitute.org). CONSeQuence was run online (http://king.smith.man.ac.uk/CONSeQuence/) using the "ANN" model. For comparison, only fully tryptic peptides without missed cleavage sites longer than 6 amino acids but with a mass of less than 6000 Da were considered. PeptideSieve, ESPPredictor, and CONSeQuence assign a global prediction score to peptides; for the comparison to PetideRank we considered all peptide scores not only those above the respective default threshold (e.g. 0.8 in the case of PetideSieve). The peptides were then ranked according to the original score of the respective prediction algorithm, which allowed us to calculate their nDCG values.

2.6. Analysis of factors potentially affecting predicted rank accuracy

Protein parameters (length, pI, grand average hydropathicity (GRAVY), topology class of membrane proteins), protein expression values (normalized spectral count abundance), as well as transcript expression values (normalized RPKM values, i.e. reads per kilobase per million mapped reads [52]) were calculated as described in [18]. The average spectral count per peptide for a protein is the mean spectral count for its peptides within the visible range. For yeast proteins, the number of sequence-modifying annotations in SwissProt/UniProtKB (version 2013_11) was extracted and counted considering the following annotations: glycosylation sites, transit-/signal-/pro-peptides, sequence conflicts, PTMs, initiator methionines, lipidation sites, cross-links, sequence variations, and splice variants.

To make sure a high predicted rank accuracy is not caused by obvious similarities (i.e. in amino acid composition) of the ranked peptides to the training-set peptides, but rather by capturing the underlying features of peptides that determine high detectability for a particular organism and experimental-setup, we also compared the distributions of minimal distances in the sequence space between testing and training peptides for the *Bartonella* dataset (see Fig. S4).

3. Results

3.1. Formulation as a ranking problem

In this study, we assessed the potential of machine learning algorithms to learn to rank peptides within a protein by using techniques similar to those used in the ranking of web documents in information retrieval. Proteins are treated like query terms and the observed peptides, ranked according to their spectral counts conceptually correspond to the ranked list of web sites resulting from a web search (Fig. 1). An organism-specific model is generated that can be used to predict the best-suited PTPs for as yet unobserved proteins (Fig. 1, orange boxes), and for proteins from organisms for which no prior experimental shotgun proteomics data exists, i.e., when using the parameter set across species boundaries.

In contrast to most existing approaches for prediction of the best-suited peptides for SRM [14,22,28,29], this approach does consider the frequency of the respective peptide identifications, instead of merely differentiating observed versus non-observed peptides within a given protein. Since the differences in observed spectral count for peptides originating from the same protein can span orders of magnitude (Fig. S1), the ranking approach is expected to benefit from utilizing this important type of information. We focus on pairwise ranking methods [38,40] which learn a scoring function for each object pair. These methods however come with a significant constraint in terms of computational cost—the number of pairs is quadratic with respect to the number of objects to be ranked. To circumvent this restriction, the objects are processed together in compatible groups, i.e. queries in the case of web searches—and proteins in our case. By grouping and comparing the peptides only within a respective protein, we achieve two additional aims: we avoid the need to correct for different protein abundances which can span several orders of magnitude within a biological sample, and we capture precisely those characteristics that make some peptides within a protein better detectable than others, which serves our purpose of selecting the best-suited peptides per protein for LC–ESI–MS/MS based SRM experiments.

3.2. Selection of training set

The selection of the training set is a crucial step in machine learning. For the binary peptide detectability prediction problem, both observed and non-observed peptides should be represented in the training set to avoid biases and over-fitting in the later learning process. Ideally, there should also be no bias against specific protein classes such as transmembrane proteins (see below).

Until recently [16–18] no complete proteome expressed under specific conditions had been identified by discovery shotgun proteomics. However, only such studies offer a more precise and complete reference set allowing the extraction of the true labels, observed and unobserved, both at the protein and more importantly at the peptide level. Since proteomics data are context dependent, i.e. only a subset of all proteins will be expressed in a given tissue or under a specific condition, for most datasets there is no known complete reference set. In addition, the selection of the precursor ions for subsequent fragmentation is a semi-stochastic process, and measuring an identical sample several times will lead to new peptide identifications within a given protein which may not have been observed in previous runs [13,53,54]. Finally, the digestion and fractionation steps also have an influence. Accordingly, the distinction between observed and non-observed is not clear-cut and is expected to largely influence the accuracy of a predictor. However, for a rank-based classifier that evaluates the peptides within a protein, this is less of an issue than for predictors that provide a global numerical prediction value such as PeptideSieve [14], ESPPredictor [28] and CONSeQuence [29].

To account for these issues, we selected shotgun proteomics datasets from four model organisms where i) a significant coverage of the predicted proteome (between 50 and 60%) had been reported [33,55,56] and ii) repeat measurement of an identical sample had been carried out along with extensive fractionation at the peptide level, leading to a high number of observed peptides per protein (see Supplemental Information). Both of these are important parameters for an accurate prediction of proteotypic peptides, which can be readily and repeatedly observed in a mass spectrometer. Such datasets should enable us to assess whether the organism-specific differences of the amino acid frequencies that we observed both in in silico tryptic digests of the respective proteomes and in the experimental datasets (Fig. S2), would have an effect on the accuracy of a rank predictor.

Finally, we explored the selection of proteins to be included in the training set. In contrast to solutions like APEX [23] which considers the 100 most abundant proteins, we selected the top 500 proteins based on spectral count. This selection increased the percentage of proteins with predicted transmembrane (TM) domains for all organisms. Importantly, the top 500 proteins included more proteins in the medium abundance range, and more proteins for which between 40 and 50% of the peptides within the visible range were observed experimentally. Among the top 100 proteins this average percentage of observed peptides was about 15% higher (data not shown). The larger training set thus supports our hypothesis that proteins for which a certain number of peptides have been observed and at the same time a similar number of peptides not observed (i.e. around 50% of the theoretically observable peptides) likely represent good examples to optimally capture the difference in peptide detectability.

3.3. Feature selection for rank prediction

To build a robust model that can be computed within a reasonable time frame, a subset of the numerical peptide features was selected that are most relevant for rank-based prediction while containing as little redundancy as possible. While the amino acid frequencies (20) and the general peptide features (10) constitute a space that exhibits only limited redundancy (data not shown), this is not true for the physicochemical properties (544). The extension of a cluster analysis (see Methods) to data from the other organisms indicated that the grouping of these parameters was very similar (data not shown). Based on the explorative visualization of their similarities (see Methods, Fig. S3A), we decided to cluster the physicochemical properties into twelve groups, which are represented as a minimum spanning tree in Fig. S3B, similar to [37]. Following procedures which have been used in feature selection for ranking [57], we selected the parameter with the highest nDCG@5 (see Methods, Table S7B) as representative physicochemical parameter for each of these twelve clusters. All subsequent analyses were carried out with the subset of 42 selected peptide features.

3.4. Shared peptides in rank prediction

The detectability of a peptide depends on its abundance in the sample, as well as many other factors including, for example the complexity of the sample, the efficiency of the tryptic digest, the ionization efficiency in the mass spectrometer, etc. shared peptides, i.e. identical peptides that can be derived from several distinct proteins, complicate both protein inference [58,59] and protein quantitation [60]. The abundance of shared peptides will be the joint contribution of all the proteins from which they are derived. While shared peptides can be used as alignment anchors to compare peptides across proteins, as described by Dost and colleagues [60], in the context of peptide predictability calculation, we excluded shared peptides from the training step.

Shared peptides can be further differentiated depending on whether the proteins they imply are encoded by one or more distinct gene models [21,61]. We only considered unambiguous peptides of evidence class 1a, which imply one specific protein isoform [21] (Table S6). For all other classes of shared peptides, we cannot unambiguously assign the respective spectral counts to a particular protein isoform (classes 1b, 2a and 2b) or to a particular gene model (classes 3a and 3b). This concept is illustrated for a class 3b peptide in the Supplementary Information. It is important to note however, that shared peptides will be ranked in the testing step and that a selection of best-suited PTPs ideally includes strategies to exclude certain types of shared peptides from assay development (see Discussion).

3.5. *Empirical results on four organisms*

We performed a thorough empirical evaluation of the proposed rank learning approach in order to compare its performance with existing PTP predictors for LC–ESI–MS/MS data, including PeptideSieve [14], ESPPredictor [28], and CONSeQuence [29]. Since we included the log-transformed spectral counts as their relevance in the evaluation process (see Methods), we could use the nDCG as evaluation measure. It indicates how relevant e.g. the top 5 peptides are with respect to the experimentally observed spectral counts. An ideal nDCG@5 score of 1 would be assigned if all 5 top peptides are ranked correctly according to the observed spectral count. Conversely, an nDCG@5 value of zero would imply that the top five predicted peptides were not experimentally observed at all. This approach fits very well with the aims of targeted proteomics, where a limited number of peptides, often three to five, are interrogated for a given protein in SRM/MRM assays.

As a first analysis, we compared the accuracy of the predictions of all rank learning algorithms that were available in a recent release of the RankLib library (version 2.1; see Methods). Among the eight algorithms, five gave consistently better results than a random ranking (Fig. 2). Among these, the LambdaMART algorithm [47] proved to provide the best performance. This result could be obtained irrespective of the organism from which the training and test datasets were derived (Fig. 2). The next-best performing rank learning algorithms were RandomForests [32] and RankBoost [44]. The LambdaMart algorithm was thus chosen for the machine learning part of PeptideRank.

In a next step, we compared PeptideRank against the previously released PTP predictors PeptideSieve, ESPPredictor and CONSeQuence. Since all of these have been trained on yeast, we trained our ranking approach on the top 500 proteins of the yeast dataset (training set) and then predicted the peptide rankings for the 566 remaining identified proteins (test set). Again, the rank-based approach was able to outperform the previously published PTP predictors (Fig. 3): the orange box plots in Fig. 3A show that PeptideRank achieved a higher median nDCG@5 score than the other predictors. This result was also reflected by the orange cumulative distribution line in Fig. 3B, where better performance (i.e. more accurate prediction) was indicated by the line being closer to an nDCG@5 value of 1.

As a final step, we performed cross-organism evaluations by using varying combinations of training set–testing set pairs. This allowed us to evaluate whether organism-specific differences in the amino acid frequencies, which likely result in differences in peptide properties, had an effect on the accuracy of the rank predictor. Panels a–d of Fig. 4 show the results of comparisons of the ranking approach using PeptideSieve MUDPIT ESI, CONSeQuence ANN and PeptideRank, when trained with data from different organisms. In this comparison, CONSeQuence consistently outperformed PeptideSieve, as has been described previously [29]. However, PeptideRank performed roughly as well as CONSeQuence, but consistently performs better when trained on suitable previous shotgun proteomics data from that organism. While the extent of the benefit varies, the orange straight line representing the cumulative distribution when training and testing with organism-specific data was always higher and thus provides more accurate results in the case of *Leptospira* (Fig. 4, upper left panel), yeast (upper right panel), *Drosophila* (lower left panel) and *Bartonella* (lower right panel). Furthermore, it can be appreciated that for yeast (upper right panel), the next-best results were obtained when testing with data from the yeast PeptideAtlas dataset (*Yeast PA*, dashed orange line), again underlining the species-specific benefit. Of particular note is the observation that the highest overall accuracy is achieved for the complete proteome dataset of *Bartonella* (Fig. 4, lower right panel), where 82.3% of the proteins exhibited an nDCG@5 value of 0.5 or more. This percentage is much lower for the other three organisms with 64.1% (*Leptospira*), 53.2% (yeast) and 57.3% for *Drosophila*.

Our results thus indicate that a significant increase in rank prediction accuracy can be achieved when training PeptideRank on suitable previous LC–ESI–MS/MS experiments from the organism for which one wants to predict PTPs for SRM-based targeted proteomics.

3.6. *Peptide rank prediction accuracy for different protein classes*

From the plots in Figs. 3 and 4 it became immediately apparent that the PeptideRank prediction of the top 5 peptides worked well for a large number of proteins, but that there are also a significant number of proteins for which there is a lower overlap with the predicted ranks. We assessed physicochemical protein parameters, the number of peptides within the visible range, and additional factors like gene/protein expression strength and post-translational modification state to explore whether we could identify factors that negatively correlate with the rank prediction accuracy, and – wherever possible – to assess whether such factors were relevant across all four organisms.

One obvious candidate class is long proteins with many tryptic peptides in the visible range; for these a correct prediction of the 5 top ranking peptides is more difficult than for short proteins. In fact, we were able to identify an overall negative Spearman rank correlation between the number of peptides within the visible range and predicted rank accuracy

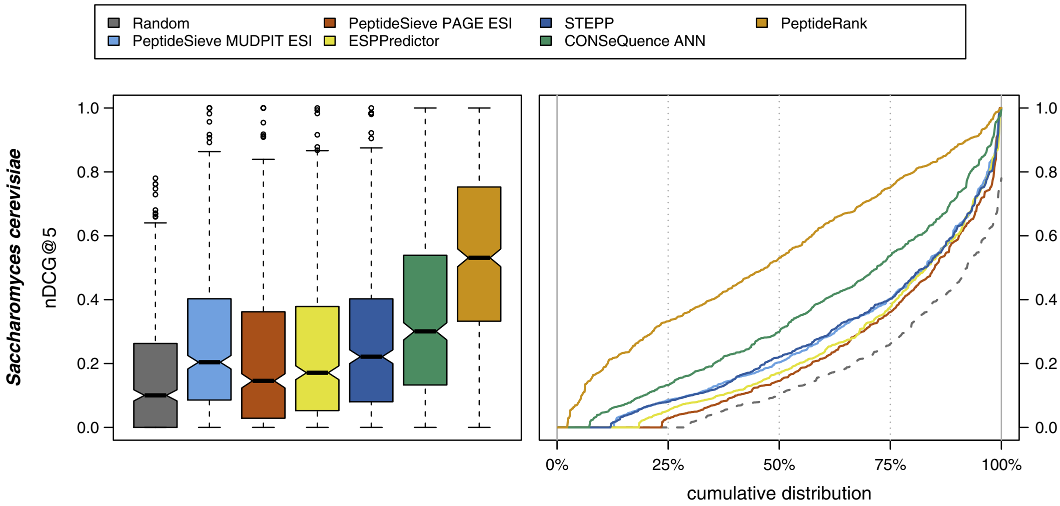

Fig. 3 – Comparison of the PeptideRank prediction results to that of other common PTP predictors for ESI–LC–MS/MS data. Results are both shown with box plots (left panel) and cumulative distributions (right panel) using the nDCG for the top 5 peptides for yeast. PeptideRank (orange) performed better on this dataset than the other approaches. See Fig. S5 for the comparison of rank prediction results on the datasets of all four organisms.

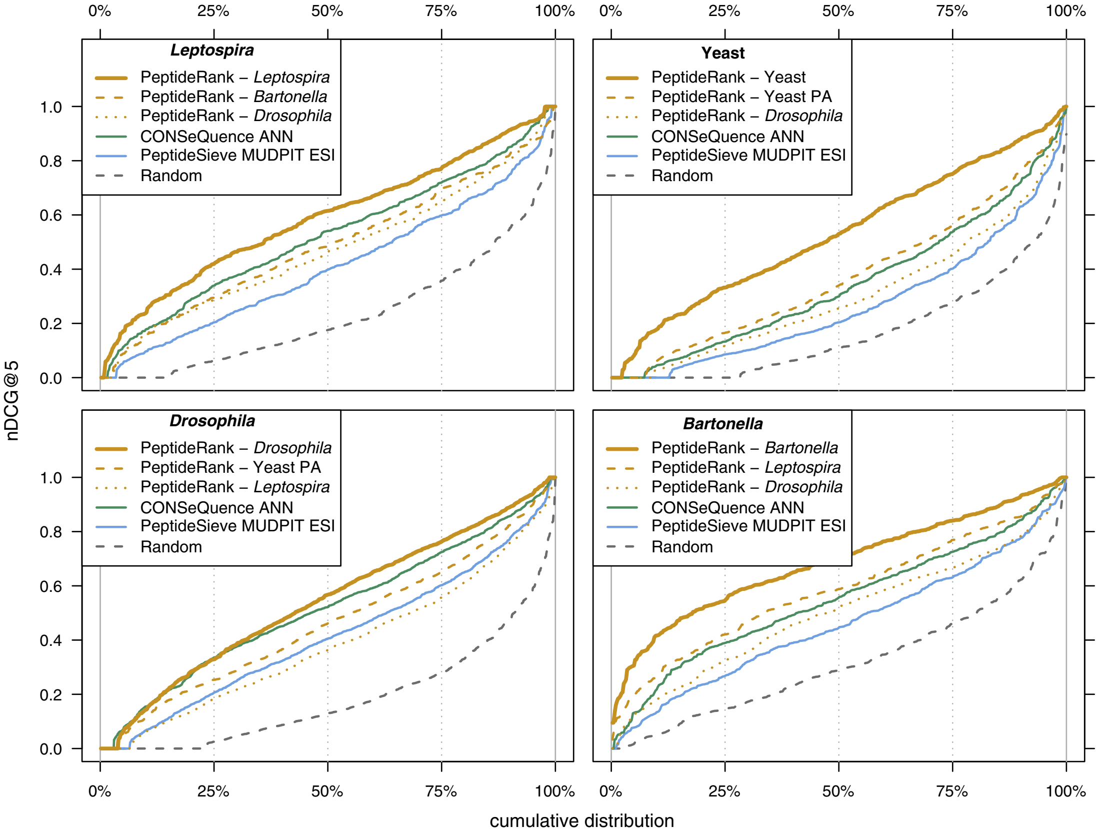

Fig. 4 – Comparison of PeptideRank results for training and testing on data from different organisms. The four panels show the results when we trained and tested with data from *Leptospira*, yeast, fly and *Bartonella*. Each panel shows the cumulative distribution of the nDCG values computed with PeptideSieve (blue), CONSeQUence (green) and PeptideRank (orange) when trained on data from different organisms. The ideal prediction line for each method would be a constant line at nDCG@5 = 1. The data show that we can obtain the best results (solid orange line) by using a within-organism rather than a cross-organism training–testing combination.

among the test proteins for all organisms (*Bartonella*: r = −0.431, n = 401; *Leptospira*: r = −0.409, n = 521; *Drosophila*: r = −0.410, n = 1353; Yeast: r = −0.329, n = 566). This negative correlation also became obvious when comparing the distribution of rank accuracies for proteins from the upper versus the lower quartile of protein lengths. Among the physicochemical parameters tested, protein length and number of peptides in the visible range showed the highest negative correlation (Fig. 5A). Conversely, very weak or no correlation was observed for other physicochemical parameters including isoelectric point (basic proteins with a high percentage of R and K residues will produce on average shorter peptides and might be hypothesized to have a different rank accuracy), and grand average hydrophobicity (membrane proteins with positive gravy values might be hypothesized to have a lower rank accuracy).

Next, we tested whether it was more difficult to accurately predict the top peptides for proteins whose genes were expressed at lower levels. For this question we could rely on the *B. henselae* dataset, for which transcriptomics data had been generated based on RNA extracted from matched samples [18]. However, when we compared the nDCG@5 rank accuracy versus the RPKM level of the encoding genes, we did not observe a notable correlation. In contrast, a strong positive correlation was observed at the spectral count level (average spectral count per visible peptide) for all organisms tested (*Bartonella*: r = 0.351; *Leptospira*: r = 0.478; *Drosophila*: r = 0.439; Yeast: r = 0.429) (Fig. 5B).

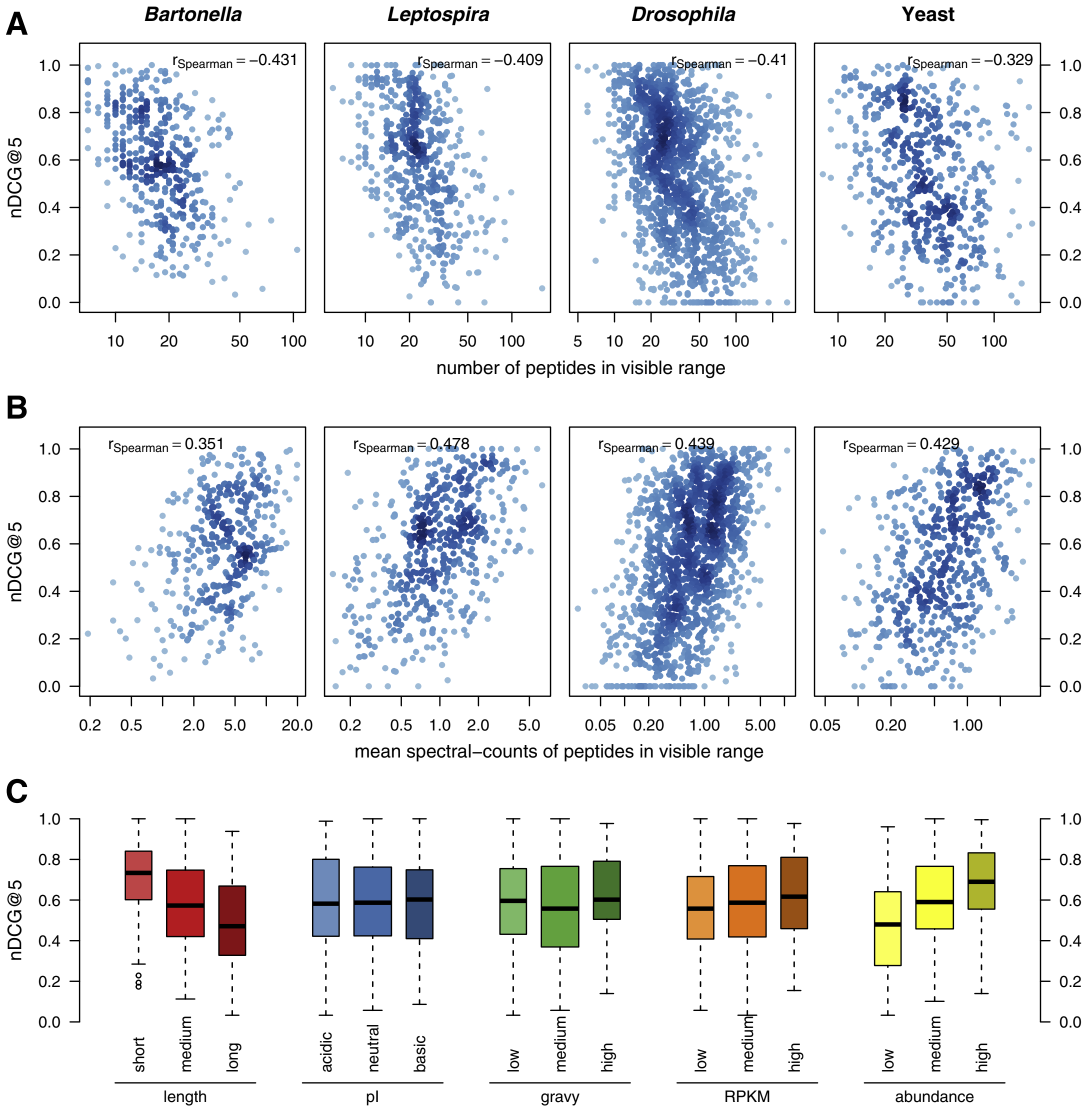

Fig. 5 – Parameters that correlate with the observed nDCG@5 accuracy. A. The number of peptides in the visible range and the protein length (not shown) display a negative correlation for all organisms tested. B. Conversely, the normalized protein abundance (mean spectral counts) correlates positively with the ranking accuracy. C. Boxplots are shown that compare the ranking accuracy for *Bartonella* proteins in the first, second plus third, and fourth quartile. Protein parameters analyzed include length (red), isoelectric point (pI, blue), grand average hydropathicity (gravy, green), and expression strength measured at the gene expression (RPKM, orange), or protein level (yellow).

The observation that the rank prediction accuracy tended to be higher for more abundantly expressed proteins was underscored by the analysis of the important protein class of transcription factors. Transcription factors are often expressed at low levels [62] and have unusual physicochemical properties (i.e. they are often very basic). Among the top 500 proteins of the *Drosophila* training set we only observed 4 of 755 transcription factors that were annotated in the *Drosophila* genome [63]. Conversely, the test dataset, which included 853 remaining proteins among the 1353 that were identified with 5 or more peptides, contained 57 transcription factors. These transcription factors displayed a lower average nDCG@5 prediction accuracy compared to all *Drosophila* proteins expressed in the dataset studied (data not shown). Importantly, however, the rank prediction accuracy of PeptideRank for the membrane proteome (i.e. proteins with predicted TM domains) was similar to the overall rank prediction accuracy for all *Bartonella* proteins (for an example of the PTP predictions for a *Bartonella* membrane protein see Fig. 6).

PTMs may have a potential influence on the rank accuracy prediction due to the fact that typically only few modifications can be considered in a database search. Therefore, heavily modified proteins might display a lower rank prediction accuracy. To test this hypothesis, we chose the UniProt annotations for yeast proteins, many of which were derived from manual SwissProt curations (see Methods). We counted all sequence-modifying annotations for the yeast test-set proteins and looked whether proteins with more annotated modifications would result in lower predicted ranking accuracy. However, no significant correlation was found. Since the overall number of PTM annotations was low, this may indicate that more data from large PTM studies under many different conditions would be required to potentially detect a signal.

4. Discussion

In this study we showed that a rank-based approach represents a valuable alternative to existing tools and can complement existing tools for prediction of the top proteotypic peptides for LC–ESI–MS/MS-based targeted proteomics measurements. Importantly, in contrast to other predictors like PeptideSieve,

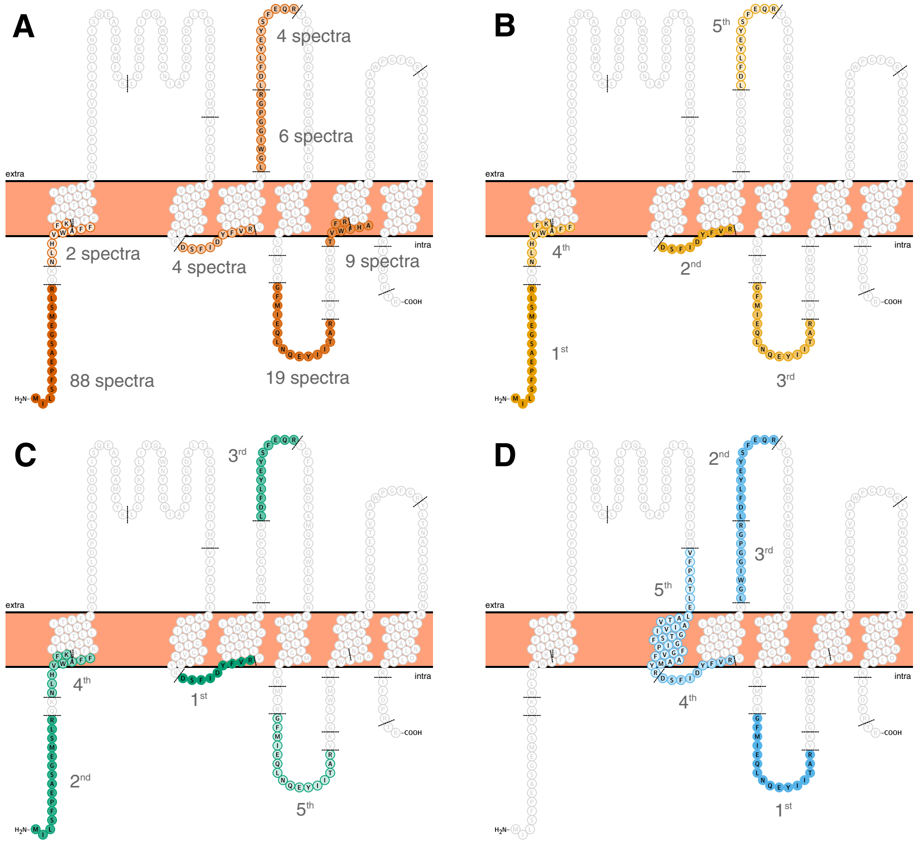

Fig. 6 – Example of a *Bartonella* membrane protein with high prediction accuracy. The Protter web application [65] was used to visualize experimental peptide evidence on top of the predicted topology of BH12170, an ABC transporter permease protein (panel A). The top 5 peptides predicted by PeptideRank, CONSeQuence (ANN method) and PeptideSieve (MUDPIT_ESI method) are shown in panels B, C and D, respectively. The rankings resulted in nDCG@5 values of 0.96 for PeptideRank, 0.65 for CONSeQuence, and 0.25 for PeptideSieve (for a high ranking performance, it is necessary to rank the two most detectable peptides up front). Importantly, the rank prediction accuracy of PeptideRank for the membrane proteome (i.e. proteins with predicted TM domains or predicted to be secreted) was similar to the overall rank prediction accuracy for all *Bartonella* proteins.

ESPPredictor, and CONSeQuence, our approach does not require the selection of positive and negative training sets, an inherently difficult and error-prone step.

The improved rank predictions are likely based on several factors in addition to overcoming the delicate issue of selecting positive and negative training sets: 1) we considered the spectral count information as a relevant parameter while most other approaches consider only binary outputs (observed, not observed); 2) in contrast to existing tools we constructed rank-based models for each protein—which model exactly the peptide selection process for targeted proteomics where prediction of the best detectable peptides for a particular protein is needed, not a global detectability score; 3) a careful selection of the training sets for each organism; 4) prediction in an organism-specific context.

To best capture the characteristics we want to predict a training set should contain a representative protein selection without noise (proteins with very low spectral counts which have a higher chance to include false positives), is not biased against specific categories such as membrane proteins, and contains many peptides observed with varying spectral counts. Combinations of many experiments including replicate analysis increase the sampling depth and are expected to allow to more accurately distinguish whether different peptide detectability versus sampling depth issues account for certain peptides to be observed or not.

Importantly, as illustrated in Fig. 4, the rank-based approach performed best when trained on LC–ESI–MS/MS data from the organism for which PTPs are to be predicted: this combination boosts the rank prediction accuracy and consistently returned better results than the other predictors that were all trained on yeast data. A similar benefit has been observed before in the context of predicting PTPs for accurate mass and time proteomics for three prokaryotic organisms [30]. This suggests that organism-specific differences in amino acid frequency and composition are relevant, and that a predictor that can capture these differences could represent a useful tool for the research community. In cases where shotgun proteomics data are not available for a specific organism, a large resource such as PeptideAtlas should be used to select the top proteins for the training set. Selecting data only from one study might not provide the desired accuracy.

While our rank-based approach performed well, there is still room for improvement: for a sizeable fraction of proteins neither the rank-based nor the other predictors performed well. Analysis of several protein parameters and other factors including the gene/protein expression level indicated that it was difficult to predict the top ranked PTPs correctly for long proteins with many peptides in the visible range, which however represents an inherent problem of a ranking approach. In contrast, the ranking was generally more accurate for more abundantly expressed proteins, in line with the underlying rationale of utilizing the information in repeatedly observed spectra, rather than a mere distinction in observed/not observed. As a consequence, transcription factors that are typically low abundant proteins are not well-represented in the training step of the ranking approach.

PTMs are also expected to affect the observed spectral count of peptides: peptides that potentially contain PTMs might be observed with low spectral counts because the search algorithms were not instructed to search for specific PTMs and such peptides might be classified as non-relevant although that might not be the case. On the other hand, this would indirectly put peptides with lower amount of PTMs at the top of the list, which should help to minimize standard deviations for the quantitation result for a protein based on several peptides in downstream targeted proteomics applications. In our analysis, the PTM status did not have a detectable effect. Since PTMs change in a context-dependent fashion (e.g. under different growth conditions, or developmental states), such an analysis requires more diverse large PTM studies.

Other factors that we did not specifically assess but that could also contribute to a lower rank prediction accuracy for certain proteins include the efficiency of the trypsin digest, i.e. are the respective peptides cleaved at all. CONSeQuence for example includes such a predictor. Also, sample complexity can change the observability of certain peptides in a probe.

Overall, we observed the best ranking predictions for *Bartonella*, i.e. the organism where we achieved the highest proteome coverage. There, the nDCG@5 values were consistently higher than for the other organisms, and the prediction accuracy was high for the important class of transmembrane proteins (Fig. 6). A comparison with PTPs predicted by PeptideSieve using the default threshold value (0.8) indicated that for 144 proteins (i.e. roughly 10% of the annotated proteome) we could experimentally identify PTPs which scored below the threshold. This implies that one should use lower cut-offs than the default provided by PeptideSieve and rely on the ranked list of peptides, in particular for proteins of interest with globally lower scores.

The prediction of peptide detectability is an important step to enable researchers to select the best-suited peptides for subsequent SRM experiments. To integrate the results of PeptideRank in typical downstream workflows, we suggest to further prioritize the list of ranked peptides by taking additional points into consideration (Fig. 7): i) while excluded from the training step, at this stage the peptide evidence class [21,61] represents a useful selection criterium: shared class 3b peptides, which can imply different proteins encoded by different gene models, should not be used for assay development. In contrast, and as requested by the Human Protein Detection and Quantitation (HPDQ) project [64] classes 2a and 2b peptides which imply a subset of protein isoforms encoded by one distinct gene model (2a), or all protein isoforms encoded by a gene model (2b), could be considered for specific questions. ii) information from predicted TM domains, signal peptides and post-translational modifications, as is available e.g. through the Protter tool [65] represents a very useful visual aid to further prioritize among predicted PTPs. iii) factors that affect the specificity and accuracy of the downstream SRM quantification step. These include information on potentially interfering transitions and other factors that have been described elsewhere [66–68]. With such a combined approach, one could e.g. use the visual representation of Protter with integrated PeptideClassifier information and highlight top ranked peptides from surface exposed proteins to score surfaceomes (Fig. 7). As demonstrated by the first description of a complete expressed membrane proteome by discovery shotgun proteomics [18], the entire set of membrane proteins (the surfaceome) – including those found expressed at the cell surface and those without

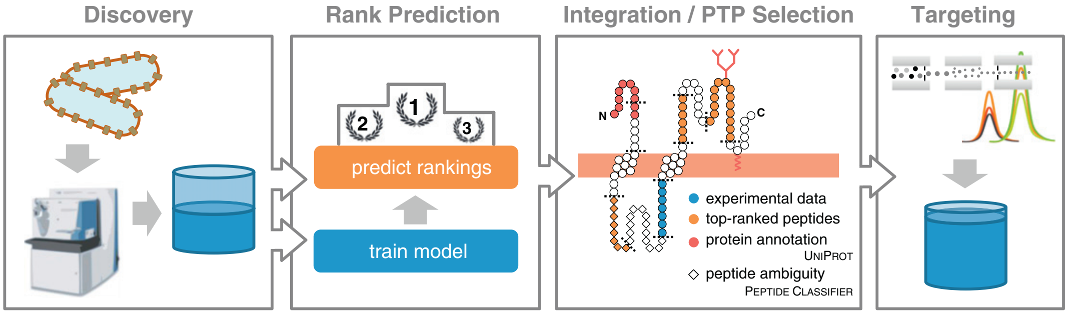

Fig. 7 – Typical case of application for the PeptideRank algorithm illustrating additional steps following the rank prediction. If experimental data from the organism of interest is available (panel Discovery), an organism-specific model can be trained; otherwise a suitable model is chosen among the library of available models to predict the ranking of best-suited PTPs (panel Rank Prediction; for simplicity the top 3 ranked peptides are shown). The output of PeptideRank can be visualized with the Protter web application [65] (panel Integration, PTP Selection) which can seamlessly integrate protein annotations from UniProt, previous experimental data, and peptide evidence classes from PeptideClassifier, all important additional criteria for the final selection of PTPs for a set of proteins that are to be used in subsequent SRM experiments (panel Targeting). Criteria relevant for SRM are not shown here.

expression evidence in these specific conditions but predicted to be localized there [69] – could now be scored under a variety of conditions, with important implications for validation in clinical cohorts [70] or for the discovery of novel targets to combat the resurgence of infectious disease [18].

A correct prediction of the top peptides and subsequent prioritization based on additional considerations is furthermore expected to lower the downstream labor and costs associated with selecting wrong or weak target peptides for subsequent assay development for proteome wide studies, which can exceed 100.000 Euro.

To allow researchers to make use of PeptideRank, we here release the rank models for the four organisms (http://wlab.ethz.ch/peptiderank/). We intend to add further organism-specific models for several important organisms in the near future.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgments

Ermir Qeli was partially funded by a grant from the University of Zurich's Research Priority Project in Systems Biology/Functional Genomics. CHA acknowledges support from the Swiss National Science Foundation SNSF under grant 31003A_130723, and the Swiss Initiative for Systems Biology SystemsX.ch under grant IPP 2011/121. SG was supported by a Marie-Heim Vögtlin fellowship (PMPDP3_122836) and EB acknowledges support by the 7th Framework Programme of the European Union (Contract no. 262067-PRIME-XS). The Functional Genomics Center Zurich and its managing director Ralph Schlapbach are gratefully acknowledged for access to the mass spectrometry facilities and continued support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2014.05.011.

REFERENCES

[1] Kuster B, Schirle M, Mallick P, Aebersold R. Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 2005;6:577–83.
[2] Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R. Generating and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol 2010;11:789–801.
[3] Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 2012;9:555–66.
[4] Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol 2010;28:710–21.
[5] Picotti P, Rinner O, Stallmach R, Dautel F, Farrah T, Domon B, et al. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nat Methods 2010;7:43–6.
[6] Surinova S, Huttenhain R, Chang CY, Espona L, Vitek O, Aebersold R. Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. Nat Protoc 2013;8:1602–19.
[7] Addona TA, Abbatiello SE, Schilling B, Skates SJ, Mani DR, Bunk DM, et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. Nat Biotechnol 2009;27:633–41.
[8] Huttenhain R, Malmstrom J, Picotti P, Aebersold R. Perspectives of targeted mass spectrometry for protein biomarker verification. Curr Opin Chem Biol 2009;13:518–25.
[9] Barnidge DR, Dratz EA, Martin T, Bonilla LE, Moran LB, Lindall A. Absolute quantification of the G protein-coupled receptor rhodopsin by LC/MS/MS using proteolysis product peptides and synthetic peptide standards. Anal Chem 2003;75:445–51.

[10] Anderson L, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 2006;5:573–88.

[11] Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104:5860–5.

[12] Lange V, Malmstrom JA, Didion J, King NL, Johansson BP, Schafer J, et al. Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. Mol Cell Proteomics 2008;7:1489–500.

[13] Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res 2010;9:761–76.

[14] Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 2007;25:125–31.

[15] Picotti P, Clement-Ziza M, Lam H, Campbell DS, Schmidt A, Deutsch EW, et al. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature 2013;494:266–70.

[16] de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 2008;455:1251–4.

[17] Peng M, Taouatas N, Cappadona S, van Breukelen B, Mohammed S, Scholten A, et al. Protease bias in absolute protein quantitation. Nat Methods 2012;9:524–5.

[18] Omasits U, Quebatte M, Stekhoven DJ, Fortes C, Roschitzki B, Robinson MD, et al. Directed shotgun proteomics guided by saturated RNA-Seq identifies a complete expressed prokaryotic proteome. Genome Res 2013;23:1916–27.

[19] Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep 2008;9:429–34.

[20] Vizcaíno JA, Reisinger F, Côté R, Martens L. PRIDE: data submission and analysis. Curr Protoc Protein Science 2010;60:25(4):25.4.1–25.4.11.

[21] Qeli E, Ahrens CH. PeptideClassifier for protein inference and targeted quantitative proteomics. Nat Biotechnol 2010;28:647–50.

[22] Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, et al. A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics 2006;22:e481–8.

[23] Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 2007;25:117–24.

[24] Alves P, Arnold RJ, Novotny MV, Radivojac P, Reilly JP, Tang H. Advancement in protein inference from shotgun proteomics using peptide detectability. Pac Symp Biocomput 2007:409–20.

[25] Li YF, Arnold RJ, Tang H, Radivojac P. The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. J Proteome Res 2010;9:6288–97.

[26] Huang T, Gong H, Yang C, He Z. ProteinLasso: a Lasso regression approach to protein inference problem in shotgun proteomics. Comput Biol Chem 2013;43:46–54.

[27] Wedge CW, Gaskell SJ, Hubbard SJ, Kell DB, Lau KW, Eyers C. Peptide detectability following ESI mass spectrometry: prediction using genetic programming. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. London, England: ACM; 2007.

[28] Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol 2009;27:190–8.

[29] Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ, Hubbard SJ. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. Mol Cell Proteomics 2011;10 [M110 003384].

[30] Webb-Robertson BJ, Cannon WR, Oehmen CS, Shah AR, Gurumoorthi V, Lipton MS, et al. A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. Bioinformatics 2010;26:1677–83.

[31] Lipton MS, Pasa-Tolic L, Anderson GA, Anderson DJ, Auberry DL, Battista JR, et al. Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. Proc Natl Acad Sci U S A 2002;99:11049–54.

[32] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[33] Balgley BM, Wang W, Song T, Fang X, Yang L, Lee CS. Evaluation of confidence and reproducibility in quantitative proteomics performed by a capillary isoelectric focusing-based proteomic platform coupled with a spectral counting approach. Electrophoresis 2008;29:3047–54.

[34] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.

[35] Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res 2010;9:1323–9.

[36] Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374.

[37] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202–5.

[38] Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. In: Smola B, Schoelkopf, Schuurmans, editors. Advances in Large Margin Classifiers. Cambridge, MA: MIT Press; 2000. p. 115–32.

[39] Yu H, Kim S. SVM tutorial—classification, regression and ranking. Handbook of Natural Computing. Springer; 2012 479–506.

[40] Joachims T. Optimizing search engines using clickthrough data. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2002. p. 133–42.

[41] Chapelle O, Keerthi SS. Efficient algorithms for ranking with SVMs. Inf Retr 2010;13:201–15.

[42] Friedman JH. Greedy function approximation: a gradient boosting machine. Technical Report, IMS Reitz Lecture, Stanford; 1999 [see also Annals of Statistics, 2001].

[43] Burges CJC, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, et al. Learning to rank using gradient descent. Proc. of ICML; 2005. p. 86–96.

[44] Freund Y, Iyer R, Schapire R, Singer Y. An efficient boosting algorithm for combining preferences. J Mach Learn Res 2003;4:933–69.

[45] Xu J, Li H. AdaRank: a boosting algorithm for information retrieval. Proc. of SIGIR; 2007. p. 391–8.

[46] Metzler D, Croft WB. Linear feature-based models for information retrieval. Inf Retr 2007;10:257–74.

[47] Wu Q, Burges CJC, Svore K, Gao J. Adapting boosting for information retrieval measures. J Inf Retr 2007;13:254–70.

[48] Cao Z, Qin T, Liu TY, Tsai M, Li H. Learning to Rank: From Pairwise Approach to Listwise Approach. ICML2007; 2007.

[49] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 2002;20:422–46.

[50] Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15:72–101.

[51] Kendall MG. A new measure of rank correlation. Biometrika 1938;30:81–9.

[52] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8.

[53] Schmidt A, Gehlenborg N, Bodenmiller B, Mueller LN, Campbell D, Mueller M, et al. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. Mol Cell Proteomics 2008;7:2138–50.

[54] Kristensen DB, Brond JC, Nielsen PA, Andersen JR, Sorensen OT, Jorgensen V, et al. Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. Mol Cell Proteomics 2004;3:1023–38.

[55] Malmstrom J, Beck M, Schmidt A, Lange V, Deutsch EW, Aebersold R. Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. Nature 2009;460:762–5.

[56] Loevenich SN, Brunner E, King NL, Deutsch EW, Stein SE, Aebersold R, et al. The *Drosophila melanogaster* PeptideAtlas facilitates the use of peptide data for improved fly proteomics and genome annotation. BMC Bioinformatics 2009;10:59.

[57] Geng X, Liu T, Qin T, Li H. Feature Selection for Ranking; 2007 407–14.

[58] Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 2005;4:1419–40.

[59] Gerster S, Qeli E, Ahrens CH, Bühlmann P. Protein and gene model inference based on statistical modeling in k-partite graphs. Proc Natl Acad Sci U S A 2010;107:12101–6.

[60] Dost B, Bandeira N, Li X, Shen Z, Briggs S, Bafna V. Shared Peptides in Mass Spectrometry Based Protein Quantification. RECOMB2009; 2009 356–71.

[61] Grobei MA, Qeli E, Brunner E, Rehrauer H, Zhang R, Roschitzki B, et al. Deterministic protein inference for shotgun proteomics data provides new insights into *Arabidopsis* pollen development and function. Genome Res 2009;19:1786–800.

[62] Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature 2003;425:737–41.

[63] Hens K, Feuz JD, Isakova A, Iagovitina A, Massouras A, Bryois J, et al. Automated protein-DNA interaction screening of Drosophila regulatory elements. Nat Methods 2011;8:1065–70.

[64] Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, Patterson SD, et al. A human proteome detection and quantitation project. Mol Cell Proteomics 2009;8:883–6.

[65] Omasits U, Ahrens CH, Muller S, Wollscheid B. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics 2014;30:884–6.

[66] Lange V, Picotti P, Domon B, Aebersold R. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol Syst Biol 2008;4:222.

[67] Abbatiello SE, Mani DR, Keshishian H, Carr SA. Automated detection of inaccurate and imprecise transitions in peptide quantification by multiple reaction monitoring mass spectrometry. Clin Chem 2010;56:291–305.

[68] Rost H, Malmstrom L, Aebersold R. A computational tool to detect and avoid redundancy in selected reaction monitoring. Mol Cell Proteomics 2012;11:540–9.

[69] Stekhoven DJ, Omasits U, Quebatte M, Dehio C, Ahrens CH. Proteome-wide identification of predominant subcellular protein localizations in a bacterial model organism. J Proteomics 2014:123–37.

[70] Huttenhain R, Surinova S, Ossola R, Sun Z, Campbell D, Cerciello F, et al. N-glycoprotein SRMAtlas: a resource of mass spectrometric assays for N-glycosites enabling consistent and multiplexed protein quantification for clinical applications. Mol Cell Proteomics 2013;12:1005–16.