Fragmentation-free LC-MS can identify hundreds of proteins.

Journal:	*PROTEOMICS*
Manuscript ID:	pmic.200900765.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	10-Jul-2010
Complete List of Authors:	Bochet, Pascal; Institut Pasteur, System Biology; CNRS, U2171 Rügheimer, Frank; Institut Pasteur, System Biology Guina, Tina; Univ. of Washington, Dept of Pediatrics Brooks, Peter; CNRS, UMR7592; Université Denis Diderot, Institut Jacques Monod Goodlett, David; Univ. of Washington, Dept of Medicinal Chemistry Clote, Peter; Boston College, Biology Dept; Ecole Polytechnique, LIX Schwikowski, Benno; Institut Pasteur, Laboratory for System Biology, Department Genomes and Genetics; CNRS, URA2171
Key Words:	Protein automated identification, Mass spectrometry, Peptide mass fingerprinting, Liquid chromatography-tandem mass spectrometry

Fragmentation-free LC-MS can identify hundreds of proteins.

Pascal Bochet[1,2], Frank Rügheimer[1,2], Tina Guina[3], Peter Brooks[4,5], David Goodlett[6], Peter Clote[7,8,9], Benno Schwikowski[1,2].

1 Institut Pasteur, Laboratoire de Biologie Systémique, Dept Génomes et Génétique, Paris, France
2 CNRS, URA2171, Paris, France
3 University of Washington, Dept of Pediatrics, Seattle, WA, USA
4 CNRS, UMR7592, Institut Jacques Monod, Paris, France
5 Université Paris 7, Paris, France
6 University of Washington, Dept of Medicinal Chemistry, Seattle, WA, USA
7 Biology Dept, Boston College, Boston, MA, USA
8 Ecole Polytechnique, LIX, Palaiseau, France
9 LRI, Université Paris XI,Orsay, France

Authors for correspondence: Pascal Bochet, Benno Schwikowski

Address:Laboratoire de Biologie Systémique, Institut Pasteur, 25, rue du Docteur Roux, 75015 Paris, France
email: pascal.bochet@pasteur.fr
Fax: 33 (0)1.40.61.37.04

Keywords

Automated protein identification, Mass spectrometry, Peptide mass fingerprinting, Liquid chromatography-tandem mass spectrometry.

Length: 8122 words

Abstract

One of the most common approaches for large-scale protein identification is liquid chromatography, followed by mass spectrometry (LC-MS). If more than a few proteins are to be identified, the additional fragmentation of individual peptides has so far been considered as indispensable, and thus, the associated costs, in terms of instrument time and infrastructure, as unavoidable. Here, we present evidence to the contrary. Using a combination of [1] highly accurate and precise mass measurements,[2] modern retention time prediction, and [3] a robust scoring algorithm, we were able to identify 257 proteins of Francisella tularensis from a single LC-MS experiment in a fragmentation-free approach (i.e., without experimental fragmentation spectra). This number amounts to 59% of the number of proteins identified in a standard fragmentation-based approach, when executed with the same false discovery rate. Independent evidence supports at least 27 of a set of 31 proteins that were identified only in the fragmentation-free approach. Our results suggest that additional developments in retention time prediction, measurement technology, and scoring algorithms may render fragmentation-free approaches an interesting complement or an alternative to fragmentation-based approaches.

1 Introduction

1.1 Protein identification by HPLC-MS

One of the key goals of proteomics experiments is the identification of proteins from samples of unknown composition. Thanks to continuous efforts in genome sequencing, genome and protein sequence information are now available for an increasing number of species [1]. For those species, the identification of proteins is accomplished by matching the results of Mass Spectrometry (MS) measurements against a sequence database of candidate proteins.

When only few protein are in the sample, a well-established procedure, Peptide Mass Fingerprinting (PMF), can be applied for protein identification [2, 3, 4, 5]. The protein is digested by an enzyme with known substrate specificity, the masses of the resulting peptides are measured by MS and the relevant sequence database is searched for proteins which generate a corresponding set of masses after enzymatic cleavage. Unfortunately, for a larger number of different proteins in a sample, the combinatorial explosion of possible mass combinations prevents reliable identification by PMF.

Additional information can be obtained by further fragmenting peptide ions isolated from the first MS analysis. The fragment masses are commonly determined in second MS steps interleaved with the first MS. The series of observed fragments is compared to predictions made from the known sequences and the most likely peptide candidates are selected [6]. One problem with this strategy is the high number of peptide ions generated by the first MS, which precludes their exhaustive analysis by fragmentation. This is partly overcome by separating the peptides prior to MS, usually on a HPLC column [7]. Most often, Reverse Phase columns are used, which separate the peptides mainly on the basis of hydrophobic interactions with the column [8]. Although HPLC separation helps alleviate the problem, not all peptides can be

analyzed and identified by fragmentation.

Apart from spreading peptides over the time domain, the preliminary HPLC analysis also provides for each of them a retention time (RT). This readily available, additional information has so far rarely been exploited for identification. Here we demonstrate that, together with highly accurate peptide mass measurements, the additional retention time information enables the identification of hundreds of proteins even without the use of MS/MS fragmentation.

1.2 Using retention time and accurate peptide mass for identification

Several previous efforts have used HPLC and peptide retention time for the identification of proteins. We will briefly discuss some of them.

1.2.1 Theoretical justification

A recent theoretical study by Norbeck *et al.* [9] has given insight into the precision of mass and retention time determination required for fragmentation-free identification of tryptic peptides. The authors evaluated to what extent the determination of mass and HPLC retention time allows unique identification of tryptic peptides. In particular, they studied the number of uniquely identifiable peptides as a function of the precision of mass and retention time and the size of the relevant proteome. The theoretical study showed that the addition of the HPLC retention time strongly enhances the identification power of MS. For instance, for a peptide mass of around 2250 Da, in the bacterium *Deinococcus radiodurans* (3167 proteins), approximately 20% of the peptides could be uniquely identified at a mass error rate of 5 parts per million (ppm). Combining the mass measurement with the HPLC retention time

(determined with a 1% error rate) increases the proportion of unique peptides from 20% to 90%. For a more complex proteome (*Homo sapiens*) the proportion of unique peptides would be 2% at 5 ppm, and would increase to 50% with the additional use of retention time prediction with 1% error. The study of Norbeck *et al.* demonstrates that taking into account the retention time of the peptides when searching for matches in the database can drastically reduce the number of candidate peptides. It does not, however, address practical implemetation aspects of this idea.

1.2.2 Determining peptide retention time

HPLC peptide retention times can be used for peptide identification by comparing experimentally obtained retention times with retention times that have been predicted, or previously measured.

Measuring retention time

A straightforward approach to obtain the retention times of a small number of peptides in a given HPLC separation system is to measure them: Synthetic peptides with the desired sequences are injected and detected when they exit from the column (see for instance [10]). However, this approach is limited to peptides that can be synthesized or highly purified and can thus not provide the hundreds or thousands of retention times required for protein identification from complex mixtures. With the advent of LC-MS/MS methods, the retention times of peptides selected for analysis by fragmentation after HPLC became available in very large amounts. These retention times can then be used directly or serve as training examples for diverse retention time prediction algorithms. But to account for unavoidable experimental variability, the retention times need to be normalized. For this purpose Strittmatter *et al.* [11], for instance, obtained data from individual runs and normalized the retention times by affine transformations optimized over the whole dataset. Still, the intrinsic

variability of HPLC places a limit on the precision that can be obtained for retention time.

Predicting retention time

An alternative to measuring peptide retention times is to predict them. Methods to predict peptide retention times from their primary sequences have been extensively reviewed recently by Baczek and Kaliszan [8].

One early approach was based on data from few injected synthetic peptides [10]. This work was limited by the amount of data available at the time, but introduced the idea that the hydrophobic interaction of a peptide with the matrix of the column could be predicted from the amino-acid composition by a linear model.

Later, other work relied on the idea of such a linear model, either using data from a rather small dataset of synthetic peptides [12] or on naturally occurring tryptic peptides [13]. Other peptides properties (e.g.its length [14]) have also been used with the linear model.

Other approaches have used artificial neural networks [15, 16] on a very large number of peptides (345000 peptides) or support vector machines [17, 18] on a much smaller dataset to obtain predictive models.

Krokhin's SSRCalc tool [19], develops the idea of a linear model further by employing different hydrophobicity coefficients for the amino acids close to the termini of the peptide, and by taking into account the length of the peptide, its likely secondary structure and effects between neighboring aminoacids. Contrary to many other models, SSRCalc does not require experiment or platform-specific training of the model. For each peptide, SSRCalc provides a hydrophobicity index which can be converted into a predicted retention time by an affine transformation.

1.2.3 Using retention time for peptide/protein identification

Regardless of how retention times are determined, several strategies have been

developed for their use in peptide and protein identification.

The Accurate Mass and Time tag (AMT) strategy, introduced by the group of R.Smith [11], uses the retention times of the peptides recorded in MS/MS experiments. Simultaneous matching of the mass and the retention times are then used to identify the peptides in a specific database and, subsequently, the proteins. The advantages of this method are high proteome coverage (at least for small proteomes like *Deinococcus radiodurans*) and short analysis time (3 hours, i.e. the typical duration of a HPLC-MS run). But this strategy requires that the retention times of thousands of peptides are recorded before the identification phase. Strittmatter *et al.* [11], for instance, recorded the retention times of more than 12000 peptides from 1067 runs of MS/MS experiments.

Instead of using previously measured retention times, only available for peptides characterized in previous experiments, several authors have used retention times predicted from the primary peptide sequence. Retention time predictions are generally available for every peptide, but the accuracy of this prediction will naturally be limited by the underlying model.

Palmblad *et al.* [13] used their composition-based, linear, retention time predictor. The deviation from predicted mass and retention time is combined into a Euclidean distance after separate normalization in each dimension. The quality of the peptide match and the identification is estimated from the value of this distance. In later work the mass and retention time determinations were combined using a likelihood-based method [20].

Similarly, using their Support Vector Machine predictor, Pfeifer *et al.* [17, 18] added to the fragmentation-based detection a filter based on the deviation of the predicted retention times from the observed and normalized values. In this approach, a potential peptide match detected on the basis of fragmentation spectra is excluded if the deviation between the observed and predicted peptide retention times is above a

given threshold. The threshold value itself is determined via a parametrized linear model that allows the acceptable absolute deviation to increase with normalized observed retention time.

Krokhin *et al.* [19] used their variant of the linear predictor in two different methods. MART analysis is based on fragmentation data from HPLC-MS/MS and uses predicted retention time to progressively eliminate peaks from the spectra, and thus speed up the computational identification process [21]. A second method is based only on HPLC-MS data. It first identifies the proteins compatible with all peptide masses detected in the MS data using the PMF search program Profound [5]. It then uses the extent of the correlation between predicted and measured retention times of the peptides belonging to each single protein to filter the results of PMF protein identification (see http://hs2.proteome.ca/SSRCalc/OKrokhin_ASMS07.pdf).

1.3 A novel fragmentation-free approach

Except for Krokhin's correlation method, all identification strategies discussed above depend upon previously measured fragmentation spectra, which are required either to directly establish peptide-specific retention times, or to train predictive models. Thus, one way to improve these methods is to rely on more reproducible HPLC or more accurate predictions. However the reproducibility of HPLC and the precision of predictions are limited, depend on the particular experimental setup, and cannot be improved beyond the intrinsic variability of the HPLC procedure.

Therefore, we designed a new approach which reduces the dependency on the retention time precision. Our approach relies on the retention *order* of the peptides rather than their retention time. By substituting the retention times by the retention order, we extract a more robust characterization of the retention process. Furthermore, we do not seek to identify isolated peptides, but rather groups of peptides coming

from the digest of the same protein: Each peptide will be identified from the experimental data only if other predicted peptides from the same candidate protein are also found in their predicted order. Unlike most of the methods described above, the proposed approach identifies proteins in one single step and not by first finding peptides and then proteins.

These ideas have been implemented in an alignment method (detailed in Methods **2.3.2**) at the core of our identification strategy based on what we have called the Ordered Peptide Match score (OPM score).

2 Material and methods

2.1 Principle

The principle of the new method is illustrated by Figure 1. For each protein in the relevant sequence database we first predict the peptides resulting from its enzymatic digest. In a second step, the HPLC retention time predictions for the peptides are obtained and the peptides are sorted by predicted retention time. In a third step we attempt to assign each predicted peptide for a given protein to a peak in a spectrum in the predicted order. We solved this alignment problem by dynamic programming.

Figure 1 near here

2.2 Experimental data

Experimental data were obtained from the species *Francisella tularensis subsp. novicida*, a Gram negative bacterium that is a causative agent of the human and

animal disease tularemia1.

2.2.1 Sample preparation

Bacteria were grown in rich medium, harvested by centrifugation during the exponential growth phase and were broken with ultrasound after several rounds of freezing in dry ice-ethanol and thawing in a water bath at 12°C. Membrane-bound and soluble proteins were separated by centrifugation. Aliquots of both preparations were solubilised in 6M urea, reduced with tris(2-carboxyethyl)phosphine (TCEP) and alkylated with iodoacetamide. Samples were then treated with DTT and digested with trypsin. Samples were desalted on C18 reverse phase columns and analyzed by HPLC-MS/MS. For more details see [23].

2.2.2 Sample analysis

Samples were loaded on a C-18 200Å pore size precolumn and separated on a C-18 100Å pore size microcolumn (75 µm × 11 cm, 5 µm)(Magic C-18 AQ beads, Michrom Bioresources, CA). Retention was achieved by a 60 min long linear gradient from 10% to 45% of solvent B (100% acetonitrile) in solvent A (0.1% formic acid, 5% acetonitrile) at a flow rate of 200nl/min. The spectra were obtained in a LTQ-FT-ICR-MS (Linear Trap Quadrupole-Fourier Transform-Ion Cyclotron Resonance Mass Spectrometer, Thermo Electron, San Jose, CA). Ions with m/z between 400 and 1800 were analyzed by FT-ICR. The most intense ions were fragmented by CID (Collision Induced Dissociation) and analyzed by LTQ. The nucleotide and protein sequences for *Francisella tularensis subsp. novicida* are publicly available from NCBI under the accession number NC_008601.1.

[1] *Francisella tularensis* is considered a potential bioterrorism agent, see [22].

2.3 Computational methods

2.3.1 Data reduction

In a first data analysis step, experimental data were preprocessed. The aim of this step was to reduce the spectra to mono-isotopic, mono-charged spectra. A method adapted from Matthiesen for his software VEMS [24] was applied, with the main difference that each spectrum was treated independently from the others.

Deisotoping was carried out as follows: a global maximum charge was chosen for the ions. The spectrum was searched for peaks with intensities above a set threshold, starting with low values of the *mass/charge* ratio. When found, such peaks were considered as putative mono-isotopic peaks and a series of matching isotopic peaks with *m/z* spacing of 1/(*charge*±*tolerance*) was searched. If found, such a series was removed from the spectrum, stored with the corresponding charge value, and the search was continued. When the higher end of the spectrum was reached, the process was repeated from the lower end with the next lower value of the ion charge. The resulting list of mono-isotopic, mono-charged peaks was stored and used for searching the predicted peptides from the protein digest.

To reduce the effect of experimental noise we only considered peaks with intensities in the top 24.5% quantile of each spectrum. In our experimental data this threshold was found to remove low intensity noise efficiently.

2.3.2 Computation of Ordered Peptide Match score (OPM)

The complete *in silico* digest was performed on all proteins predicted by the database. Only peptides at least four amino acids long were kept. For each protein these peptides were sorted according to their predicted order of retention from the HPLC column. For the results reported here we have used the hydrophobic predictions in TFA of the SSRCalc software, version 3.2, available on-line at

http://hs2.proteome.ca/SSRCalc/SSRCalc32.html [21, 25] .

For each protein, the algorithm described in Figure 2 computes the optimal alignment of the predicted peptides with the ions found in the experimental spectra. This maximizes the OPM score. The score optimized in the alignment is defined as follows: the mass of each predicted peptide is compared to the closest mass in each spectrum after deisotoping and decharging. If Δ, the difference between the two masses is less than a given threshold, the contribution of the match to the score is +1, otherwise a penalty is counted (-1). The alignment is then built in a progressive way, as shown by Figure 2. A value of 0.008 Da for the mass threshold has been chosen because 95% of the peptides detected by fragmentation are within this distance from the true mass (see below **2.4**).

For the particular choice of scoring function used here, the optimal OPM score is an affine transform of the number of predicted peptides matched to the experimental data. Therefore in the rest of the text we use this property to simplify the analysis and the figures by using the number of matching peptides.

Figure 2 near here

2.3.3 Interpretation of the Ordered Peptide Match scores

To evaluate the statistical significance of the OPM scores obtained for the proteins from the sequence databases, they were compared to the OPM scores of decoy proteins obtained by random permutation of the sequences of the real ones. Each of these decoy proteins had the same length, number of tryptic peptides and amino acid composition as its naturally occurring counterpart. In the database the number of proteins with the largest content of digest peptides was low. To obtain a sufficiently accurate estimate of the quantile limits for these large proteins (with more than 56

peptides), several permutations were performed for each of them.

Because the alignment scores depend on the number of tryptic peptides (see Section **3.1** and Figure 3), and in the absence of a substantial theoretical justification for the use of a particular parametric distribution model, we decided to apply quantile regression [26, 27] to describe the empirical distribution of scores for the decoy proteins. Briefly, for each value of the variable (here, the number of tryptic peptides in the protein) quantile regression uses a weighting function to construct a local score distribution for the randomized proteins. The score of each real protein was compared to that local distribution for the same number of tryptic peptides, and the corresponding quantile was reported as the significance of the match.

2.3.4 Estimation of the false discovery rate

The observed distribution of scores for real proteins is modeled as a mixture generated by drawing from two populations: the proteins absent from the sample and those present. Since neither the relative contributions of each group nor the actual distribution of scores for proteins in the sample are known, we applied the following approximation. For each quantile limit, we considered that all proteins with scores below the limit were absent from the sample. We assumed a score distribution for these absent proteins identical to that of the decoy proteins, and that no real protein effectively present in the sample had a score below the quantile limit. This amounts to a conservative choice of the largest possible number of absent proteins for a given total number of proteins. Based on this distribution we calculated the number of absent real proteins expected to have a score above the quantile limit (false positives). For each quantile limit this allowed us to estimate the false discovery rate (FDR) as the proportion of false positives among the total positives with scores above the limit.

2.4 Comparison with fragmentation

As the experimental data we used also included fragmentation spectra, we were able to compare the set of proteins identified by the OPM method with the set of proteins obtained by the standard fragmentation-based approach on the same experiments. Note that this analysis is likely to slightly underestimate the power of a pure OPM approach: If no time were devoted to the acquisition of fragmentation spectra, the additionally acquired MS spectra could be expected to improve the performance of the OPM method. Since this effect may not be very large, and since the comparison on the same dataset avoids statistical differences in all other respects, we chose to use the same dataset to evaluate both approaches. Fragmentation spectra were analyzed with the Mascot software [4]. Version 2.2.0 was used with the following parameters: sequence database: Francisella NC_008601 (NCBI), mass type: monoisotopic, fixed modifications: Carbamidomethyl (C), no variable modifications, enzyme: Trypsin, no missed cleavages, peak mass tolerance: 0.6, precursor mass tolerance: 1.2. For comparability, the false discovery rate (FDR) of Mascot was estimated by using a decoy database as described by [28] and the number of positive proteins retained from Mascot was set at 436 to obtain a FDR equal to that of the OPM method (5.7%).

3 Results

3.1 Correlation of peptide number and score

Figure 3 shows, for each *Francisella* protein and each decoy protein, the OPM scores as a function of the number of their tryptic peptides (data in Table 1). The basis of the analysis is the comparison between the scores of the proteins encoded by the genome of the bacteria (1719 real proteins) and the decoy database made of 2205 fictive

proteins.

For both groups of proteins, the score is, in a first approximation, proportional to the number of tryptic peptides in the protein. However, the slope of the associated least-squares regression line is higher for the real proteins (slope: 0.164 ± 0.003) than for the randomized ones (0.116 ± 0.002). This reflects the fact that the number of peptides matching with the experimental data is higher for the real proteins than for the decoy proteins. Also noteworthy is the larger deviation from linearity of scores for real proteins (R^2 = 0.567) than for the decoy proteins (R^2 = 0.673). This reflects the presence of real proteins with a large number of matching peptides and a high score absent from the decoy database.

Figure 3 near here

3.2 Detection of proteins and false discovery rate

In this section we make use of the OPM score described above to classify proteins as either present or absent from the experimental sample. Because the alignment scores depend on the number of peptides in the proteins we could not use it directly for this purpose. Therefore, we used a quantile regression method. By building a conditional, empirical, local distribution of the scores for each number of peptides in the proteins, quantile regression allows the comparison of scores for real proteins with those of corresponding randomized proteins. The proteins from the experiment shown in Figure 3 were classified as described in methods (section **2.3.3** and see quantile assignment in Table 1). For several quantile values, the number of proteins with score above the quantile threshold is indicated in Table 2, top. Table 2 shows that the number of detected proteins (real proteins in the table) is much larger than the corresponding number from randomized proteins (random proteins in the table). The

proportion of randomized proteins above each threshold (random proteins) is consistent with the corresponding quantile limit, thus showing the absence of any large systematic bias in the local quantile estimator.

Estimates of the FDR, the proportion of false positives among the positives, are also shown in Table 2. This estimated proportion decreases with the stringency of the detection from 34% for a quantile of 80% (721 proteins detected) to 5.7% for 99% (257 proteins detected).

Table 2 near here

3.3 Comparison of ordered and unordered peptide scores.

When compared to classical peptides mass fingerprinting (PMF), the OPM scoring approach introduced here additionally considers the degree of consistency between the orders of measured and predicted retention times. The rationale for the present work is the idea that the order of elution of peptides based on their predicted retention times will introduce additional discriminatory power for matching the peptides. To verify the impact of this additional element, we performed the same analysis as above, but without taking into account the elution order of the peptides. Specifically, we created a fictive data set, where all experimental peaks retained after decharging and deisotoping (cf. Section **2.3.1**) were present in every spectrum. The alignment procedure was then performed against this new data set, effectively neutralizing the order information present in the original data.

As with the ordered peptides, the scores are proportional to the number of tryptic peptides. As expected, the alignment scores obtained (unordered peptide match

(UPM) scores) were larger than the OPM scores computed before. They were higher for the real proteins (slope: 0.427 ± 0.005) than for the randomized ones (slope: 0.356 ± 0.003). The deviation from proportionality is less than that for the ordered peptides (R^2 = 0.807 for real proteins, R^2 = 0.858 for randomized). In summary, the score difference between the real and the randomized proteins, which is the basis for the protein recognition procedure introduced here, is much weaker without the order than with it.

This is also reflected in the detection of proteins as shown at the bottom of Table 2. In this table we can see that for every value of the quantile limit, the number of real proteins labelled as present in the sample is lower than that obtained with the ordered peptides match score. In addition, the estimated FDR is larger. This is further illustrated in Figure 4 A, where the FDR is plotted against the number of detected proteins: For almost all numbers of recognized proteins (obtained by varying the quantile limit) the estimated FDR is substantially smaller when based on the OPM score than the UPM score.

Figure 4 near here

3.4 Comparison of ordered peptide matched score with fragmentation based identification.

We determined how many of the proteins identified using the OPM method based on retention time and mass were also detected by Mascot on the basis of the fragmentation and MS/MS data from the same experiment. With our Mascot settings (see Methods), 436 proteins were identified in the experiment, out of 1719 annotated

in the genome of *F. tularensis* (see Table 3). Figure 4B shows the ROC curves obtained by comparing the identification results based on ordered or unordered peptide match scores to the proteins identified by Mascot. It can be seen that order information consistently improves the performance of the protein detector for values of the false positive rate (FPR) up to FPR = 0.4 (corresponding to a True Positive Rate of 0.91, at the low quantile limit of 0.66 (OPM), 0.654 (UPM)). This improvement is coupled to a significant reduction in the estimated false discovery rate (cf Figure 4A.). Table 2, Top also shows that the OPM method can detect proteins not identified from the fragmentation data: At the 99% quantile limit, 257 proteins were identified, of which 226 had been also detected by fragmentation and Mascot, revealing 31 newly detected proteins. Even when peptides of these 31 proteins had been identified they did not support the identification of the protein by Mascot under the settings used. For 6 of the 31 newly found proteins, no peptide had been identified by Mascot. For 8 of them, only one peptide had been identified by Mascot, and, for the remaining 17 proteins, at least 2 and up to 10 peptides had been identified. In addition, in a more extensive characterization of proteins expressed in *Francisella tulerensis* (1080 proteins detected), 27 of these 31 proteins have been detected independently by Rohmer *et al.* [29]. This overlap corresponds to a p-value of $3.4*10^{-6}$ under the assumption of a random choice of the 31 proteins.

Finally, the estimated FDR of 5.7 % of the OPM-based analysis corresponds to 15 false positives out of the 257 proteins detected. Therefore even in the unlikely case where all proteins common to Mascot are real positives, such that all 15 false positive are clustered in the 31 additional proteins, at least 16 of these would be real new hits.

4 Discussion

4.1 Summary of results

In the present paper we describe an alignment approach that exploits both peptide mass and retention order for the identification of proteins. The main result of this work is that a combination of highly accurate MS, modern retention time prediction and robust scoring algorithm can be used to identify a large number of proteins, of the same order of magnitude as fragmentation-based methods. The scoring algorithm introduced here exploits both peptide mass and retention order for the identification of proteins. The new method tends to detect the same proteins as those detected in the fragmentation-based approach. It evaluates peptide peaks only within the context of the candidate proteins from which they arise.

Using this method we were able to identify 257 proteins from *Francisella tularensis*. This is more than twenty times the 12 proteins repeatedly detected by Palmblad *et al.* from membranes of *Yersinia pestis* [20] in the only work that we are aware of, that reports concrete identification results from fragmentation-free MS and predicted HPLC retention times. Relative to the total number of proteins in the samples (at most 1719 in Francisella, and 456 in the membrane fraction of *Yersinia pestis* [30], we detect a much higher proportion of the proteins present in the sample (15% vs 2.6%) and are able to estimate a confidence level for these protein identifications.

4.2 Robustness of the scoring algorithm with regard to prediction errors

As our scoring algorithm only relies on prediction *order*, it is robust against differences between predicted and measured retention time. This relaxes the accuracy requirement on the retention time predictor used: Differences between actual and predicted retention times will not change the outcome as long as they do not alter the retention order.

Furthermore, if the predicted retention order for two peptides is inverted, both can still contribute to the OPM score: As long as they elute in the overlapping sections of their retention profiles, the peaks corresponding to these peptides can be found by the alignment, despite their wrong order.

Finally, a correct retention order used is only relevant when comparing peptides resulting from the digest of the same protein. An incorrect retention order for peptides from other proteins has no influence on the OPM score.

To test these ideas, we have simulated errors by adding Gaussian-distributed noise to each of the predicted peptide retention times and rerunning the analysis. We evaluated the effect of this modification on the estimated false discovery rate and on the number of proteins also found by fragmentation and Mascot. The magnitude of this noise is expressed in the same units as the hydrophobic index used by SSRCalc and described in Section **1.2.2**.

For a Gaussian noise with a standard deviation below 4 units, the FDR only slightly increased for all numbers of positive proteins. Similarly the ROC curve was only slightly modified. However when the standard deviation of the noise reached 5 units and more, the FDR increased to the levels obtained when the order of peptides was not taken into account and even beyound that level. The same degradation was seen on the ROC curve. For comparison purposes we computed the intervals between

the predicted retention times of subsequent tryptic peptides belonging to the same proteins for all proteins in the database. Over the whole database, we found that the median of these intervals was 1.18 SSRCalc units and the 3rd quartile 2.84 units. This is consistent with the magnitude of the introduced noise for which the performance of our alignment-based detection method starts to be disrupted.

4.3 Impact of the ion pairing reagent

Commonly used ion pairing agents for the HPLC of peptides include Trifluoroacetic Acid (TFA) and Formic Acid (FA). The retention time predictor used here [21, 25] had been optimized by its authors for TFA on 100Å pore-size columns. On the other hand, the data that we used have been obtained with FA. Although the authors emphasized that the performance of their predictor was lower in FA than in TFA (the correlation coefficient between predictions and measures was only $R^2 = 0.94$ with FA as opposed to a typical $R^2 = 0.99$ with TFA on different columns), we still obtained an improvement of protein identification with the retention order in TFA. We interpret this result as evidence for the robustness of the OPM method.

4.4 Impact of sample complexity and database size

The performance of most protein identification approaches depends on the sample complexity and the size of the database: The more complex the sample, and the larger the database used for identification, the less one will be able to infer from MS peaks. We believe this to be true for the OPM approach as well. The examples shown here were obtained with samples of medium complexity, namely soluble proteins of whole cell lysate, from a prokaryote with a relatively small proteome: With its estimated 1719 proteins, *Francisella tularensis* has a simple proteome: 262 (13.9%) of the 1886

entries listed in the Integr8 [1] database of complete genomes (version 110, May 26th, 2010) include a number of proteins equal or smaller to that value. Similarly, 50% of the genomes in Integr8, V 110 have less than 3187 proteins.

It is noteworthy that other strategies that require limited fragmentation data were also developed on target proteomes of relatively low complexity. For instance, the AMT strategy was developed on *Deinococcus radiodurans*, which has a proteome of 3085 proteins [11] and the *Yersinia pestis* membrane fraction used by Palmblad [20] contains only 456 proteins.

4.5 Post-translational modifications and missed clivages.

Individual amino acids in proteins are often modified from their canonical form after protein synthesis. These post-translational modifications (PTMs) imply modifications in the measured masses and retention times. They can either affect every instance of the relevant amino acid (fixed PTM) or only a subset of *a priori* unknown size (variable PTM). Similarly, some predicted enzymatic cleavages may not occur experimentally (missed cleavage).

As presented here, the OPM approach does not accommodate PTMs or missed cleavages, but it can easily be extended for both. Conceptually, fixed missed cleavages merely lead to a different set of peptides whose retention times can also be predicted. Variable missed cleavages are more problematic: because of the global nature of the scoring scheme, the match of several peptides could be affected by the modification of only one of them and the optimal alignment needs to be recomputed.

The introduction of a PTM induces changes in both peptide mass and retention time. Any fixed PTM taken into account by the retention time predictor (as is the case here for the alkylation by iodoacetamide) only translates into a different set of ordered

peptides as input for the OPM method. In the case of variable PTMs, two possibilities (with and without the PTM) must be considered individually. As in other methods, the number of cases to consider will increase exponentially with the number of independent, variable PTMs in a given protein.

The case of phosphorylations is of particular importance. According to Steen *et al.* [31], the retention time of peptides is, in many cases, not drastically modified upon phosphorylation, particularly if the number of basic residues in the peptide and the positive charges that they carry can counterbalance the additional negative charge induced by the phosphorylation. This would happen with monophosphorylated tryptic peptides which almost always have a basic residue (lysine or arginine) at their C-terminal end. Therefore as the OPM method is robust to minor errors in the retention time it could readily accomodate phosphorylation using existing retention time prediction methods.

4.6 Incorporating additional information

The interest of a protein identification method based on HPLC and single MS could be questioned in view of the increasing availability of mass spectrometers with fragmentation and MS/MS capacity.

However, even with the current generation of fast instruments, it is still not possible to fragment all the peptide ions from the first MS. Furthermore, not all the precursor ions selected for fragmentation provide spectra leading to the identification of peptide sequences. Therefore, the primary scans constitute a more complete representation of the peptides present in the samples than the series of assigned fragmentation spectra. The present method allows the use of data from this faithful representation of the digest peptides.

Finally, the OPM approach can be used to incorporate additional information into

fragment-based analysis: For instance, in HPLC-MS/MS experiments, some proteins only give rise to a single fragmentation spectrum, which is frequently not considered as sufficient for unambiguous identification. Additional evidence about the presence of such a protein could be obtained by searching for peaks from its other peptides in the primary spectra, even if these peptides have not been fragmented. The scoring method described here can be modified for this purpose: If a peptide has been detected by fragmentation analysis, the alignment score for matching with this peptide is increased strongly above the default matching score. The rest of the alignment is executed as described above, rewarding matches of additional predicted peptides with the primary spectra. With the appropriate score settings, the resulting optimal alignment will incorporate the peptide recognized after fragmentation and the protein identification will thus be based on evidence from both primary and secondary spectra. Alternatively mass and retention time information from one experiment could be used in further experiments to guide the choice of precursor peptides, and thus provide more informative fragmentation steps.

5 Conclusions

Our main conclusion is that highly accurate mass measurements, modern retention time prediction, and a new scoring algorithm can be combined into a fragmentation-free approach that is able to identify hundreds of proteins in a medium-complexity sample. This demonstrates for the first time that, supported by suitable informatics approaches, fragmentation-free mass spectrometry may, in the future and depending on sample complexity, deliver protein identification performance on a level similar to that of fragmentation-based approaches. In this respect, new instruments enable this possibility, for example the Thermo Exactive mass spectrometer (http://www.thermo.com/Exactive).

The OPM method introduced here provides a novel approach of using the

information on peptide retention time for the identification of peptides by HPLC-MS. Contrary to existing methods [11] it does not require previous retention time measurement. Furthermore, it relies on the retention *order* of peptides, and not on their exact retention times. Therefore, it can use predictors not trained with data obtained in exactly the same conditions and on the same HPLC system used for the analysis. In the practical application which we studied here, the OPM approach identified a significant fraction of the proteins found by a fragmentation-based approach, and some additional proteins.

Finally, the alignment method underlying the OPM approach may also be used in combination with fragmentation data in at least two ways: 1) The alignment for one specific protein can include one peptide recognized from MS/MS spectra and other peptides from the same protein matching peaks in primary scans. 2) The alignment results (mass and retention time of the peptides) can direct the choice of ions which will be submitted to fragmentation in further LC/MS-MS experiments with the same samples.

6 Acknowledgments

Work of P. Clote was partially funded by the Digiteo Foundation, as well as by the National Science Foundation grants DBI-0543506 and DMS-0817971. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Frank Rügheimer was supported by a grant of the European Union (FP6, BaSysBio, grant LSHG-CT-2006-037469). The authors would like to thank Sébastien Li-Thiao-Té for discussions and suggestions and Thomas Rolland for comments.

7 Conflict of interest

The authors declare no conflict of interest.

References

[1] Kersey, P., Bower, L., Morris, L., Horne, A., *et al.* Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res* 2005. *33*, D297–D302.

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

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

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

[5] Zhang, W., Chait, B. T. Profound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* 2000. *72*, 2482–2489.

[6] Eng, J. K., McCormack, A. L., III, J. R. Y. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 1994. *5*, 976 – 989.

[7] McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S., *et al.* Direct analysis and identification of proteins in mixtures by lc/ms/ms and database searching at the low-femtomole level. *Anal Chem* 1997. *69*, 767–776.

[8] Baczek, T., Kaliszan, R. Predictions of peptides' retention times in reversed-

phase liquid chromatography as a new supportive tool to improve protein identification in proteomics. *Proteomics* 2009. *9*, 835–847.

[9] Norbeck, A. D., Monroe, M. E., Adkins, J. N., Anderson, K. K., *et al.* The utility of accurate mass and lc elution time information in the analysis of complex proteomes. *J Am Soc Mass Spectrom* 2005. *16*, 1239–1249.

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

[11] Strittmatter, E. F., Ferguson, P. L., Tang, K., Smith, R. D. Proteome analyses using accurate mass and elution time peptide tags with capillary lc time-of-flight mass spectrometry. *J Am Soc Mass Spectrom* 2003. *14*, 980–991.

[12] Guo, D., Mant, C. T., Taneja, A. K., Rodges, R. S. Prediction of peptide retention times in reversed-phase high-performance liquid chromatography ii. *Journal of Chromatography A* 1986. *359*, 519 – 532.

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

[14] Mant, C. T., Zhou, N. E., Hodges, R. S. Correlation of protein retention times in reversed-phase chromatography with polypeptide chain length and hydrophobicity. *J Chromatogr* 1989. *476*, 363–375.

[15] Petritis, K., Kangas, L. J., Ferguson, P. L., Anderson, G. A., *et al.* Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal Chem* 2003. *75*, 1039–1048.

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

[17] Pfeifer, N., Leinenbach, A., Huber, C. G., Kohlbacher, O. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinformatics* 2007. *8*, 468.

[18] Pfeifer, N., Leinenbach, A., Huber, C. G., Kohlbacher, O. Improving peptide identification in proteome analysis by a two-dimensional retention time filtering approach. *J Proteome Res* 2009. *8*, 4109–4115.

[19] Krokhin, O. V. Sequence-specific retention calculator. algorithm for peptide retention prediction in ion-pair rp-hplc: application to 300- and 100-a pore size c18 sorbents. *Anal Chem* 2006. *78*, 7785–7795.

[20] Palmblad, M., Ramström, M., Bailey, C. G., McCutchen-Maloney, S. L., *et al.* Protein identification by liquid chromatography-mass spectrometry using retention time prediction. *J Chromatogr B Analyt Technol Biomed Life Sci* 2004. *803*, 131–135.

[21] Krokhin, O. V., Ying, S., Cortens, J. P., Ghosh, D., *et al.* Use of peptide retention time prediction for protein identification by off-line reversed-phase hplc-maldi ms/ms. *Anal Chem* 2006. *78*, 6265–6269.

[22] Oyston, P. C. F., Sjostedt, A., Titball, R. W. Tularaemia: bioterrorism defence renews interest in francisella tularensis. *Nat Rev Microbiol* 2004. *2*, 967–978.

[23] Guina, T., Radulovic, D., Bahrami, A. J., Bolton, D. L., *et al.* Mgla regulates francisella tularensis subsp. novicida (francisella novicida) response to starvation and oxidative stress. *J Bacteriol* 2007. *189*, 6580–6586.

[24] Matthiesen, R. Extracting monoisotopic single charge peaks from liquid chromatography electrospray ionization mass spectrometry. *Methods Mol Biol* 2007. *367*, 37–48.

[25] Spicer, V., Yamchuk, A., Cortens, J., Sousa, S., *et al.* Sequence-specific retention calculator. a family of peptide retention time prediction algorithms in reversed-phase hplc: applicability to various chromatographic conditions and

columns. *Anal Chem* 2007. *79*, 8762–8768.

[26] Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, New York, NY, 2005.

[27] Koenker, R. *quantreg: Quantile Regression*, 2008. R package version 4.24.

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

[29] Rohmer, L., Guina, T., Chen, J., Gallis, B., *et al.* Determination and comparison of the francisella tularensis subsp.novicida u112 proteome to other bacterial proteomes. *J Proteome Res* 2008. *7*, 2016–2024.

[30] Pieper, R., Huang, S.-T., Clark, D. J., Robinson, J. M., *et al.* Integral and peripheral association of proteins and protein complexes with yersinia pestis inner and outer membranes. *Proteome Sci* 2009. *7*, 5.

[31] Steen, H., Jebanathirajah, J. A., Rush, J., Morrice, N., Kirschner, M. W. Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. *Mol Cell Proteomics* 2006. *5*, 172–181.

8 Figures and table legends

Figure 1: Principle of the identification method: for each protein the predicted tryptic peptides are sorted according to their predicted retention time (bottom). They are then matched to ions from the spectra in the same order. Each trapezoid represents one experimental spectrum, and the masses of the peptides are symbolized by the height of the lines. Intensities of the peaks are not considered. Solid lines: Matching peptides; Dashed lines: Non matching peptides.

Figure 2: This dynamic programming algorithm iteratively computes the score $s_{i,j}$ of the best alignments between a prefix $(p_1,...,p_j)$ of the list of peptide masses and a prefix $(S_1,...,S_i)$ of the list of spectra. Here the values of the *no-peptide score* and the *no-spectrum score* are 0 and −1 respectively.

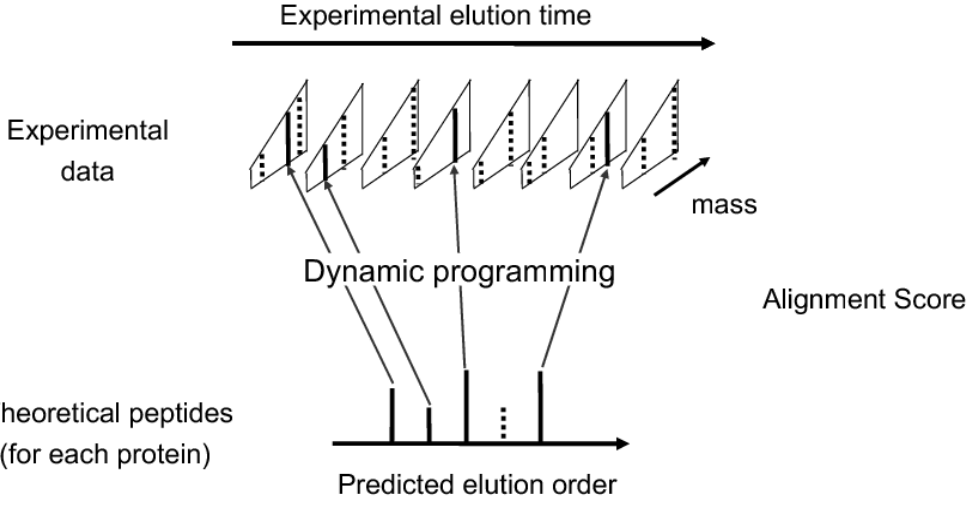
Figure 3: Distribution of number of matched peptides for proteins from *Francisella tularensis subsp. novicida* with one HPLC-MS experiment. Each point represents a protein from the sequence database. For visualization only, the positions of the points are shifted by a small gaussian random amount (σ=0.5). The lines represent the best linear fits for the data (see text). Abscissa : number of peptides longer than 4 amino acids per protein; Ordinate: number of matched peptides.

Figure 4: Effect of incorporating the order of the peptides. A : Values of the estimated false discovery rate (FDR) for different numbers of proteins identified above the threshold, with or without taking the peptide order into account. B : Receiver Operating Characteristic (ROC) curves for the comparison of the novel method with results from peptide fragmentation and Mascot. The ROC curves are obtained using the output from Mascot as the reference set.

Table 1: OPM scores and quantile assignment of *F. tularensis* proteins. For each protein this table lists the number of tryptic peptides longer than 4 amino acids and the number of these peptides matched to the experimental MS1 peaks by the alignement procedure. In the last column the quantile assignment of the protein in shown. This corresponds to the position of the alignment score in the distribution of all the scores for decoy proteins with the same number of tryptic peptides.

Table 2: Number and proportion of total proteins detected for several quantile limits. Top: with experimental peptides matching in the predicted order; Bottom: without ordering the peptides. Real proteins: number of real proteins detected (portion of total proteins); Random proteins: number of randomized proteins from the decoy database detected; FDR estimate: estimated proportion of absent proteins from the protein with scores above the quantile limit; Frag. overlap: detected proteins also identified from the fragmentation data with the Mascot software. The total number of proteins in the database is 1719 and 436 of them were identified by Mascot from the fragmentation spectra.

Table 3: Proteins identified by fragmentation and Mascot. This table lists the accesssion numbers and Mascot scores for the 436 proteins detected by Mascot on the basis of the fragmentation (MS-MS)spectra. Based on the comparison with a decoy data base this number of proteins was retained in order to obtain a false discovery rate of 5.7%.


Experimental elution time
Experimental
data
mass
Dynamic programming
Alignment Score
Theoretical peptides
(for each protein)
Predicted elution order

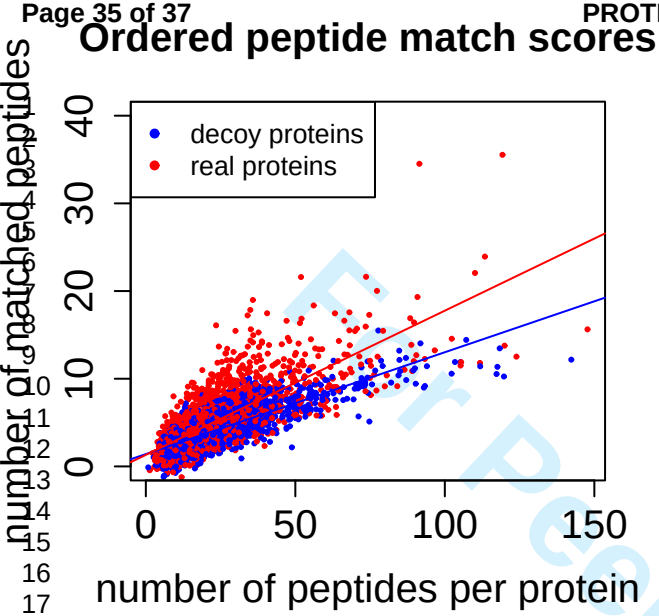
Algorithm OrderedPeptideMatchScore

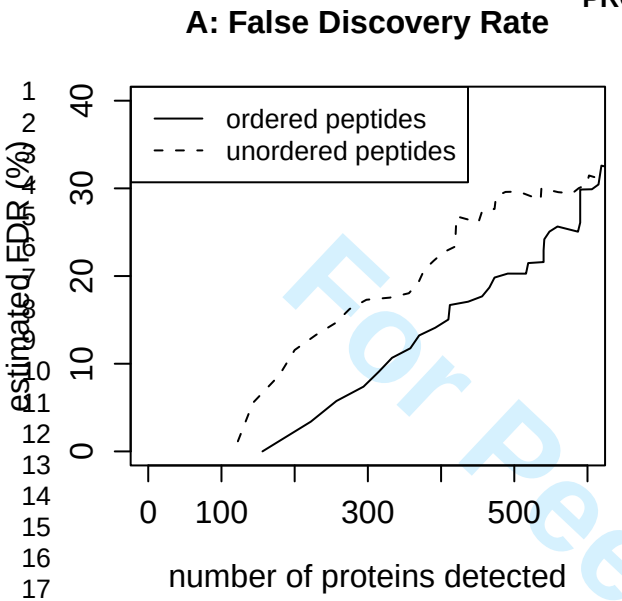
Input:

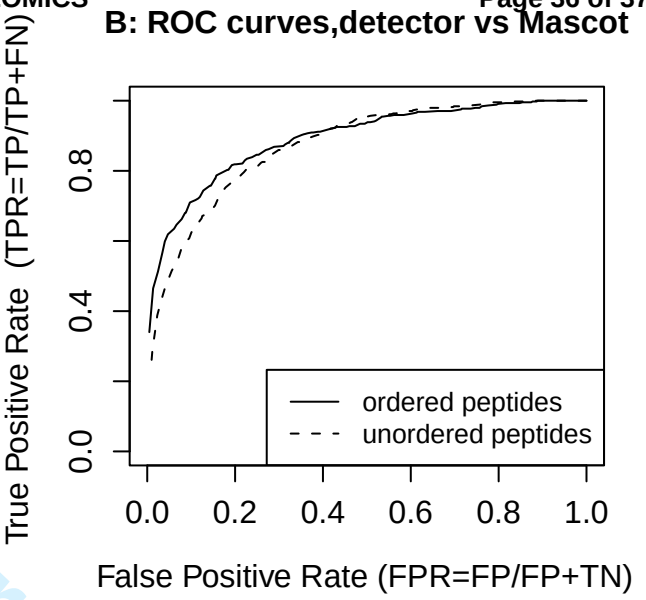
- A protein P
- A list $(S_{1,}.., S_m)$ of decharged, deisotoped spectra
- A list of masses $(p_{1,}.., p_n)$ of the tryptic peptides of P, ordered by predicted elution time
- A *peak match threshold* Δ
- Constant *no-peptide score nps*, *no-spectrum score nss*, and a *peptide-spectrum match score ps(q)*

1. For $i=1,\ldots,m$
2. For $j=1,\ldots,n$
3. If S_i has a peak in the interval $(p_j-\Delta, p_j+\Delta)$, set $ps=+1$; otherwise set $ps=-1$.
4. Compute, $s_{i,j}=max(s_{i-1,j-1}+ps, s_{i-1,j}+nps, s_{i,j-1}+nss)$ under the convention that $s_{i,j}=0$ if $i=0$ or $j=0$

Output: $s_{m,n}$







Ordered peptide match score

Quantile	0.80	0.90	0.95	0.99
Real proteins	721 (41.9%)	531 (30.9%)	407 (23.7%)	257 (14.9%)
Random proteins	372 (21.6%)	174 (10.1%)	93 (5.4%)	15(0.91%)
FDR estimate	34.60%	24.80%	16.90%	5.7%
Frag. overlap	376	337	295	226

Unordered peptide match score

Quantile	0.80	0.90	0.95	0.99
Real proteins	701 (40.8%)	479 (27.9%)	357 (20.8%)	181 (10.5%)
Random proteins	350 (20.4%)	168 (9.8%)	85 (4.9%)	16 (0.9%)
FDR estimate	36.3%	28.8%	20.0%	8.6%
Frag. overlap	367	301	256	158