Research Article

Rapid Communications in Mass Spectrometry

Received: 27 August 2012 Revised: 1 November 2012 Accepted: 2 November 2012 Published online in Wiley Online Library

Rapid Commun. Mass Spectrom. **2013**, 27, 391–400
(wileyonlinelibrary.com) DOI: 10.1002/rcm.6462

Combination of Edman degradation of peptides with liquid chromatography/mass spectrometry workflow for peptide identification in bottom-up proteomics

Anna A. Lobas[1,2], Anatoly N. Verenchikov[3], Anton A. Goloborodko[1,2,4], Lev I. Levitsky[1,2] and Mikhail V. Gorshkov[1,2]*

[1]Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia
[2]Moscow Institute of Physics and Technology (State University), Dolgoprudny, Moscow region, Russia
[3]Mass Spectrometry Consulting Ltd., Bar, Montenegro
[4]Department of Physics, Massachusetts Institute of Technology, Boston, MA, USA

RATIONALE: High-throughput methods of proteomics are essential for identification of proteins in a cell or tissue under certain conditions. Most of these methods require tandem mass spectrometry (MS/MS). A multidimensional approach including predictive chromatography and partial chemical degradation could be a valuable alternative and/or addition to MS/MS.
METHODS: In the proposed strategy peptides are identified in a three-dimensional (3D) search space consisting of retention time (RT), mass, and reduced mass after one-step partial Edman degradation. The strategy was evaluated *in silico* for two databases: baker's yeast and human proteins. Rates of unambiguous identifications were estimated for mass accuracies from 0.001 to 0.05 Da and RT prediction accuracies from 0.1 to 5 min. Rates of Edman reactions were measured for test peptides.
RESULTS: A 3D description of proteolytic peptides allowing unambiguous identification without employing MS/MS of up to 95% and 80% of tryptic peptides from the yeast and human proteomes, respectively, was considered. Further extension of the search space to a four-dimensional one by incorporating the second N-terminal amino acid residue as the fourth dimension was also considered and was shown to result in up to 90% of human peptides being identified unambiguously.
CONCLUSIONS: The proposed 3D search space can be a useful alternative to MS/MS-based peptide identification approach. Experimental implementations of the proposed method within the on-line liquid chromatography/mass spectrometry (LC/MS) and off-line matrix-assisted laser desorption/ionization (MALDI) workflows are in progress.

Protein discovery is a rapidly evolving field of science in the post-genome era in which proteomics is a technological platform to effectively characterize proteins on a large scale. To learn about the basic mechanisms associated with cell functioning, protein expression has to be measured under continuously changing external conditions and/or in pathological states of the organism. The dynamic nature of the proteome is one of the main challenges in proteomics that stimulate the current trend towards development of high-throughput strategies. The most widely used strategy for protein identification is bottom-up proteomics based on the identification of proteolytic peptides, obtained from certain cell or tissue protein mixtures by enzymatic digestion.[1–7] In this strategy mass spectrometry (MS) provides data on the sequences of identified peptides. This sequence information is obtained from fragmentation mass spectra of the peptides using a variety of methods of tandem mass spectrometry (MS/MS). A two-dimensional array of mass spectrometric data (peptide mass, peptide fragment masses) is further used to identify peptides in databases using search algorithms,[8–11] or for *de novo* sequencing.[12] However, due to the sequential selection of precursor ions in the first stage of the MS process and due to spreading of signal intensity between multiple fragment peaks, the combination of speed and sensitivity of MS/MS experiments is limited, and the majority of minor species (i.e. the species present in the original mixtures in minute amounts) are lost from the analysis.[13] In addition, due to well-known reasons, such as incomplete fragmentation pattern, peptide ion losses during the in-trap isolation for subsequent fragmentation, 'chimera' spectra, arising from simultaneous isolation and fragmentation of several precursor ions,[13,14] etc., this strategy does not provide unambiguous peptide identifications and the search engines return a number of possible sequence candidates resulting in wrong sequence assignments. Because of the above reasons (and others not mentioned) it is tempting to search for approaches to unambiguous peptide identifications alternative to MS/MS.

* *Correspondence to:* M. V. Gorshkov, Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, Moscow, Russia.
E-mail: mike.gorshkov@gmail.com

When the mass of a peptide is measured with a sufficient precision such that it is unique among all the peptides predicted from a given genome sequence, it can be used as a biomarker for the identification of its parent protein by mass only.[15] Even at sub-ppm mass accuracy this one-dimensional 'Accurate Mass Tag' technique is useful only for small protein databases, such as baker's yeast (*S. cerevisiae*).

To improve further the efficiency of identification strategy with lower requirements for the mass measurement accuracy (MMA) any complementary information about peptide sequences becomes a valuable addition to the MS data. One possible strategy employs the fact that a typical bottom-up proteomics workflow incorporates a separation of a mixture of proteolytic peptides using liquid chromatography (LC) prior to MS analysis.[16] Since the retention time (RT) of a peptide depends on its primary structure, LC can provide additional complementary information about peptide sequence. This makes possible the strategies based on a two-stage approach: at the first stage, standard LC/MS/MS analyses are undertaken on fractionated protein digests to yield tentative peptide identifications followed by the generation of a database containing the calculated masses based on putative peptide sequences and their corresponding chromatographic retention times. This database is subsequently validated in an LC/MS experiment measuring the accurate masses of the detected peptides at the normalized retention times observed for their initial identification. At the second stage, these accurate mass and time (AMT) tags are used in a series of LC/MS measurements as biomarkers of the presence of a given protein without resorting systematically to MS/MS for identification. The above strategy has been extensively explored in recent years and is known as the AMT-tag approach.[17] In brief, instead of searching the database using purely mass spectrometric data, peptides are identified in a two-dimensional space of (mass, RT), in which the chromatographic dimension provides sequence-related information.

AMT database generation with LC/MS/MS requires extensive experimental work using large sets of peptides. Alternatively, these databases can be generated *in silico* by employing peptide retention times predicted for the specified HPLC conditions.[18] A number of retention time prediction models have been developed recently allowing a rather high level of prediction accuracy with the correlation between predicted and experimental values of R^2 ranging from 0.85 to 0.95. They include empirical models such as SSRCalc;[19] additive models, accounting for amino acid composition of peptides;[20,21] as well as physical models, BioLCCC (Liquid Chromatography of Biomacromolecules under Critical Conditions),[22–24] and QSRR (Quantitative Structure-Retention Relationship).[25] Also there is a number of algorithms for retention time prediction based on machine learning, employing a kernel-based support vector machine approach[26] or artificial neural networks.[27] While the AMT-tag approach is a highly promising strategy for high-throughput identification and quantification of peptides, it has been successfully applied in proteome-wide scale analyses for rather small genomes.[17] Indeed, the 2D search space (mass, RT) does not provide unambiguous peptide identification for large proteomes. To improve further the unambiguous identification rate both mass measurement (MMA) and RT prediction accuracies have to be elevated beyond the edge of current technologies. Alternatively, peptide identification may be enhanced by obtaining additional information on peptides with partial sequencing.

The Edman degradation reaction (Fig. 1(A)) is a long-known method for peptide sequencing, in which amino acid residues are analyzed by chromatography.[28,29] It is used to cleave the peptide bond at the N-terminal amino acid residue thus producing analogues of y_{n-1} fragments (so-called reduced peptides). The reaction rate is known to have little dependence on peptide sequence except when proline is the first residue. The combination of Edman degradation with mass spectrometry is known as ladder sequencing implemented previously for *de novo* sequencing of peptides.[30,31] A decade ago the single-step degradation was proposed to obtain additional mass spectrometric information about peptide sequences for subsequent database search.[32–34] In these earlier works a combination of chemical degradation with mass spectrometry was tested experimentally for single protein identifications,[32,33] and considered *in silico* for the whole proteomes of *E. coli*, *H. influenzae* and *C. elegans*.[34] Mass spectrometry data in resulting 2D search space (peptide mass, reduced peptide mass) give higher identification rates, however, with lesser requirements for the completeness of the fragmentation pattern and/or the accuracy of peptide mass measurements. For example, *in silico* studies have shown that up to 90% peptides can be unambiguously identified for relatively small proteomes using the partial degradation technique and peptide masses only.[34]

The purpose of this work was to evaluate the feasibility of unambiguous peptide identifications without employing the MS/MS technique by combining three complementary peptide descriptors obtained experimentally in an LC/MS run. The proposed strategy can be considered as an extension of the above described approaches and relies on a combination of predictive chromatography, mass spectrometry, and partial chemical degradation. It can be called an 'extended AMT-tag' method for shotgun proteomics in which peptides are identified in a 3D space (peptide mass, peptide RT, reduced peptide mass). Using *in silico* simulations the rates of unambiguous identifications of tryptic peptides from yeast and human proteomes were evaluated. A number of possible protocols for the Edman reaction were tested experimentally to prove the feasibility of on-line implementation of the proposed method.

EXPERIMENTAL

All calculations were performed using two databases: *S. cerevisiae* (baker's yeast) and human proteomes, downloaded from the UniProt Knowledge Base.[35] The yeast proteome *S. cerevisiae* database contains 1877 entries (release-2011_10) and is a common, widely used small-sized model system for proteomic studies. The human proteome database contains 56392 entries (release-2011_07) and was selected to assess the capabilities of the proposed method for a large database.

The software used for the calculations was written in Python programming language and based on the annotated library called 'Pyteomics',[36] developed in our laboratory and freely available for download.[37] This library allows the processing of FASTA input files and generation of protein

Rapid Communications in Mass Spectrometry

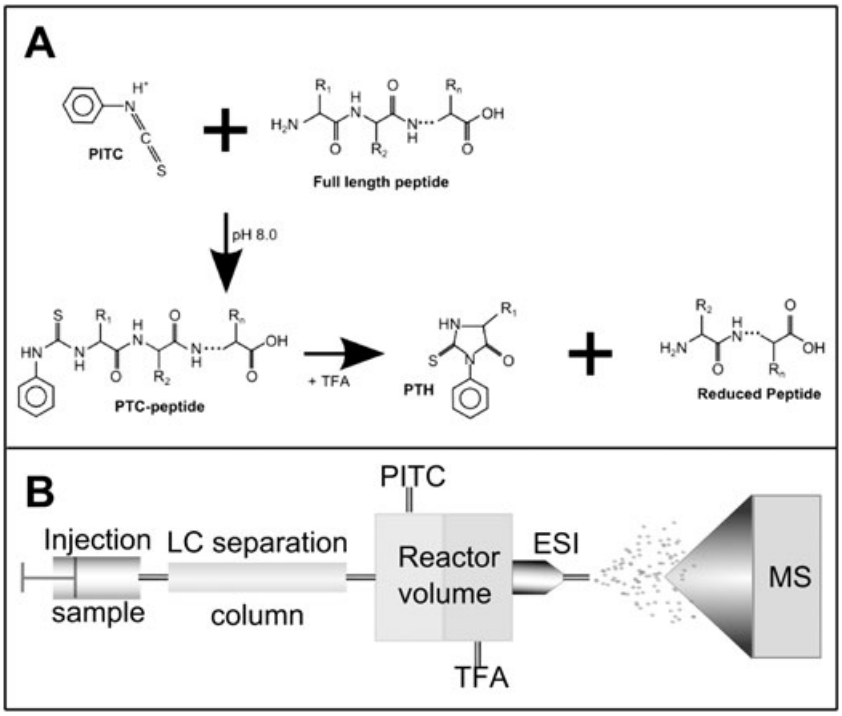

Figure 1. (A) Schematic of the Edman degradation reaction. First stage: phenyl isothiocyanate (PITC) binding to the N-terminus of the peptide at high pH. Second stage: cleavage of the N-terminal amino acid residue at low pH. (B) Schematic of the possible experimental setup for on-line implementation of the proposed strategy. The peptide mixture after trypsin digestion is injected into the LC separation column. Then, the separated peptide fractions enter the Edman degradation reactor, consisting of two chambers: the first one operated at high pH, in which PITC is added to the peptide and attached to its N-terminus, and the second one with low pH. PITC is added to the peptide fraction in the former chamber. In the latter chamber, TFA is added to the mixture to provide the cleavage of the N-terminal amino acid residue. After chemical degradation the mixture of full-length and reduced peptides is transferred to the ESI source, followed by MS measurements.

tryptic digests including peptide modifications *in silico*. Another in-house developed Python open-source library 'Pyteomics.biolccc' was used for peptide retention time prediction based on the BioLCCC peptide separation model.[22–24] This library is available on-line.[38] Standard nano-LC separation conditions were used for RT calculations: column length of 150 mm, inner diameter of 0.075 mm, flow rate of 0.3 μL/min, pore size of 100 Å and a 60 min linear gradient from 2 to 40% ACN with 0.1% trifluoroacetic acid (TFA) as an ion-pairing agent. For highly hydrophobic peptides which are not eluted under these typical gradient conditions the upper ACN concentration was automatically prolonged until the elution. Monoisotopic masses were calculated for all peptides and one possible missed cleavage was allowed when generating the tryptic peptides list. The effects of post-translational modifications of amino acids, including N-terminal acetylation, methionine oxidation and phosphorylation of serine, threonine and tyrosine residues, as well as phenyl isothiocyanate (PITC) modification of the lysine side chain taking place during the Edman reaction, were assessed in a separate set of simulations. A peptide was considered as unambiguously identified if it had no 'neighbors' within certain mass and time accuracy windows in the three-dimensional search space (peptide mass, reduced peptide mass, peptide RT).

The peptides used in experiments were purchased from Sigma-Aldrich (St. Louis, MO, USA). Edman degradation consisting of two steps was performed as follows: for binding PITC to the N-terminus of the peptide it was dissolved in the mixture of water/ethanol/pyridine (1:1:1) containing 5% of PITC at the concentration of 1 mg/mL. The reaction was held at 45 or 50 °C for 1 to 30 min. For cleavage of the N-terminal residue an equal volume of TFA was added to the mixture, which was subsequently incubated at 45 to 60 °C for 1 to 10 min. For the chromatographic analysis the mixtures were dissolved in four volumes of water and centrifuged for 10 min at 18 000 *g*.

LC separations were performed using a gradient HPLC system (Rainin Dynamax SD-200; Mettler Toledo, Greifensee, Switzerland) in binary water/acetonitrile solutions: solution A (water, 0.1% formic acid (FA)), solution B (95% acetonitrile, 5% water, 0.1% FA). Two types of elution gradients were used: linear 5–50% B in 30 min, as well as a step gradient, including 10 min wash with 5% B and 20 min with 45% B.

Mass spectrometric analyses of the peptides were performed using an in-house built hybrid 1.25 Tesla Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer based on the permanent magnet and equipped with a heated metal capillary electrospray ionization (ESI) source and a linear quadrupole

radio-frequency (rf) ion trap for ion accumulation and isolation,[39] and a commercial rf ion trap ESI-LC/MS instrument (Amazon SL; Bruker Daltonics, Bremen, Germany).

RESULTS AND DISCUSSION

Database analyses

In the first round of *in silico* studies, the databases used for the estimation of the efficiency of proposed approach were analyzed. The yeast protein database, containing 1877 entries, after *in silico* trypsin digestion with one possible missed cleavage, produced 122 637 peptides (without repeats). Then, the peptides with the sequence lengths exceeding 4 amino acid residues and with masses less than 3000 Da were selected. For the selected group of 99 900 peptides, the distributions by masses, retention times and N-terminal amino acid residues were calculated. The same procedure was employed for the human proteins, for which the database for the analysis consisted of 1 490 728 tryptic peptides. The resulted distributions are shown in Figs. 2(A), 2(B), and 2(C) and reveal similar patterns for both yeast and human proteomes. Mass distributions (Fig. 2(A)) have a decaying exponential form on a crude scale, and there are ca. 1 Da mass clusters on a fine scale. This clustering was previously described by Gay *et al.*[40] RT distributions are shown in Fig. 2 (B) and typically reveal two maxima at 35 min and in the low RT region for the used gradient conditions. Finally, the N-terminal residue distribution is shown in Fig. 2(C). Note that the uniformity of this distribution (relative to the other physicochemical properties of the residues) is an important factor for the successful use of N-terminal residue information as an additional search space dimension. Although, as one can see in Fig. 2(C), this distribution is not uniform, the improvements in the successful peptide identification rate are expected since many of the residues are present at the N-terminus of tryptic peptides with almost equal probabilities. The most abundant N-terminal amino acid is leucine, and, since it cannot be distinguished from isoleucine by MS, their contributions were summed. About 15% of all tryptic peptides under consideration start from either a leucine or an isoleucine residue. For these peptides we expect minimal improvement in unambiguous identification rate from the additional information about the N-terminus. Note, however, that leucine and isoleucine are chromatographically distinguishable residues.[24] Maximum improvement is predicted for peptides starting with tryptophan (W) or proline (P) residues. According to this distribution we expect a shortening of the list of peptide candidates by a factor of 3 to 30 compared to a standard two-dimensional AMT-tag strategy.

Two-dimensional (mass, RT) distributions of yeast tryptic peptides with N-terminal leucine/isoleucine and tryptophan residues are shown in Figs. 2(D) and 2(E), respectively. These distributions have almost identical patterns that further illustrate the absence of a strong correlation between N-terminal amino acid residues and masses or retention times of peptides. This figure also demonstrates the orthogonality between mass and RT for the peptides under study.

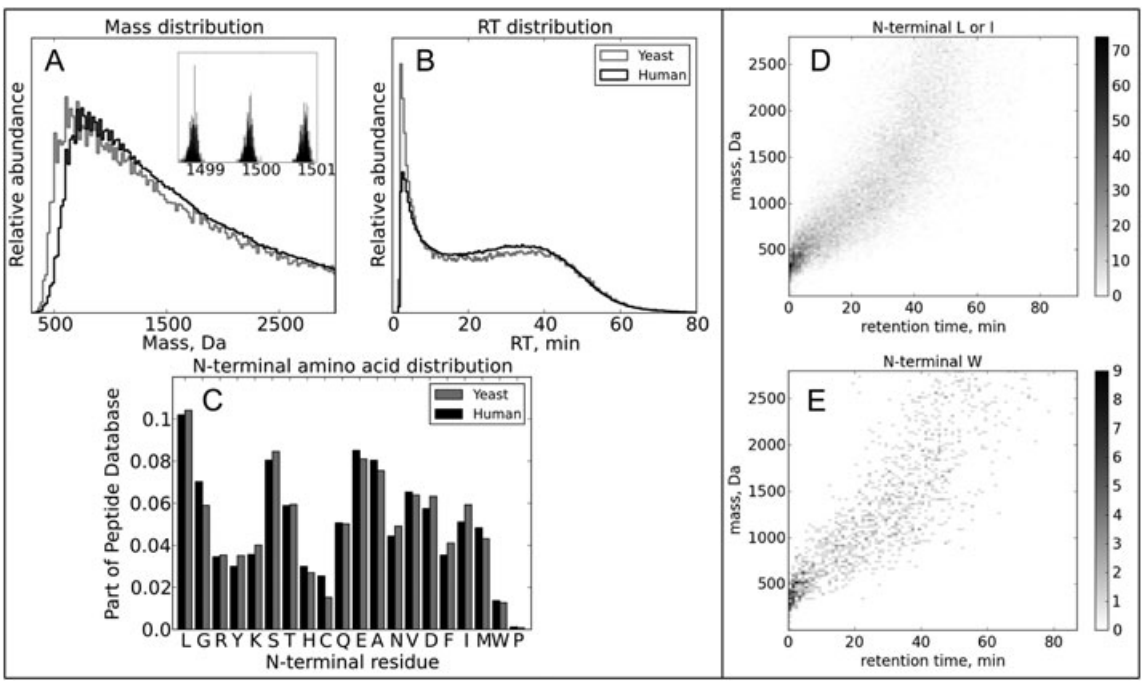

Figure 2. Tryptic digest of model protein databases generated *in silico*. Black plots correspond to the baker's yeast proteome (99 900 peptides); gray plots represent the human proteome (1 490 728 peptides). (A) Normalized mass distributions of peptides. (B) Normalized distributions of peptides by retention times. (C) Normalized distributions of peptides by N-terminal amino acid residue. Analyses of tryptic digest of yeast proteins. Two-dimensional (mass, RT) distributions of yeast tryptic peptides with N-terminal (D) leucine/isoleucine and (E) tryptophan residue. The size of the cell is 25 Da × 1 min. The bar with varying dark color intensity stands for the number of peptides per cell.

Rapid Communications in Mass Spectrometry

Evaluation of the rate of unambiguous peptide identifications in 3D search space

At the next step of the analysis we assessed the utility of combining the partial chemical degradation via Edman reaction with LC and MS data for peptide identifications. The rate of unambiguous identifications of tryptic peptides was defined as the ratio between the number of peptides identified unambiguously and the number of all peptides under consideration. This rate was calculated for a wide range of mass measurement accuracies from 0.001 to 0.05 Da. These values correspond to 1 to 50 ppm of relative mass measurement accuracy for tryptic peptides with the average *m/z* of 1000. The lower end of this MMA range is attainable by high-performance mass spectrometry such as LC/HRT,[41] FTICR, or Orbitrap FTMS, while the upper end is typically achieved when using low-resolution MS instruments based on rf quadrupole ion traps. The range of retention time accuracies was from 0.1 to 5 min. Note that the lower end of this range corresponds to the experimental precision of modern HPLC systems. None of the existing models and/or algorithms for peptide retention time prediction are capable of delivering this level of accuracy.[18,42] The upper end of the considered retention time prediction accuracy range corresponds to more realistic performance of retention prediction algorithms. For example, the statistically determined standard error of the retention time prediction using the BioLCCC model for the gradient conditions used in this study is ~2 min.[22,23]

The results of the evaluation of the unambiguous peptide identification rate in the proposed 3D search space are shown in Figs. 3(A) and 3(B). As shown in this figure one should expect more than 60% of unambiguous peptide identifications using the proposed 3D search space for the human protein database, and more than 85% for the yeast proteome, with rather modest values of mass measurement and retention time prediction accuracies of 0.01 Da and 2 min, respectively. This demonstrates a significant improvement compared with the 2D search space (mass of peptide, mass of reduced peptide) proposed previously.[33]

Figures 3(C) and 3(D) show the improvement in peptide identifications provided by addition of information about the N-terminal amino acid residue compared with the standard AMT-tag approach (which is also two-dimensional). In this figure the plots marked as '3D' correspond to the proposed 3D search space. For comparison the figure also contains the results from the standard AMT-tag approach for mass measurement accuracy of 0.01 Da. The corresponding

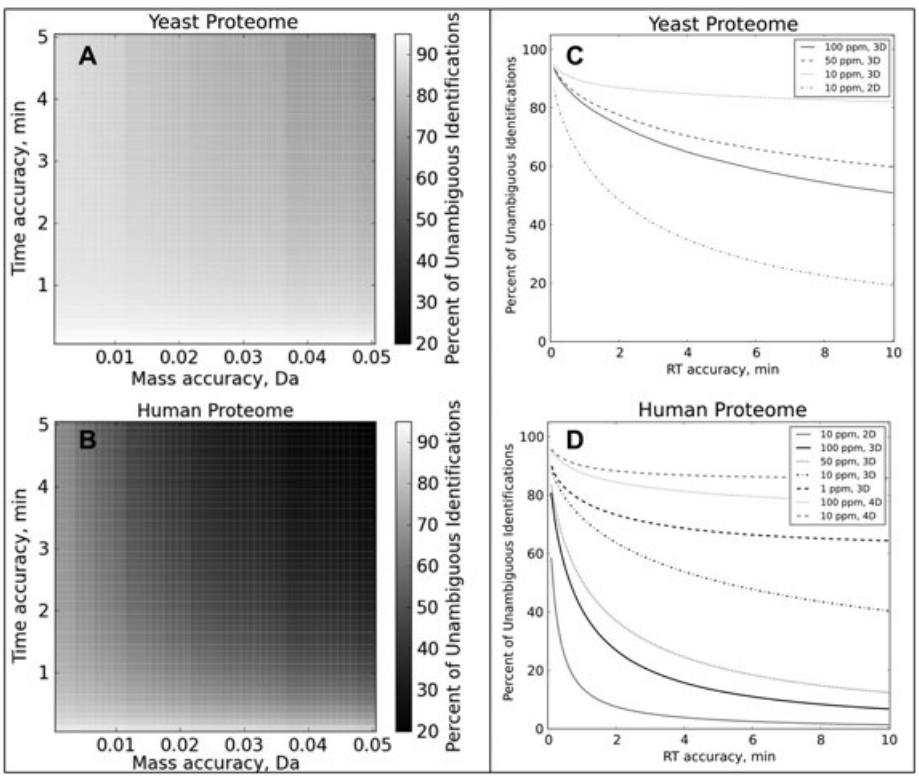

Figure 3. Percent ratio of unambiguous identifications for (A) yeast and (B) human tryptic peptides in three-dimensional search space (mass of full-length peptide, mass of reduced peptide, RT) as a function of the accuracies of mass measurements and retention time predictions. Dependence of the rate (in %) of unambiguous identifications of peptides on the values of time and mass accuracies. The values of mass accuracies in ppm correspond to the average peptide mass of 1 kDa: (C) yeast peptides, (D) human peptides database. '2D' designation stands for the standard two-dimensional (mass, RT) AMT-tag strategy, '3D' designation stands for the proposed 'extended' AMT-tag strategy in which the information on the N-terminal amino acid residue is added into the (mass, RT) search space. Finally, '4D' stands for the same proposed strategy, in which two N-terminal residues are known.

plots are depicted as '2D' in Figs. 3(C) and 3(D) illustrating a significant increase in the rate of unambiguous identifications by incorporating the third search dimension. For example, for the human peptide database (Fig. 3(D)) the increase in the identification rate is 1.5-fold for the highest (and actually not realistic) retention time accuracies, while it can be as high as 30-fold for the non-accurate retention time predictions (10 min). The increase in successful identification rate can be as large as 4-fold achieving 90% for the yeast proteome, as shown in Fig. 3(C).

In addition to the 3D search space, the contribution of the information about the second N-terminal residue as the fourth dimension was also considered. The results of this evaluation are shown in Fig. 3(D) by the plots marked as '4D': even with 100 times worse MMA of 100 ppm the rate of unambiguous identifications is higher for the 4D search space compared to the 3D one at the MMA of 1 ppm. Moreover, obtaining the information about two N-terminal residues, which is certainly an instrumental challenge, will significantly reduce the requirements for the accuracy of predicted chromatographic retention times. This will allow the use of less accurate physical peptide retention models, such as additive, or BioLCCC, rather then the empirical ones. The latter provide better accuracy yet in the limited range of separation conditions and require large sets of known peptides for the model training.

Modified peptides

Proteins in eukaryotic organisms are usually subject to post-translational modifications (PTMs). The multi-dimensional search space strategy proposed in this study was tested for the modified peptides as well. In the first step, N-terminal acetylation of the proteins as a variable modification was taken into account only. This modification is one of the easiest for identification. Firstly, it can appear only at the N-terminus of the protein, which means that no more than two acetylated peptides correspond to each protein in the case of one possible missed cleavage. Secondly, acetylated peptides are not subject to the Edman degradation reaction as PITC can only bind to an NH_2- group of the peptide. Therefore, there will be no reduced pairs for the acetylated peptides, which can be further distinguished from the non-acetylated ones within the proposed 3D search space. As a result, the addition of the N-terminal acetylation as a variable modification does not decrease the rate of unambiguous identifications as shown for both the yeast and human protein databases. The lengths of peptides under consideration were from 4 to 30 amino acid residues. PITC modification of lysine residues was also taken into account as a variable one. For the yeast protein database the set of peptides under study consisted of 80 242 entries without acetylation and 1591 acetylated peptides. Up to 69 910 of 80 242 (87.1%) non-acetylated peptides, 1508 of 1591 (94.8%) acetylated peptides when analyzed separately, and 71 418 of 81 833 peptides in the whole set (87.3%) were unambiguously identified in the 3D search space for mass and time accuracies of 0.01 Da and 2 minutes, respectively. This actually means that all of the acetylated peptides were distinguished from non-acetylated ones. In the case of human proteins, 702 962 of 1 152 560 entries without acetylation (61.0%), 20 455 of 25 973 acetylated (78.7%), and 723 417 of 1 178 533 (61.4%) peptides in the whole set were unambiguously identified.

In the next round of calculations, phosphorylation of serine, threonine and tyrosine, as well as methionine oxidation, were taken into account. The lengths of peptides were in a range from 4 to 30 amino acid residues and all possible combinations of modified and non-modified residues were considered. Accounting for the above modifications resulted in a set of 1 499 546 peptides for the yeast database that is an almost 20-fold expansion of the database compared to the non-modified peptides (80 242 entries). The size of this database is very close to the size of the human peptide database without modifications (1 490 728 entries) which makes it easy to compare the results for peptide identifications in both cases. The results are shown in Fig. 4(D). The rate of unambiguous identifications of modified yeast peptides is rather low; it has a weak dependence on mass accuracy and a strong dependence on retention time prediction accuracy compared with the human non-modified peptide database. This observation is expected and the explanation for this difference is that there are many more isomers in the set of modified yeast peptides when multiple modification sites are considered. For instance, 55 881 of 80 242 peptides in the yeast database have more than one possible phosphorylation site. This results in the increase in the number of peptides which have the same full-length peptide and reduced peptide masses, except the case of N-terminal amino acid residue modification. Thus, the only dimension in which these peptides can be distinguished is the retention time which is known to be sequence- and modification-site specific.[43] Note also that additive models for retention time prediction[21] are not applicable because they are not sequence-specific. However, even for sequence-specific models, such as BioLCCC and SSRCalc, the required retention time prediction accuracy of 1 min or less is beyond their current capabilities. Moreover, the modified yeast peptides have higher densities of mass and RT distributions, as shown in Figs. 4(A) and 4(B), which further complicates their unambiguous identification. The large shift in the mass distribution of the modified peptides to the higher mass region is explained by the multiple modifications (Fig. 4 (C)). Note that the higher is the mass of a peptide, the more sites of possible modification it has, and the more modified peptides it produces. If the number of variable modification sites in a peptide is n, it gives rise to 2^n modified peptides.

Experimental implementation

The ultimate goal of the present study is the MS/MS-free peptide identification through a combination of a standard AMT-tag approach for full-length peptides and partial chemical degradation followed by MS analysis of reduced-length peptides. In previous reports,[32,33] the Edman degradation reaction was performed separately, and the mixture of full-length and reduced peptides was then analyzed by MS. The major differences between the system used in the studies by Van der Rest *et al.*[32] and the one proposed here are the addition of the LC separation stage in the workflow and chemical cleavage of the N-terminal residue instead of in-trap ion fragmentation. For the off-line setup certain fractions of full-length peptides eluting from the HPLC column are collected, subjected to partial chemical degradation one after another, and then transferred to a mass spectrometer. This approach is easier for the practical implementation and similar to the one described by Van der Rest *et al.*[32] The off-line workflow

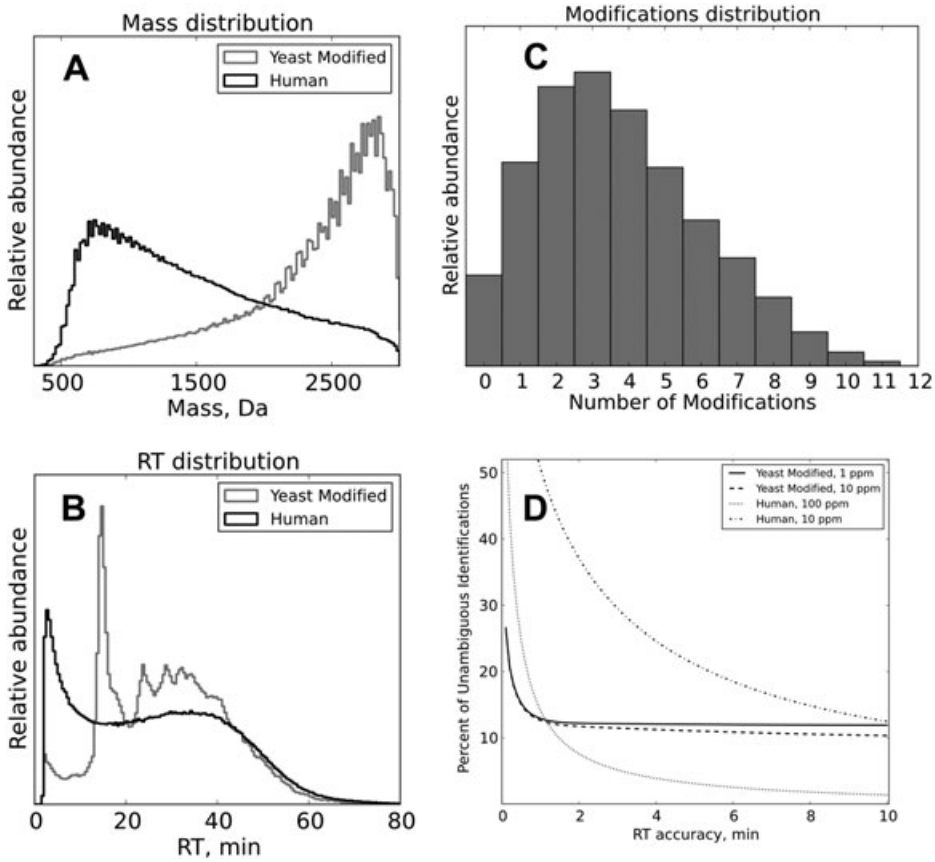

Figure 4. Mass and retention time distributions for PTM-modified yeast peptides and unmodified human peptides. The distributions for modified peptides have sharper peaks, providing lower unambiguous identification rate compared to the human peptide database. (A) Mass distribution of PTM-modified yeast peptides is strongly shifted to the higher mass region. (B) Retention time distribution has several sharp peaks. (C) The distribution of modified yeast peptides by the number of modifications. Most of the peptides under study are multiply modified. (D) Comparison of the efficiency of the use of proposed three-dimensional search space for identification of yeast peptides with PTMs and human peptides without modifications. Dependence of the rate (in %) of unambiguous identifications of peptides on the value of time accuracy for modified yeast peptides and for non-modified human peptides with various values of mass accuracy.

may also be automated using a fraction collector and a robotic system to allow addition of the required chemical reagents. Using peptide immobilizing membranes, several steps of the Edman degradation may be performed in series thus adding more sequence-specific information.[30] Besides, using the volatile reagents, such as TFEITC instead of PITC,[31] TFA and pyridine, would reduce the chemical poisoning that sacrifices the ionization efficiency in the ESI source. The off-line strategy requires rapid fraction analysis, such as robotized rapid injections with the ESI source or scanning using the MALDI source. The on-line workflow is shown schematically in Fig. 1(B). In this setup, the peptides separated by HPLC come into a reactor by fractions with their subsequent chemical degradation, followed by the injection of the degradation products into the ESI source of a mass spectrometer. The Edman degradation reaction consists of two steps: phenylisothiocyanate (PITC) binding to the NH_2 group of the peptide, performed at high pH value, and the cleavage of the PITC-modified N-terminal amino acid residue, performed at low pH value. Note that most of the reported side reactions related to Edman degradation affect the released PTH-amino acids, which is important in the case of a regular sequencing contrary to ladder sequencing. In the latter case the reduced peptides are only analyzed. For this reason the reaction protocols may be simplified by eliminating the need for nitrogen blowing.[31] The main challenges are in designing the two-chamber reactor for on-line implementation of the Edman reaction, optimizing the speed of the reaction to make it applicable for on-line chemical degradation, and adjusting the mixture of reaction products to the conditions suitable for the operation of the atmospheric pressure ionization source of the mass spectrometer.

Here we present the results of the experiments with the off-line implementation of the Edman degradation. Angiotensin I (DRVYIHPFHL, monoisotopic mass 1295.7 Da) was used as a model peptide. The reaction protocols were optimized to make it compatible with the on-line setup. The products of the reactions were analyzed using HPLC, which also helped in removing the pyridine present in the reaction mixture. Pyridine is one of the main components used in the Edman

degradation protocol and because of its high proton affinity it dampens the peptide ionization in the ESI source, even if present in minute amounts. The results of these analyses are shown in Figs. 5(A)–5(C). Figures 5(A) and 5(B) show chromatograms of the products of the first and second stages of the reaction, respectively; and Fig. 5(C) contains mass spectra of the fractions, in which triply charged species are present. The first stage of the reaction is PITC binding to the N-terminus of the peptide under elevated temperature conditions. The duration of the incubation was varied from 0 to 30 min. As shown in Fig. 5(A), 10-min incubation with PITC at 45 °C is sufficient to completely modify angiotensin I (fraction 3 contains modified peptide PTC-angiotensin I (phenyl thiocarbamoyl)). To further simplify the reaction

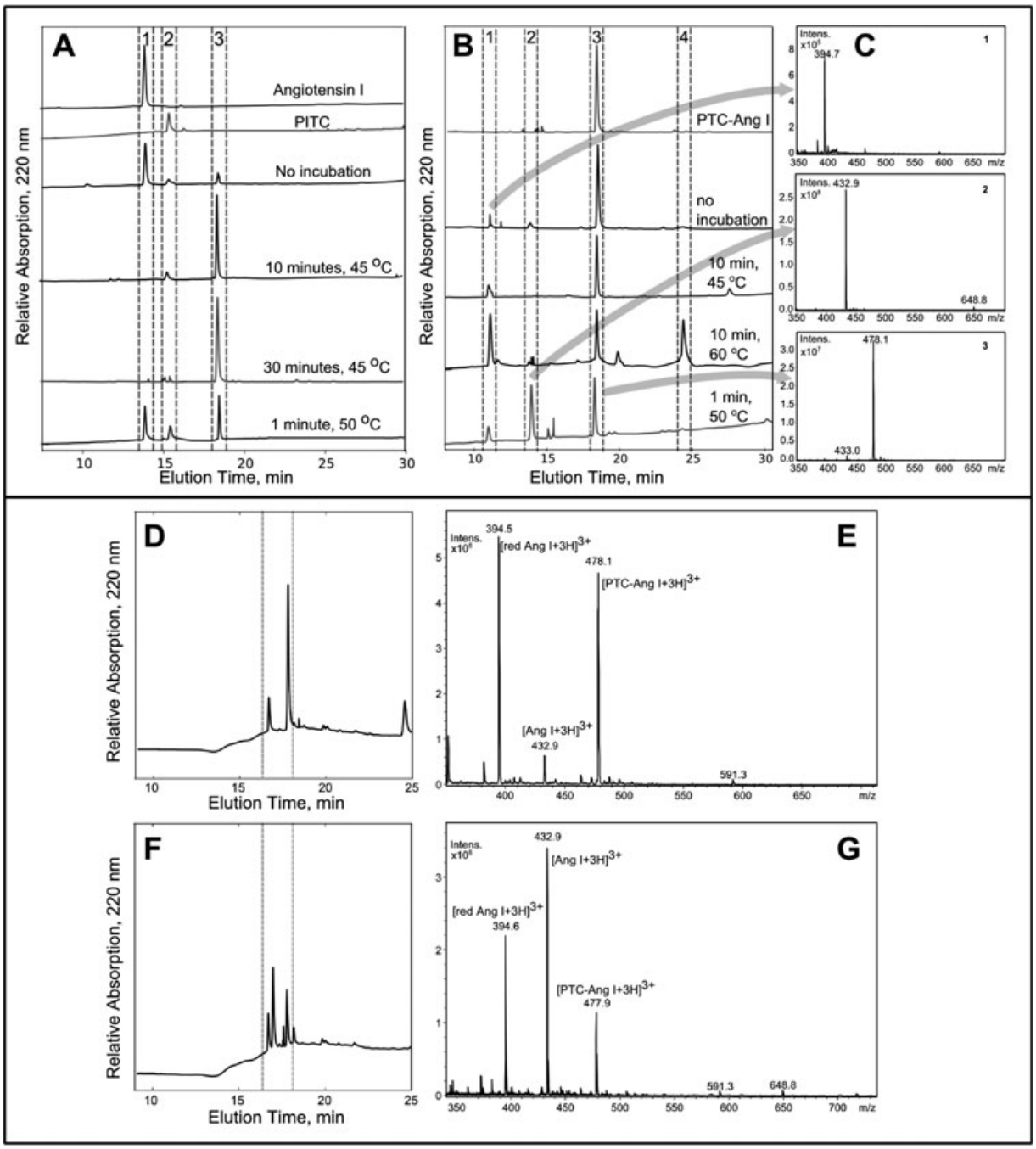

Figure 5. Optimization of the incubation conditions for Edman degradation. LC/MS analyses of the reagents and reaction products: (A) Chromatograms corresponding to the first stage, PITC binding to the N-terminus of angiotensin I. (B) Chromatograms corresponding to the second stage, cleavage of the N-terminal residue. (C) Mass spectra of the fractions collected after LC. LC/MS analyses of products of Edman degradation eluted in a single fraction in step gradient. (D, E) Chromatogram and mass spectrum after long incubation (30 min with PITC, 10 min with TFA at 45 °C). (F, G) Chromatogram and mass spectrum after short incubation (1 min with PITC, 1 min with TFA at 50 °C).

protocol, the incubation time was shortened to 1 min and the temperature was raised to 50 °C. Under these conditions, up to 50% of the full-length peptides were modified with PITC. The cleavage of the N-terminal residue was performed by addition of an equal volume of TFA to the reaction mixture with further incubation. The incubation for 30 min at 45 °C was insufficient for complete transformation of the peptide into its reduced form. Further increase in the temperature to 60 °C increased the yield of the reduced peptide (fraction 1 in Figs. 5(B) and 5(C)). However, the incubation at elevated temperature caused the generation of the reaction byproducts which are marked as fraction 4 in Fig. 5(B). These byproducts have presumably low ionization efficiency in ESI since they were not identified using MS. To check the possibility of rapid degradation, the reaction mixture after 1-min incubation with PITC was dissolved in an equal volume of TFA and incubated for 1 min at 50 °C. The yield of the reduced peptide was comparable to the longer incubation periods at 45 °C.

The yields of the products of Edman degradation in the case of long and short incubation periods were measured using step-wise gradient HPLC, in which all the peptides were collected in a single fraction shown by dashed lines in Figs. 5 (D) and 5(F). Expectedly, for longer incubations (Fig. 5(D) and 5(E)) the amount of unmodified angiotensin I was lower, while for the shorter ones (Fig. 5(F) and 5(G)) it was the major component in the collected LC fraction. However, in both cases the peaks in the mass spectra corresponding to unmodified (full-length) and reduced-length peptides have comparable heights. Hence, we believe that prolonged incubation periods are not necessary to measure unambiguously the masses of both full-length and reduced-length peptides. Note also that for on-line implementation it is essential that each peptide arrives at the mass analyzer at the same time as its reduced counterpart. The design of the on-line setup is still under development and may not incorporate the additional LC column after the reactor chamber.

CONCLUSIONS

In the present work a new three-dimensional approach combining the AMT-tag strategy with partial chemical degradation for bottom-up proteomics was proposed. It can be used for unambiguous MS/MS-free peptide identification by employing novel high-performance and throughput MS technologies such as LC-HRT. The results of *in silico* studies on yeast and human tryptic peptides, as well as yeast peptides with modifications, demonstrated both possible benefits and constraints of this extended AMT approach. The feasibility of rapid Edman degradation with the reaction rate compatible with on-line implementation of the proposed method was also shown experimentally.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (Grant No. 11-04-00515) and Prot-HiSPRA Project No.282506 funded by the 7th Framework Program of the European Commission. The authors also thank Prof. Albert T. Lebedev from Moscow State University for useful comments and discussions on the results.

REFERENCES

[1] R. Aebersold, M. Mann. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198.
[2] W. P. Blackstock, M. P. Weir. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **1999**, *17*, 121.
[3] J. R. Yates III. Database searching using mass spectrometry data. *Electrophoresis* **1998**, *19*, 893.
[4] J. R. Yates III. Mass spectrometry. From genomics to proteomics. *Trends Genet.* **2000**, *16*, 5.
[5] D. Fenyö, J. Qin, B. T. Chait. Protein identification using mass spectrometric information. *Electrophoresis* **1998**, *19*, 998.
[6] A. Shevchenko, O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie, M. Mann. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 14440.
[7] O. N. Jensen, A. Podtelejnikov, M. Mann. Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1371.
[8] J. Eng, A. L. McCormack, J. R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.
[9] D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551.
[10] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3*, 958.
[11] R. Craig, R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466.
[12] J. Allmer. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteomics* **2011**, *8*, 645.
[13] A. Michalski, J. Cox, M. Mann. More than 100 000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res.* **2011**, *10*, 1785.
[14] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N. G. Ahn, W. M. Old. Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomics studies. *J. Proteome Res.* **2011**, *9*, 4152.
[15] T. P. Conrads, G. A. Anderson, T. D. Veenstra, L. Pasa-Tolić, R. D. Smith. Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* **2000**, *72*, 3349.
[16] C. C. Wu, M. J. MacCoss. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther.* **2002**, *4*, 242.
[17] R. D. Smith, G. A. Anderson, M. S. Lipton, L. Pasa-Tolić, Y. Shen, T. P. Conrads, T. D. Veenstra, H. R. Udseth. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2002**, *2*, 513.
[18] T. Baczek, R. Kaliszan. 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.
[19] O. V. Krokhin, S. Ying, J. P. Cortens, D. Ghosh, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, J. A. Wilkins. Use of peptide retention time prediction for protein identification by off-line reversed-phase HPLC-MALDI MS/MS. *Anal. Chem.* **2006**, *78*, 6265.

[20] C. T. Mant, T. W. Lorne Burke, J. A. Black, R. S. Hodges. Effect of peptide chain length on peptide retention behaviour in reversed-phase chromatography. *J. Chromatogr.* **1988**, *458*, 193.
[21] M. Gilar, H. Xie, A. Jaworski. Utility of retention prediction model for investigation of peptide separation selectivity in reversed-phase liquid chromatography: impact of concentration of trifluoroacetic acid, column temperature, gradient slope and type of stationary phase. *Anal. Chem.* **2010**, *82*, 265.
[22] A. V. Gorshkov, I. A. Tarasova, V. V. Evreinov, M. M. Savitski, M. L. Nielsen, R. A. Zubarev, M. V. Gorshkov. Liquid chromatography at critical conditions: comprehensive approach to sequence-dependent retention time prediction. *Anal. Chem.* **2006**, *78*, 7770.
[23] A. V. Gorshkov, V. V. Evreinov, I. A. Tarasova, M. V. Gorshkov. Applicability of the critical chromatography concept to proteomics problems: Dependence of retention time on the sequence of amino acids. *Polym. Sci. B* **2007**, *49*, 93.
[24] I. A. Tarasova, A. V. Gorshkov, V. V. Evreinov, C. Adams, R. A. Zubarev, M. V. Gorshkov. Applicability of the critical chromatography concept to proteomics problems: Experimental study of the dependence of peptide retention time on the sequence of amino acids in the chain. *Polym. Sci. A* **2008**, *50*, 309.
[25] R. Kaliszan, T. Baczek, A. Cimochowska, P. Juszczyk, K. Wiśniewska, Z. Grzonka. Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *Proteomics* **2005**, *5*, 409.
[26] N. Pfeifer, A. Leinenbach, C. G. Huber, O. Kohlbacher. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinformatics* **2007**, *8*, 468.
[27] K. Petritis, L. J. Kangas, P. L. Ferguson, G. A. Anderson, L. Pasa-Tolić, M. S. Lipton, K. J. Auberry, E. F. Strittmatter, Y. Shen, R. Zhao, R. D. Smith. Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.* **2003**, *75*, 1039.
[28] P. Edman. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **1950**, *4*, 283.
[29] P. Klemm, in *Methods in Molecular Biology: Proteins*, (Ed: J. M. Walker). The Humana Press Inc., Clifton, **1984**, pp. 243–254.
[30] B. T. Chait, R. Wang, R. C. Beavis, S. B. Kent. Protein ladder sequencing. *Science* **1993**, *262*, 89.
[31] M. Bartlet-Jones, W. A. Jeffery, H. F. Hansen, D. J. Pappin. Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 737.
[32] G. Van der Rest, F. He, M. R. Emmett, A. G. Marshall, S. J. Gaskell. Gas-phase cleavage of PTC-derivatized electrosprayed tryptic peptides in an FT-ICR trapped-ion cell: mass-based protein identification without liquid chromatographic separation. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 288.
[33] F. L. Brancia, A. Butt, R. J. Beynon, S. J. Hubbard, S. J. Gaskell, S. G. Oliver. A combination of chemical derivatisation and improved bioinformatic tools optimises protein identification for proteomics. *Electrophoresis* **2001**, *22*, 552.
[34] K. S. Sidhu, P. Sangvanich, F. L. Brancia, A. G. Sullivan, S. J. Gaskell, O. Wolkenhaue, S. G. Oliver, S. J. Hubbard. Bioinformatic assessment of mass spectrometric chemical derivatisation techniques for proteome database searching. *Proteomics* **2001**, *1*, 1368.
[35] Available: http://www.uniprot.org/.
[36] A. A. Goloborodko, L. I. Levitsky, M. V. Ivanov, M. V. Gorshkov. Pyteomics – a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J. Am. Soc. Mass Spectrom.* **2012**. DOI: 10.1007/s13361-012-0516-6.
[37] Available: http://pypi.python.org/pypi/pyteomics.
[38] Available: http://pypi.python.org/pypi/pyteomics.biolccc.
[39] A. N. Vilkov, C. M. Gamage, A. S. Misharin, V. M. Doroshenko, D. A. Tolmachev, I. A. Tarasova, O. N. Kharybin, K. P. Novoselov, M. V. Gorshkov. Atmospheric pressure ionization permanent magnet Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2007**, *18*, 1552.
[40] S. Gay, P.-A. Binz, D. F. Hochstrasser, R. D. Appel. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis* **1999**, *20*, 3527.
[41] C. F. Klitzke, Y. E. Corilo, K. Siek, J. Binkley, J. Patrick, M. N. Eberlin. Petroleomics by ultrahigh-resolution time-of-flight mass spectrometry. *Energy Fuels* **2012**, *26*, 5787.
[42] V. I. Babushok, I. G. Zenkevich. Retention characteristics of peptides in RP-LC: Peptide retention prediction. *Chromatographia* **2010**, *72*, 781.
[43] T. Y. Perlova, A. A. Goloborodko, Y. Margolin, M. L. Pridatchenko, I. A. Tarasova, A. V. Gorshkov, E. Moskovets, A. R. Ivanov, M. V. Gorshkov. Retention time prediction using the model of liquid chromatography of biomacromolecules at critical conditions in LC-MS phosphopeptide analysis. *Proteomics* **2010**, *10*, 3458.

