Published in final edited form as:

Anal Chem. 2012 May 15; 84(10): 4383–4395. doi:10.1021/ac203337s.

Defining intact protein primary structures from saliva: A step toward the human proteome project

F. Halgand^{1,*}, V. Zabrouskov², S. Bassilian¹, P. Souda¹, J.A. Loo³, K.F. Faull¹, D.T. Wong⁴, and J.P. Whitelegge¹

¹The Pasarow Mass Spectrometry Laboratory, NPI - Semel Institute for Neuroscience and Human Behavior, David Geffen School of Medicine, University of California Los Angeles, 760 Westwood Plaza, Los Angeles, CA 90024

²Thermo Scientific Corporation, 355 River Oaks Parkway, San Jose, California 95134

³Department of Chemistry and Biochemistry and Department of Biological Chemistry, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095

⁴School of Dentistry, University of California Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095

Abstract

Top-down mass spectrometry has been used to investigate structural diversity within some abundant salivary protein families. In this study we report the identification of two isoforms of protein IB8a that was best satisfied by including a mutation and a covalent modification in the C-terminal part, and the assignment of a sequence of a previously unreported protein of mass 10433 Da. The final characterization of Peptide P-J was achieved and the discovery of a truncated form of this peptide was reported. The first sequence assignment was done at low resolution using a hybrid quadrupole time of flight instrument to quickly identify and characterize proteins and data acquisition was switched to FTICR for proteins that required additional sequence coverage and certainty of assignment. High-resolution and high mass accuracy mass spectrometry on a FTICR-MS instrument combined with ECD provided the most informative datasets, with the more frequent presence of 'unique' ions that unambiguously define the primary structure. A mixture of predictable and unusual post-translational modifications in the protein sequence precluded the use of shotgun-annotated databases at this stage, requiring manual iterations of sequence refinement in many cases. This led us to propose guidelines for an iterative processing workflow of MS and MSMS datasets that allow researchers to completely assign the identity and the structure of a protein.

Keywords

Salivary proteins; Mass spectrometry; Top-down proteomics; Protein sequence polymorphisms

Introduction

Saliva is a body fluid involved in numerous biological processes such as digestion, lubrication and protection of teeth, and the non-specific immune protection of the mouth. All of these biological functions involved several salivary protein families that have now been

* Author to whom correspondence should be sent: Dr Frédéric Halgand., Plate-forme Protéomique Biogenouest, Bâtiment 24 - Campus de Beaulieu - 35042 Rennes cedex FRANCE, frederic.halgand@univ-rennes1.fr, Phone: +33 (0)2 2323 5283 Fax: +33(0)2 2323 5282.

studied for more than 20 years. Recently, because it has been shown that the protein content of saliva reflects human health status ^{1, 2} this biological fluid has been viewed as a potential source of disease biomarkers and therapeutic targets. To achieve this goal, it is necessary to determine the biochemical composition and to describe the qualitative and quantitative variations of this body fluid. To that purpose different proteomics approaches have been used to characterize salivary proteins that are produced along their secretion pathways ³⁻⁵. Such approaches have also been useful for characterizing quantitative changes of diurnal saliva protein composition ⁶, inter-individual variability ⁷ and to elucidate salivary protein secretion and transit ^{8, 9} or for the discovery of novel endogenous peptides ^{5, 10}. This strategy has also been shown to be successful in correlating changes in the expression levels of mucins and calgranulins with pathological processes like cancer ¹¹⁻¹³ and with the finding of five potential biomarkers of oral squamous cell carcinoma ¹⁴.

Salivary proteins have also been studied from the genetic ¹⁵ and structural biology points of view, with the idea of building an exhaustive inventory of their structures and post translational variations that are revealed by conventional techniques ¹⁶⁻²⁰. Recent progress in this field has allowed researchers to propose that low phosphorylation levels of salivary protein could be correlated with autism disorders ²¹ and pre-term newborns ²². However, the structural characterization of salivary proteins remains a challenging task, due to a number of properties that make it difficult to study. First is the presence of protein super families, such as the abundant proline-rich proteins (PRPs) or histatins, whereby many different protein products derive from a limited number of genes ¹⁶. This leads to a high degree of sequence homology, exacerbated by the presence of a high rate of repetitive sequence units. Second, most of the time, these proteins are proteolytically processed at both the N- and C-termini so that the use of trypsin for further proteolysis is, ineffective most of the time, since all authorized trypsin cleavages were already processed. Third, single nucleotide polymorphisms, alternative splicing and post-translational modifications all considerably increase the heterogeneity of human saliva proteins as well. Therefore, if the objective is to catalogue the presence of different gene products this goal can be easily accomplished, but if the objective is to characterize exactly which protein forms are present and to fully determine their primary structures the task becomes considerably more complex. The difficulties in assessing primary structures and properties of salivary proteins have been discussed in two recent reviews ^{23, 24}. The efforts made and the advances achieved in this field along with detailed descriptions of artifacts, and unexpected or unknown mass increments, that are commonly encountered when studying salivary protein by MS are reported in a recent article ²⁴. Such difficulties in unraveling saliva protein heterogeneity and polymorphisms are exemplified, for example, by the recent assignment of Peptide P-J structure ⁸, or with the requirement for definitive sequence identification or the 10434, 23462 and 29415 Da salivary proteins or for slight molecular weight changes of as little as 1 Da observed in various cases ²⁵.

With this in mind, we proposed that combination of intact mass tag (IMT's) measurements with top-down MSMS experiments could be a complementary approach that would be helpful for characterizing unknown compounds and subtle changes. Indeed, top-down proteomics with software analysis proved to be a powerful tool for the determination of the match or mismatch of peptide fragments to a database containing sequences and post translational modifications ²⁶⁻³⁴. Such an approach would be complementary to computational identification of protein using their accurate IMT's only ³⁵. In the current study, experiments for the characterization of salivary proteins were conducted on a qTof instrument for primary identification while ambiguities related to subtle changes could only be solved by using ultra-high resolution and highly accurate MSMS experiments on a FTICR instrument. The combination of these techniques was shown to be extremely powerful for identifying new or unknown compounds and for revealing the complexity of

the human proteome. Pitfalls encountered with salivary proteins data handling in this study led us to propose a semi-automated workflow analysis to overcome these difficulties and achieve reliable identification.

Experimental Procedures

Material

Chemicals—All solvents (HPLC grade and otherwise), buffers and reagents (guanidinium HCl, acetonitrile, protease-inhibitor cocktail and trifluoroacetic acid (TFA)) used were purchased from Sigma Aldrich.

Sample collection—Parotid saliva samples were obtained from 5 donors that were recruited for the Human Salivary Proteome Project. Donors were in good health and exhibited normal salivary function. Parotid saliva secretions were harvested using a saliva collector³⁶ fitted with a sterile 100 μ L pipette tip. Stimulation of the salivary glands was provided by repeated topical application of a mild solution of citric acid (2%) to the dorsal surface of the tongue. Care was taken to keep the acid solution away from the collection area. Collection volumes were 500–2000 μ L/donor. The collected samples were centrifuged ($2,600 \times g$, 15 min, 4°C), and re-centrifuged (20 min) if the supernatant did not appear to be clear. Supernatants were transferred to new containers to which aprotinin (1 μ L/ml of saliva, 10 mg/ml), sodium orthovanadate (3 μ L/ml of saliva, 400 mM) and phenylmethyl sulfonyl fluoride (10 μ L/ml of saliva, 10 mg/mL) were promptly added prior to storage at -80°C .

Sample fractionation by LCMS+—Individual parotid saliva samples (between 500 μ L and 1 mL) were dried by centrifugal evaporation, re-dissolved in aqueous guanidinium-HCl (6 M, 100 μ L), centrifuged ($10,000 \times g$, 5 minutes, room temp) and fractionated by liquid chromatography coupled to mass spectrometry with on-line fraction collection according to a procedure previously described³⁷. Each sample was loaded onto a reverse-phase HPLC column (40 $^{\circ}\text{C}$, PLRP/S 5 μ m, 300 Å, 2.1 mm $\times$ 150 mm, Agilent Technologies) previously equilibrated in 95% A, 5% B (A, 0.1% TFA in water; B, 0.1% TFA in CH₃CN) and eluted with a compound linear gradient from 5% B at 5 min after injection, through 20% B at 10 min, 50% B at 70 min, and 90% B at 90 min. The eluent was passed through a UV detector (280 nm) prior to a liquid-flow splitter with fused silica capillaries to transfer liquid to the ESI source (50 cm) and the fraction collector (25 cm). Fractions (1 min) were collected into microcentrifuge tubes and stored at -20°C for further analysis. The IonSprayTM source was connected to a triple quadrupole mass spectrometer (API III+, Applied Biosystems) tuned and calibrated as previously described³⁸ scanning from m/z 600–2300 (orifice voltage ramped with m/z from 60–120, 6 sec/scan). Data were processed using MacSpec 3.3, Hypermass and BioMultiview 1.3.1 software (Applied Biosystems).

Electrospray Ionization Mass Spectrometry—Primary identification of salivary protein sequences were obtained by acquiring nano-ESI CID MSMS spectra of compounds previously separated using chromatography (Q-Star XL, Applied Biosystems, CA, USA). Instrument parameters were optimized to achieve the best signal to noise ratio. Collision energy as well as collision gas pressure in the collision cell were optimized to achieve the best fragmentation yield of the parent ion. Typical collision energy was approximately 40 eV to 90 eV for low ($\sim$ 4 kDa) to higher molecular weight proteins ($\sim$ 17 kDa), respectively, and gas pressure was set to 3 to 5, accordingly. Proteins were fragmented using various charge state to obtain complementary information. Calibration was achieved in the tandem MSMS mode using commercially-available Glu Fibrinopeptide (Sigma Aldrich). Typical mass errors on MS and MSMS data were approximately 30 to 50 ppm, respectively, except when there were poor ion statistics where mass accuracy was closer to 100 ppm. MSMS

data were then both manually and software analyzed with Prosight PTM (<https://prosigthptm.scs.uiuc.edu>). Data mining with Prosight PTM was performed using a 50-ppm mass tolerance for experiments performed on the Qstar except when mentioned.

High-resolution top-down mass spectrometry and tandem mass spectrometry

—These experiments were performed as previously described³² on a 7 Tesla hybrid linear ion trap-FTICR mass spectrometer (LTQ-FT Ultra, Thermo Fisher Corporation, San Jose, CA) fitted with an off-line nanospray source. Ion transmission into the linear trap and then to the FTICR cell were automatically controlled to a 2×10^6 ion count target for both the full scan- and MS²-FTICR experiments. The m/z resolving power of the FTICR mass analyzer was set to either 100,000 or 750,000 (defined by $m/\Delta m_{50\%}$ at m/z 400). Individual charge states of the multiply-protonated protein molecular ions were selected for isolation and collision-induced dissociation in the linear ion trap followed by the detection of the resulting fragments in the FTICR cell. For the FTICR-MSMS experimental parameters were chosen to fragment the full isotopic mass of the most abundant charge state in order to increase detection of product ions while checking for homogeneity of the selected peak. For the collisionally induced dissociation (CID) studies, the precursor ions were activated using 30 to 35 % normalized collision energy at the default activation q -value of 0.25. Additional studies were conducted in which the precursor ions were guided to the FTICR cell and further fragmented using electron capture (ECD) using the following instrument settings: 5 to 10 % normalized collision energy, 50 ms delay and 10 msec duration. For the infrared multiphoton dissociation (IRMPD) experiments, instrument settings were: 50% normalized collision energy, 50 ms delay and 20 msec duration. In both cases, the fragmentation efficiency was optimized to maximize the product ion signal intensity.

FTICR spectra, from an average of 50 – 500 transient signals, were examined with a combination of manual and automatic procedures. Monoisotopic mass lists ($s/n = 1.1$, fit 0%, remainder 0%, averaging table set to averaging) were prepared using XtractAll (Xcalibur 2.0, Thermo Fisher, Bremen, Germany). Prosight PTM (<https://prosigthptm.scs.uiuc.edu>) software was used with a threshold of 15 ppm and the delta mass feature deactivated, with custom post-translational modifications as required. Interpretation was a manual, iterative process as different sequences and post-translational modifications were independently tested to maximize the number of product ions matched. Nomenclature for assignment of peptide/protein ions was according to Roepstorff and Fohlman³⁹. Pscore values reflect the match of the proposed primary structure with the peaklist data; the lower the score the higher the confidence in the proposed sequence is⁴⁰. We also used a manual Pscore that similarly reflects the confidence of the data interpretation, but relies on the masses of product ions that matched the sequence, updated with product ions manually identified in the MS/MS spectra. Manual Pscores were calculated in Prosight PTM using the Manual Single Protein Mode. Extracted peak lists are provided in the Supplemental materials.

Results

Studied proteins were first fragmented by CID MSMS using a standard quadrupole-time of flight instrument (Qstar XL, Applied Biosystems) for primary identification. MSMS datasets were used for protein identification and detection of post-translational modifications in the databases. When ambiguities and/or subtle changes were noted, FTICR-MS and MSMS experiments (CAD, ECD, IRMPD) were carried out at high resolution and mass accuracy. All information about identified proteins reported in this article are summarized in Table 1.

Identifications of peptide P-B, P-D, P-H, P-E and IB-1 protein (Table 1 of supplementary material)

These small, well known, saliva proteins were identified without ambiguity. For these proteins, the IMT's were in agreement with the expected calculated masses and the MSMS data were in good agreement with the sequences retrieved from Swiss-Prot entries. See Table 1 of supplementary material. Manual analysis of the MSMS spectra of these species did not reveal any noticeable modifications or variants. However, during our study we detected a new truncated form of peptide P-D in one of the individuals studied (Individual 3). Analyses of the CID MSMS datasets were performed using the peptide P-D sequence retrieved from Swiss-Prot (P10163). Automated searches with Prosight PTM clearly showed that the modification was located in the C-terminal portion as shown by the lack of y ion matching the expected sequence. Data fitting allowed us to propose that this peptide corresponded to a new Peptide P-D form that was missing the final 16 C-terminal amino-acids. The hypothesized sequence assignment was confirmed by manual analysis of the MSMS data and assignment of the sequence from residues 169 to 222 of the P10163 entry was in agreement with the experimentally-measured IMT of 5268.73 ± 0.5 Da (Calculated averaged mass = 5268.88 Da).

Identification of Histatin 1, Peptide P-F and P-J (Table 2 of supplementary material)

Histatin 1 was directly identified based on its sequence collected from Swiss-Prot. The MSMS data fitted with the reported sequence (P34084) spanning from residues 20 to 57 with the presence of one phosphorylation. The presence of one phosphorylation was further confirmed with the observation of a +79.979 Da mass increment and a -97.9767 Da losses on several fragment ions. However, we did not succeed in clearly identifying the phosphorylation site. Based on CID MSMS data we could only propose that Tyr30 and Ser21 are probably both phosphorylated. Indeed, if fragment ions b28 and b29 supported the presence of a phosphorylation on Ser21 they unfortunately also matched with the masses of internal fragment ions leading to an ambiguous interpretation. In contrast the observation of b24 and b25 ions that were shown to be unique and allowed us to locate the phosphorylation site to Ser21 without ruling out the possibility that Tyr30 could also be modified. A unique ion is an unambiguously assigned product ion that cannot be explained by internal fragmentation, or loss of water, ammonia etc. However data interpretation should be cautious since it was demonstrated that the phosphoryl group can move from one site to another during collisionally activated dissociation process^{41, 42}. Consequently, additional ECD MSMS experiments are required to distinguish whether Histatin 1 is a mixture of two phosphorylation products. (see Table 2 of the supplementary section).

Peptide P-F and P-J identifications

Measurements of Peptide P-F (Theoretical $MW_{av} = 5843.5 \pm 0.5$ Da) and Peptide P-J (Theoretical $MW_{av} = 5943.5 \pm 0.5$ Da) IMT's were expected to provide a direct identification of these peptides. In fact, Peptide P-F shares a high sequence homology (~95 %) with Peptide P-J since they belong to the same gene product sequence (P02812), which also contains tandem sequence repeat units (Peptide P-J and Peptide P-F sequences corresponding to amino-acids 79–139 and 265–325, respectively). Peptides P-F and P-J only differ by 3 amino acids which are located in the C-terminal portion of the sequence (Peptide P-F sequence: SPPGKPQGPP---PPQGGSKSRS **A**; Peptide P-J sequence: SPPGKPQGPP---PPQGDNKSRS **S**). See Figure 1. In order to determine whether top down MSMS using a standard Q-ToF instrument can rapidly provide a reliable identification of these compounds the CID MSMS was acquired. Unfortunately, the MSMS datasets we obtained did not allow a complete discrimination between these two compounds (Figure 1A). This was related to the absence of detection of unique ion matching the C-terminal part where sequence

differences (3 amino acids) are located. Additional investigations were carried out by acquisition of MSMS data using CID and IRMPD fragmentation using on FTICR-MS instrument (Figure 1B). However, neither CID nor IRMPD MSMS ($R = 100\,000$) permitted discrimination Peptide P-F and P-J. This was attributed to the presence of numerous internal fragment ions originating from multiple collisional events and having close raw chemical compositions to the normal b or y product ions, leading then to misinterpretation. Distinguishing between Peptide P-F and P-J was finally achieved by generating highly resolved and accurate ECD MSMS data. With ECD the presence of large c ions (Figure 1C) and unique fragments that could only come from c or z⁺ ions (data not shown) with a RMS of less than 5 ppm finally allowed us to prove that protein at 5841.98 ± 0.6 Da mass in average was Peptide P-F. The sequence reported here for Peptide P-J is in full agreement with results obtained by Cabras et al.⁸.

The study of Peptide P-F also revealed that standard identification of proteins using trypsin cleavage can result in the detection of false positives. Peptide P-F is the precursor of Peptide F, a small peptide of 2.1 kDa that is released by tryptic cleavage of the C-terminus at Lys 290 (P02812). This cleavage was observed despite the fact that tryptic cleavage is not usual due to the presence of a proline residue at n+1 position^{20, 43}. Nonetheless, this particular tryptic cleavage has been reproduced *in vitro* and was confirmed in this experiment (Experimental measured mass = 2495.5 ± 0.25 Da, in agreement with the calculated molecular weight = 2495.78 Da). The identity of Peptide P-F was then fully confirmed by MSMS sequencing on the Qstar instrument with a Pscore of 2.81×10^{-62} (Data not shown).

Identification of Protein II-2, Protein IB8a and of the unknown protein of 10433 Da (Table 3 of supplementary material)

Protein II-2—In the literature protein II-2 is reported to have an average molecular weight of 7608 Da in close agreement with those deduced from our measurements (7607.4 ± 0.8 Da < $IMT_{(av.)} > 7608.5 \pm 0.8$ Da)¹⁶. Sequence information retrieved from the Swiss-Prot entry (C38355) did not match with the calculated molecular weight of the full length protein (7642.38 Da) which was 34.44 Da heavier, on average, than the experimental masses. Knowing that many salivary proteins have a pyroglutamate (−17.0265 Da) on the N-terminus²³ we first considered this possibility but this would give a calculated mass of 7625.35 Da that was still 17.9 and 16.8 Da heavier than our experimentally-measured masses. Literature searches revealed a variant of this protein lacking a proline residue in position 39⁴⁴ which would lead to a calculated molecular weight of 7545.249 Da. Addition of a pyroglutamate modification in N-terminus to this latter species would lead to an average calculated molecular weight of 7528.223 Da which is now 79.18 and 80.28 Da lighter than the experimental measured masses. This mass difference could be attributed to the presence of a phosphoryl group, giving a theoretical MW of 7607.7192 Da which is in agreement with our measured experimental molecular weights.

Top down MSMS spectra were then recorded for several charge states of the protein II-2 using a quadrupole time of flight instrument (Qstar). The presence of protein II-2 lacking a proline residue in position 39, bearing a pyroglutamic acid in N-terminus and a phosphoryl group on serine 8 (form 1) was confirmed (Figure 2A). MSMS datasets analysis with Prosight PTM also suggested that protein II-2 could be a mixture of two forms..

To demonstrate whether protein II-2 was a mixture or not, highly accurate and highly resolved MSMS experiments (CID, IRMPD, and ECD) were recorded on a FTICR-MS instrument. Results obtained at 100 000 resolution at m/z 400 showed that protein II-2 was a mixture of the two forms. Notably, the presence of these two forms was observed in each of the individual studied. The best fit was achieved with the ECD data for protein II-2 form 1, with the detection of unique and large fragment ions (see Figure 2C). In that case, structural

characteristics of the first isoform of protein II-2 fitted with the deletion of the proline in position 39, the presence of a pyroglutamate in N-terminus, and a phosphorylation located on serine in position 8. In contrast, ECD data obtained for the second form of protein II-2 only allowed identifying the presence of a pyroglutamic acid in N-terminus as well as the presence of the proline 39 that was found deleted in protein II-2 isoform 1. However IRMPD data (Figure 2B) led to the unequivocal identification of the pyroglutamic modification in N-terminus, the presence of the proline 39, and the detection of a dehydroalanine residue in position 8. The detection of unique ions in the IRMPD dataset bracketing the proposed modifications (presence of the proline 39 and of the dehydroalanine residue) was of crucial importance to validate the presence of these modifications since ECD data only yielded one unique *c* ion (C73). To conclude protein II-2 seemed to be a mixture of two forms with monoisotopic experimental molecular weights of 7603.7787 Da (theoretical: 7603.7048; $\Delta m = 9.7$ ppm) and 7602.7806 Da (theoretical: 7602.78095; $\Delta m = 0.05$ ppm) for isoforms 1 and 2, respectively. Little decrease in mass accuracy observed for the form 1 of protein II-2 was attributed to the overlapping of the ^{13}C isotope of form 2 with the monoisotopic peak of form 1. It is clear that dehydroalanine residue observed for isoform 2 of protein II-2 could not come from a loss of the phosphoryl group found on isoform 1 since these two proteins differed by proline 39 (97 Da). However, it is not clear whether the presence of the dehydroalanine residue in protein II-2 resulted from an, *in vivo*, beta elimination of the phosphoryl group or from an artifactual loss during our experiment from its phosphorylated form. Surprisingly, the phosphorylated form 2 of protein II-2 (expected MW of 7682.7169) was never observed in any of our samples. This result was striking since other identified phosphorylated proteins were shown to keep their phosphoryl groups intact. Additional data supporting our interpretation are given in the Supplementary Section (See Tables 5 to 8).

Protein Ib8a—According to the literature ⁴⁴, protein Ib8a has an average molecular weight of 11894 ± 1.1 Da, and has been shown to contain a glucose molecule. The translated DNA sequence mentioned in this article allowed the calculation of a theoretical molecular weight of 11714 ± 1.1 Da for Ib8a protein ⁴⁴. Taking into account the presence of the glucose molecule, the expected calculated mass was of 11876 Da, which was ~ 18 Da lighter than experimentally measured mass (Table 3). To explain this mass discrepancy top down MSMS experiments using CID (on both the qToF and the FTICR), and IRMPD and ECD on FTICR-MS were performed. Without allowing any chemical modifications, preliminary analysis of the MSMS data confirmed most of the sequence as reported in Stubbs's paper (P02812, AA 265–382); (see Figure 3). Nonetheless, the detection of only few *y* and *z* fragment ions led us to hypothesize that there could be modifications in the C-terminal part of the protein (Figure 3A). We therefore introduced different mass increments into various positions in the C-terminal part of the protein to match the measured IMT with the MSMS data. The MSMS data was in agreement with the insertion of a Q to H mutation in position 115 and glucose molecule on Serine 120. This was confirmed by the detection of unique and large *c* ions (*c*116 at m/z 11197.704; *c*117 at m/z 11268.782; *c*118 at m/z 11355.828; *c*119 at m/z 11483.912) and *z* ions (*z*62 at m/z 6154.158; *z*63 at m/z 6282.223; *z*65 at m/z 6497.343; *z*85 at m/z 8468.341). At this stage the calculated IMT for this species was still ~ 14 Da heavier than the observed molecular weight. Hypothesizing that this mass increment could be linked to the presence of a methyl group, MSMS data analysis suggested that the methyl group could either be located on the C-terminal carboxylic group or on the side chain of the penultimate arginine residue. This result was confirmed by manual analysis of MSMS data, with the detection of unique fragment ions, a high “manual” Pscore of 9.86×10^{-109} and a RMS of 2.7 ppm obtained from the ECD MSMS (Figure 3B). Detailed results are given in Table 9 of the Supplementary Section, and show fragment ion lists matching our findings,

and a sequence chart showing fragment ions matching our experimental MSMS data and IMT.

Unknown protein of ~ 10433.5 Da—An unknown protein of 10.4 kDa was recently reported in literature¹⁶ and the information gathered allowed it to be classified within the basic proline rich protein group. Following the same procedure as above we first attempted to identify this protein in databases using the automated procedure in Prosight PTM. Unfortunately this approach failed to provide identification. Stretches of the amino acid sequence obtained by this automated search were used to mine salivary protein sequences databases. This attempt failed too. Finally, manual analysis of all MSMS spectra were then performed which allowed the identification of several sequence tags that were used to further investigate salivary protein sequences. This approach allowed us to propose that this protein could be derived from the P04280 sequence entry. However, fitting the experimental data to the proposed sequence was unsatisfactory since most of the assigned fragment peaks were found to match with internal fragment ions. To reach a better agreement we confronted our MSMS datasets to amino acids sequences deduced from mRNA (K03205) using Clustal W software (<http://www.ebi.ac.uk/Tools/clustalw2/>). The best-fitting MSMS data was obtained with the amino acid sequence deduced from the first mRNA open reading frame (K03205) indicating that the 10.4 kDa protein corresponded to amino acids 98 to 198 of P04280 entry and did not contain any post translational modification. The high “manual” Pscore of 1.9012×10^{-146} with a RMS of 2.21 ppm (Pscore achieved with monoisotopic peak list was 1.99636×10^{-81}) which we obtained for this sequence, together with the assignment of unique fragment ions, strengthened the confidence in our data interpretation (see Figure 4). Additional supporting data, such as the peak list of ions matching our findings (Table 10) with sequence charts and unique ions, are provided in the Supplementary Material.

PRP1/PRP2, Db-f, Db-s and IB-6 proteins (Table 4 of supplementary material)

—Other examples of protein identifications concerned PRP1/PRP2, Db-f, and Db-s proteins that all derived from the sequence corresponding to the P02810 Swiss-Prot entry. All of these proteins were found with a pyroglutamic acid in the N-terminus. The protein of 15514.1 Da was assigned either to the diphosphorylated protein PRP-1 or PRP-2, but not Pif-s due to the presence of an aspartic acid in position 4. Phosphorylations were not precisely localized, but ECD experiment showed a high mass z^+ fragment ion ($z149$) and a $c51$ ion confirming the presence of two phosphorylations in the N-terminal portion of the protein. Experiments performed on the 13279.34 Da molecular weight protein confirmed the identity of the diphosphorylated Db-f variant of PRP-1 with the replacement of Q97 by the QGGQQQQGPPPPQGKPQGPPQQ sequence (See data for P02810 accession number). Similarly, only the first phosphoserine (Ser8) was localized without ambiguity. For the protein of ~ 17632.5 Da, the MSMS data obtained revealed the limitations of the MSMS capabilities of the qToF instrument. The data obtained were sufficient to identify the protein as a diphosphorylated Db-s protein but not to determine the location of phosphoryl groups. Protein IB-6 that was detected only for individuals 3 and 4 did not show any modification (data not shown). Information gathered for these proteins are summarized in Table 4 of the Supplementary Material section.

Summary of previous findings on Peptide P-C, PRP3 and Cystatin SN proteins

Our previous data obtained on Peptide P-C allowed us to reveal the presence of two new isoforms differing by ~ -0.94 Da and ~ + 0.98 Da, respectively. The lighter form was assigned to the presumed occurrence of an alternative splicing event leading to the replacement of the QQGPPP sequence by PRPPR. The heavier form was assigned to protein

sequence polymorphisms (PSP's) with replacement of glutamine 14 with a glutamic acid³³. Likewise, data recorded for cystatin SN permitted us to confirm the presence of PSP's with replacement of Proline 11 toward a Leucine residue. This modification was in agreement with the recently-proposed SNP's³². This work on cystatins was pursued and allowed describing in detail chemical and genetic polymorphisms³⁴. Finally top down MSMS experiments performed on a protein with molecular weight ~ 10 999 Da, was shown to be the diphosphorylated PRP-3 protein lacking the C-terminal arginine. In addition two PSPs were also found on PRP-3 protein lacking its C-terminal arginine with replacements of both aspartic residues in positions 4 and 50 toward an Asn. CID and ECD MSMS experiments also allowed localizing the two phosphorylation sites of PRP-3 protein lacking its C-terminal protein on Ser8 and Ser22 rather than Ser7 as expected from literature³². Information collected from these experiments are summarized in Table 4 of Supplementary Material.

Discussion

In this study we used an unbiased top down approach to decode the salivary proteome with the view not to only itemize a list of the proteins but also to embrace polymorphisms of these proteins. As stated in the Introduction section, the salivary proteome is particular in the sense that salivary proteins are represented both by acidic and basic proteins within a wide range of molecular weights in addition to particular structural properties (sequence homology, complete tryptic processing, genetic polymorphisms, etc) that made them challenging to achieve their extensive characterization. Using top down MSMS experiments performed on both a qTof and a FTICR instruments 17 major proteins from parotid saliva from 5 individuals were analyzed. While some of these proteins were easily identified, the pitfalls and difficulties encountered to fully characterize the others allowed us to determine essential points that needed to be considered for a complete and reliable description of protein covalent structure.

Experimental pitfalls

For primary structure assignment, the usual procedures are based on peptide tandem mass spectra and rely upon predicted peptides from genomic sequences in databases. Sequence polymorphisms can be taken into account for known or predictable post-translational modifications as well as sequence variations from alternative splicing through appropriate software solutions. However if such strategies are satisfactory for "bottom-up" approaches, unprejudiced data analysis is necessary for complete documentation of protein heterogeneity. This is critical for the characterization of subtle changes such as low-mass increment modifications (~ 1 Da) and characterization is further complicated when such alterations result in a mixture of (nearly) isobaric protein isoforms. To overcome this limitation the reliable assignment of fragment ions identities in MSMS spectra is mandatory. However, this task is difficult when internal (secondary) fragment ions are encountered in MSMS spectra. These fragments that arise from multiple MSMS collision events could lead to ambiguous interpretations due to chemical formulas which give masses that are close to or identical to, those from the primary fragment ions. To surmount this obstacle, the acquisition of highly accurate and highly resolved MSMS spectra using several fragmentation modes such as CID, ECD, and IRMPD, were necessary. Moreover, IRMPD and ECD experiments are also of particular interest when proteins are difficult to fragment, since these two collision modes can be coupled to give extra vibrational energy to the selected parent ion prior to ECD fragmentation, leading to increased fragmentation yield⁴⁵. Finally, in order to rule out ambiguous data interpretation due to the presence of internal fragment ions and to obtain complete protein identification the concept of unique ions was introduced. Unique ions are defined as primary b, y, c or z fragment ions whose masses did

not match any predicted mass of internal fragment including water and ammonia losses. Identification of these unique ions in the MSMS spectra allowed us to confirm portion of the complete sequence or even the entire protein sequence. In our study, we noticed that ECD was the collision mode that allowed us to obtain the highest number of unique fragment ions. This could be explained by the non ergodic character of the ECD fragmentation process in relationship with the low energy (~ 0.2 eV) of electrons used to cleave the N-C α of the amide backbone. Fragmentation of the peptidic backbone in which the intra-molecular energy randomization is slower than ECD cleavages allows obtaining an extensive fragmentation of the peptidic backbone with a very low probability of internal fragment ion production^{46, 47}. Charge neutralization events in the ECD process that considerably reduce the number of ions that could yield fragment ions were also observed, as previously reported^{46, 48}. The detection of *b* and *y* ions in the ECD MSMS spectra sometimes compensates for the lack of cleavage on the N-terminal side of proline residues during the sequence assignment process⁴⁹. In summary, the information retrieved from all of these collisional modes, recorded for several charge states, were found to be extremely complementary and assisted in achieving the most comprehensive data analysis.

From semi-automated treatment to a processing workflow to achieve the most complete data analysis

Data analysis of the MS and MSMS spectra rely on software allowing the calculation of peptides, proteins and fragment ion masses based on the determination of peak charge states. Such software (TRASH, Xtract, Modificomb/Prosight PTM and PC) have been greatly improved during the past years and were shown to be very helpful for data handling. This approach was described for human sub-proteome “shotgun” databases that incorporate known post-translational modifications and combinations thereof, thus permitting automated analysis of a subset of histone isoforms⁵⁰. Within this scope Prosight PTM software was used for interpretation of our “top-down” datasets in a semi-automated manner. Deconvolution software is essential since all searches are based on the production of mass lists. However, complete peak assignment is sometimes impossible. Lack of peak detection due to low ion signal statistics on ions (particularly for monoisotopic peak), incorrect charge state determination related to incomplete isotopomer profiles or mis-assignment of isotopomer profiles due to overlapping isotopic profiles, and competition in ion protonation and desorption are some of the difficulties encountered. These problems are amplified since analytical software cannot distinguish between internal and normal fragment ions based only on mass detection. Consequently, manual charge state and peak assignments are still required. Such an approach was necessary for the detection of novel isoforms that might be overlooked or potentially mis-assigned when multiple isobaric or isomeric isoforms are present, and for checking for the presence of fragment ions that were hypothesized to carry modifications. The software development issue has been recently addressed, and the Pevzner group has successfully identified proteins with modifications (post-translational modifications, insertions, and deletions) from top down experiments using a new computational tool based on spectral alignments^{51, 52}. However, further software development efforts are required. In this final section we will propose developments that appeared important to us to improve protein characterization. Ideally, software for the automated assignment of top-down data needs to consider all possible primary structure permutations at each residue, using some type of unbiased iterative process to refine the best match. To avoid massively complicated computation, it is suggested that this interpretation be split into two parts: identification, followed by unbiased primary structure assignment, with a mechanism to score each unique solution. In addition amino acid sequences deduced from genetic data (cDNA and mRNA) should be taken into account in order to incorporate known variants and isoforms originating from other open reading frames in order to increase the information coverage and consequently the amount of matching data. This is illustrated

both with Peptide P-C data and with a 10.4 kDa protein where only mRNA translated sequences allowed us to fully explain our experimental data. Except for the fact that genetic data and deduced sequences should be taken into account, the first processing step described in Figure 6 is identical to solutions present in Prosight PTM and PC software. To complete the process of MSMS spectrum analysis, however, a manual validation of identified peaks is required. Two steps have been shown to be essential for us to validate our data interpretation. First, differentiation of primary fragment ions from internal fragment ions is required. Experimental mass lists should be compared with theoretical *in silico* calculated fragments, taking into account not only the presence of b, y, c or z⁺ ions but also chemical losses such as ammonia or water as well as all the presence of internal fragments. In this step, unique ions will be labeled in the mass list and can be used to validate protein sequences and proposed modifications. Next, an isotopic profile of each identified ion should be determined and compared to its experimental counterpart. This step is similar to the one used in metabolomics studies⁵³. From the chemical formulas, the theoretical isotopic profiles of regular fragments should be calculated, allowing the comparison of these with their experimental counterparts. This process should reveal species with small changes in molecular weight of ~ 1 Da that lead to the overlapping of isotopic profiles and result in higher apparent abundances of chemical isotopes such as ¹³C or ¹⁵N³⁰. By repeating these steps, we should be able to unambiguously annotate MSMS spectra and label “confident peaks”. Then, these annotated peaks should be “removed” allowing simplification of the MSMS spectra and revealing the presence of unidentified peaks (peaks corresponding to overlapping clusters and to unidentified ones). Dedicated analyses of these peaks could then be conducted using an iterative process, proposing modifications based on literature and experimental data. Repeating steps 1 to 3 (Figure 6), we should be able to annotate most of the peaks allowing us to finally focus on unknown remaining peaks. The last purged MSMS spectrum should only show chemical noise. At this point the complete exploitation of data has been achieved.

Biology

From a biological point of view, our findings need further investigations in order to connect this new structural information with biological activities and genetic evidence. This is of importance when protein isoforms coexist *in vivo* (for example, Peptide P-C and protein II-2). Protein II-2 isoforms in particular display strikingly different post translational modifications. Isoform 1 that corresponds to the presence of a phosphate group on Serine 8 with the lack of proline 39 could be classified in the group of proteins playing a role in the calcium homeostasis or phosphate buffering in mouth⁵⁴. In contrast, isoform 2 that possesses a dehydroalanine residue in place of Serine 8 could belong to anti-bacterial and anti-fungal protein families. This could be in agreement with current knowledge showing that such modification is enzyme-dependent and displays anti-bacterial and anti-fungal activities^{55–57}. Further support in literature showed that dehydroalanine (Dha) is found in a number of proteins and nonribosomal natural products, and typically arises from post-translational modifications of serine or cysteine^{58–60}. Consequently, detection of this modification for protein II-2 should arise from an *in vivo* processing of serine 8. Such processes could be viewed as a way to increase the protein’s functional diversity. Meanwhile, further investigation of protein II-2 heterogeneity is required to fully assess the presence of isoform 2 in particular. However, the task is even more complicated when new proteins are characterized, as shown with the 10.4 kDa unknown protein, where information gathered from literature allowed classifying it as a basic PRP, based on its chromatographic behavior¹⁶.

Truncation is also another modification regularly encountered, as we observed in our study of peptide P-D and as reported in literature^{61, 62}. The role of these truncated forms of

peptides or proteins remain to be addressed. Finally, many measured masses were shown to present mass increments of about 0.5 Da to 1 Da higher than calculated. Such errors are too large for the mass accuracy range of our experiment. This could be attributed to the occurrence of Single Nucleotide Polymorphism such as D to N, Q to E, I/L to N, or in a more complex manner changes involving a combination of several mutations or chemical changes. Further investigation at high resolution and ultra-high mass accuracy are required to document these mass shifts in detail.

Conclusion

We have demonstrated in this study that a combination of LC-MS fractionation with top down MSMS experiments can produce reliable information on the polymorphism and post translational modifications of human salivary proteins and led to the discovery of new protein isoforms, using lower resolution instruments for primary identification of proteins. Nonetheless it is clear that highly resolved and ultra-high accuracy MSMS experiments are necessary to describe subtle changes in proteins that seem to occur in a much wider extent than previously envisioned. We also showed that front-end sample preparation, chromatography, and down-stream data handling still remain critical for achieving meaningful results.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledged financial support obtained from NIH-NIDCR (U01 DE016275-01) is. The authors also thank Robert Barkovich and David Horn (Thermo Scientific Corp.) for assistance with software.

Abbreviations

IMT	Intact Mass Tag
MS/MS	Tandem mass spectrum
FTICR-MS	Fourier-transform Ion Cyclotron resonance mass spectrometry
PRP	proline rich protein
CAD	collisionally activated dissociation
ECD	electron-capture dissociation
IRMPD	Infrared multiphoton dissociation
PSP	protein sequence polymorphisms

References

1. Kaufman E, Lamster IB. The diagnostic applications of saliva--a review. *Crit Rev Oral Biol Med.* 2002; 13(2):197–212. [PubMed: 12097361]
2. Streckfus CF, Bigler LR. Saliva as a diagnostic fluid. *Oral Dis.* 2002; 8(2):69–76. [PubMed: 11991307]
3. Hardt M, Thomas LR, Dixon SE, Newport G, Agabian N, Prakobphol A, Hall SC, Witkowska HE, Fisher SJ. Toward defining the human parotid gland salivary proteome and peptidome: identification and characterization using 2D SDS-PAGE, ultrafiltration, HPLC, and mass spectrometry. *Biochemistry.* 2005; 44(8):2885–99. [PubMed: 15723531]

4. Walz A, Stuhler K, Wattenberg A, Hawranke E, Meyer HE, Schmalz G, Bluggel M, Ruhl S. Proteome analysis of glandular parotid and submandibular-sublingual saliva in comparison to whole human saliva by two-dimensional gel electrophoresis. *Proteomics*. 2006; 6(5):1631–9. [PubMed: 16402355]
5. Denny P, Hagen FK, Hardt M, Liao L, Yan W, Arellanno M, Bassilian S, Bedi GS, Boontheung P, Cociorva D, Delahunty CM, Denny T, Dunsmore J, Faull KF, Gilligan J, Gonzalez-Begne M, Halgand F, Hall SC, Han X, Henson B, Hewel J, Hu S, Jeffrey S, Jiang J, Loo JA, Ogorzalek Loo RR, Malamud D, Melvin JE, Miroshnychenko O, Navazesh M, Niles R, Park SK, Prakobphol A, Ramachandran P, Richert M, Robinson S, Sondej M, Souda P, Sullivan MA, Takashima J, Than S, Wang J, Whitelegge JP, Witkowska HE, Wolinsky L, Xie Y, Xu T, Yu W, Ytterberg J, Wong DT, Yates JR 3rd, Fisher SJ. The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. *J Proteome Res*. 2008; 7(5):1994–2006. [PubMed: 18361515]
6. Hardt M, Witkowska HE, Webb S, Thomas LR, Dixon SE, Hall SC, Fisher SJ. Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal Chem*. 2005; 77(15):4947–54. [PubMed: 16053308]
7. Quintana M, Palicki O, Lucchi G, Ducoroy P, Chambon C, Salles C, Morzel M. Inter-individual variability of protein patterns in saliva of healthy adults. *J Proteomics*. 2009; 72(5):822–30. [PubMed: 19470415]
8. Cabras T, Castagnola M, Inzitari R, Ekstrom J, Isola M, Riva A, Messana I. Carbachol-induced in vitro secretion of certain human submandibular proteins investigated by mass-spectrometry. *Arch Oral Biol*. 2008; 53(11):1077–83. [PubMed: 18617156]
9. Messana I, Cabras T, Pisano E, Sanna MT, Olianias A, Manconi B, Pellegrini M, Paludetti G, Scarano E, Fiorita A, Agostino S, Contucci AM, Calo L, Picciotti PM, Manni A, Bennick A, Vitali A, Fanali C, Inzitari R, Castagnola M. Trafficking and postsecretory events responsible for the formation of secreted human salivary peptides: a proteomics approach. *Mol Cell Proteomics*. 2008; 7(5):911–26. [PubMed: 18187409]
10. Siqueira WL, Salih E, Wan DL, Helmerhorst EJ, Oppenheim FG. Proteome of human minor salivary gland secretion. *J Dent Res*. 2008; 87(5):445–50. [PubMed: 18434574]
11. Hu S, Denny P, Denny P, Xie Y, Loo JA, Wolinsky LE, Li Y, McBride J, Ogorzalek Loo RR, Navazesh M, Wong DT. Differentially expressed protein markers in human submandibular and sublingual secretions. *Int J Oncol*. 2004; 25(5):1423–30. [PubMed: 15492834]
12. Robinovitch MR, Ashley RL, Iversen JM, Vigoren EM, Oppenheim FG, Lamkin M. Parotid salivary basic proline-rich proteins inhibit HIV-I infectivity. *Oral Dis*. 2001; 7(2):86–93. [PubMed: 11355444]
13. Giusti L, Baldini C, Bazzichi L, Ciregia F, Tonazzini I, Mascia G, Giannaccini G, Bombardieri S, Lucacchini A. Proteome analysis of whole saliva: a new tool for rheumatic diseases--the example of Sjogren's syndrome. *Proteomics*. 2007; 7(10):1634–43. [PubMed: 17436266]
14. Hu S, Arellano M, Boontheung P, Wang J, Zhou H, Jiang J, Elashoff D, Wei R, Loo JA, Wong DT. Salivary proteomics for oral cancer biomarker discovery. *Clin Cancer Res*. 2008; 14(19):6246–52. [PubMed: 18829504]
15. Oppenheim FG, Salih E, Siqueira WL, Zhang W, Helmerhorst EJ. Salivary proteome and its genetic polymorphisms. *Ann N Y Acad Sci*. 2007; 1098:22–50. [PubMed: 17303824]
16. Messana I, Cabras T, Inzitari R, Lupi A, Zuppi C, Olmi C, Fadda MB, Cordaro M, Giardina B, Castagnola M. Characterization of the human salivary basic proline-rich protein complex by a proteomic approach. *J Proteome Res*. 2004; 3(4):792–800. [PubMed: 15359733]
17. Messana I, Loffredo F, Inzitari R, Cabras T, Giardina B, Onnis G, Piludu M, Castagnola M. The coupling of RP-HPLC and ESI-MS in the study of small peptides and proteins secreted in vitro by human salivary glands that are soluble in acidic solution. *Eur J Morphol*. 2003; 41(2):103–6. [PubMed: 15621864]
18. Inzitari R, Cabras T, Onnis G, Olmi C, Mastinu A, Sanna MT, Pellegrini MG, Castagnola M, Messana I. Different isoforms and post-translational modifications of human salivary acidic proline-rich proteins. *Proteomics*. 2005; 5(3):805–15. [PubMed: 15693058]

19. Inzitari R, Cabras T, Rossetti DV, Fanali C, Vitali A, Pellegrini M, Paludetti G, Manni A, Giardina B, Messina I, Castagnola M. Detection in human saliva of different statherin and P-B fragments and derivatives. *Proteomics*. 2006; 6(23):6370–9. [PubMed: 17080484]
20. Helmerhorst EJ, Sun X, Salih E, Oppenheim FG. Identification of Lys-Pro-Gln as a novel cleavage site specificity of saliva-associated proteases. *J Biol Chem*. 2008; 283(29):19957–66. [PubMed: 18463091]
21. Castagnola M, Messina I, Inzitari R, Fanali C, Cabras T, Morelli A, Pecoraro AM, Neri G, Torrioli MG, Gurrieri F. Hypo-phosphorylation of salivary peptidome as a clue to the molecular pathogenesis of autism spectrum disorders. *J Proteome Res*. 2008; 7(12):5327–32. [PubMed: 19367726]
22. Inzitari R, Vento G, Capoluongo E, Boccacci S, Fanali C, Cabras T, Romagnoli C, Giardina B, Messina I, Castagnola M. Proteomic analysis of salivary acidic proline-rich proteins in human preterm and at-term newborns. *J Proteome Res*. 2007; 6(4):1371–7. [PubMed: 17341109]
23. Helmerhorst EJ, Oppenheim FG. Saliva: a dynamic proteome. *J Dent Res*. 2007; 86(8):680–93. [PubMed: 17652194]
24. Messina I, Inzitari R, Fanali C, Cabras T, Castagnola M. Facts and artifacts in proteomics of body fluids. What proteomics of saliva is telling us? *J Sep Sci*. 2008; 31(11):1948–63. [PubMed: 18491358]
25. Messina I, Kasicka V. Analysis of peptides by separation and mass spectrometric methods. *J Sep Sci*. 2008; 31(3):425–6. [PubMed: 18266257]
26. Forbes AJ, Patrie SM, Taylor GK, Kim YB, Jiang L, Kelleher NL. Targeted analysis and discovery of posttranslational modifications in proteins from methanogenic archaea by top-down MS. *Proc Natl Acad Sci U S A*. 2004; 101(9):2678–83. [PubMed: 14976258]
27. Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol Cell Proteomics*. 2006; 5(5):935–48. [PubMed: 16439352]
28. Roth MJ, Forbes AJ, Boyne MT 2nd, Kim YB, Robinson DE, Kelleher NL. Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol Cell Proteomics*. 2005; 4(7):1002–8. [PubMed: 15863400]
29. Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL. New and automated MSn approaches for top-down identification of modified proteins. *J Am Soc Mass Spectrom*. 2005; 16(12):2027–38. [PubMed: 16253516]
30. Zabrouskov V, Han X, Welker E, Zhai H, Lin C, van Wijk KJ, Scheraga HA, McLafferty FW. Stepwise deamidation of ribonuclease A at five sites determined by top down mass spectrometry. *Biochemistry*. 2006; 45(3):987–92. [PubMed: 16411774]
31. Wu S, Lourette NM, Tolic N, Zhao R, Robinson EW, Tolmachev AV, Smith RD, Pasa-Tolic L. An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. *J Proteome Res*. 2009; 8(3):1347–57. [PubMed: 19206473]
32. Whitelegge JP, Zabrouskov V, Halgand F, Souda P, Bassilian S, Yan W, Wolinsky L, Loo JA, Wong DT, Faull KF. Protein-Sequence Polymorphisms and Post-translational Modifications in Proteins from Human Saliva using Top-Down Fourier-transform Ion Cyclotron Resonance Mass Spectrometry. *Int J Mass Spectrom*. 2007; 268(2–3):190–197. [PubMed: 19050733]
33. Halgand F, Zabrouskov V, Bassilian S, Souda P, Wong DT, Loo JA, Faull KF, Whitelegge JP. Micro-heterogeneity of human saliva Peptide P-C characterized by high-resolution top-down Fourier-transform mass spectrometry. *J Am Soc Mass Spectrom*. 2010; 21(5):868–77. [PubMed: 20185333]
34. Ryan CM, Souda P, Halgand F, Wong DT, Loo JA, Faull KF, Whitelegge JP. Confident Assignment of Intact Mass Tags to Human Salivary Cystatins Using Top-Down Fourier-Transform Ion Cyclotron Resonance Mass Spectrometry. *J Am Soc Mass Spectrom*.
35. Holmes MR, Giddings MC. Prediction of posttranslational modifications using intact-protein mass spectrometric data. *Anal Chem*. 2004; 76(2):276–82. [PubMed: 14719871]
36. Wolff A, Begleiter A, Moskona D. A novel system of human submandibular/sublingual saliva collection. *J Dent Res*. 1997; 76(11):1782–6. [PubMed: 9372796]

37. Whitelegge JP, Zhang H, Aguilera R, Taylor RM, Cramer WA. Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein: cytochrome b(6)f complex from spinach and the cyanobacterium *Mastigocladus laminosus*. *Mol Cell Proteomics*. 2002; 1(10):816–27. [PubMed: 12438564]
38. Whitelegge JP, Gundersen CB, Faull KF. Electrospray-ionization mass spectrometry of intact intrinsic membrane proteins. *Protein Sci*. 1998; 7(6):1423–30. [PubMed: 9655347]
39. Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*. 1984; 11(11):601. [PubMed: 6525415]
40. LeDuc RD, Taylor GK, Kim YB, Januszyk TE, Bynum LH, Sola JV, Garavelli JS, Kelleher NL. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res*. 2004; 32(Web Server issue):W340–5. [PubMed: 15215407]
41. Palumbo AM, Reid GE. Evaluation of gas-phase rearrangement and competing fragmentation reactions on protein phosphorylation site assignment using collision induced dissociation-MS/MS and MS3. *Anal Chem*. 2008; 80(24):9735–47. [PubMed: 19012417]
42. Palumbo AM, Tepe JJ, Reid GE. Mechanistic insights into the multistage gas-phase fragmentation behavior of phosphoserine- and phosphothreonine-containing peptides. *J Proteome Res*. 2008; 7(2):771–9. [PubMed: 18181561]
43. Rodriguez J, Gupta N, Smith RD, Pevzner PA. Does trypsin cut before proline? *J Proteome Res*. 2008; 7(1):300–5. [PubMed: 18067249]
44. Stubbs M, Chan J, Kwan A, So J, Barchynsky U, Rassouli-Rahsti M, Robinson R, Bennick A. Encoding of human basic and glycosylated proline-rich proteins by the PRB gene complex and proteolytic processing of their precursor proteins. *Arch Oral Biol*. 1998; 43(10):753–70. [PubMed: 9796777]
45. Whitelegge J, Halgand F, Souda P, Zabrouskov V. Top-down mass spectrometry of integral membrane proteins. *Expert Rev Proteomics*. 2006; 3(6):585–96. [PubMed: 17181473]
46. Zubarev RA. Reactions of polypeptide ions with electrons in the gas phase. *Mass Spectrom Rev*. 2003; 22(1):57–77. [PubMed: 12768604]
47. Cooper HJ, Akbarzadeh S, Heath JK, Zeller M. Data-dependent electron capture dissociation FT-ICR mass spectrometry for proteomic analyses. *J Proteome Res*. 2005; 4(5):1538–44. [PubMed: 16212404]
48. Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem*. 2000; 72(3):563–73. [PubMed: 10695143]
49. Lee S, Han SY, Lee TG, Chung G, Lee D, Oh HB. Observation of pronounced b*, y cleavages in the electron capture dissociation mass spectrometry of polyamidoamine (PAMAM) dendrimer ions with amide functionalities. *J Am Soc Mass Spectrom*. 2006; 17(4):536–43. [PubMed: 16490359]
50. Pesavento JJ, Kim YB, Taylor GK, Kelleher NL. Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry. *J Am Chem Soc*. 2004; 126(11):3386–7. [PubMed: 15025441]
51. Frank AM, Pesavento JJ, Mizzen CA, Kelleher NL, Pevzner PA. Interpreting top-down mass spectra using spectral alignment. *Anal Chem*. 2008; 80(7):2499–505. [PubMed: 18302345]
52. Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, Whitelegge JP, Bafna V, Pevzner PA. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol Cell Proteomics*. 2010; 9(12):2772–82. [PubMed: 20855543]
53. Draper J, Enot DP, Parker D, Beckmann M, Snowdon S, Lin W, Zubair H. Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour ‘rules’. *BMC bioinformatics*. 2009; 10:227. [PubMed: 19622150]
54. Humphrey SP, Williamson RT. A review of saliva: normal composition, flow, and function. *J Prosthet Dent*. 2001; 85(2):162–9. [PubMed: 11208206]
55. Kuipers OP, Rollema HS, Yap WM, Boot HJ, Siezen RJ, de Vos WM. Engineering dehydrated amino acid residues in the antimicrobial peptide nisin. *J Biol Chem*. 1992; 267(34):24340–6. [PubMed: 1447185]

56. Karakas Sen A, Narbad A, Horn N, Dodd HM, Parr AJ, Colquhoun I, Gasson MJ. Post-translational modification of nisin. The involvement of NisB in the dehydration process. *Eur J Biochem.* 1999; 261(2):524–32. [PubMed: 10215865]
57. Wiedemann I, Breukink E, van Kraaij C, Kuipers OP, Bierbaum G, de Kruijff B, Sahl HG. Specific binding of nisin to the peptidoglycan precursor lipid II combines pore formation and inhibition of cell wall biosynthesis for potent antibiotic activity. *J Biol Chem.* 2001; 276(3):1772–9. [PubMed: 11038353]
58. Langer B, Rother D, Retey J. Identification of essential amino acids in phenylalanine ammonia-lyase by site-directed mutagenesis. *Biochemistry.* 1997; 36(36):10867–71. [PubMed: 9283076]
59. Chatterjee C, Paul M, Xie L, van der Donk WA. Biosynthesis and mode of action of lantibiotics. *Chem Rev.* 2005; 105(2):633–84. [PubMed: 15700960]
60. Okesli A, Cooper LE, Fogle EJ, van der Donk WA. Nine post-translational modifications during the biosynthesis of cinnamycin. *J Am Chem Soc.* 2011; 133(34):13753–60. [PubMed: 21770392]
61. Castagnola M, Congiu D, Denotti G, Di Nunzio A, Fadda MB, Melis S, Messana I, Misiti F, Murtas R, Olianias A, Piras V, Pittau A, Puddu G. Determination of the human salivary peptides histatins 1, 3, 5 and statherin by high-performance liquid chromatography and by diode-array detection. *J Chromatogr B Biomed Sci Appl.* 2001; 751(1):153–60. [PubMed: 11232845]
62. Castagnola M, Inzitari R, Rossetti DV, Olmi C, Cabras T, Piras V, Nicolussi P, Sanna MT, Pellegrini M, Giardina B, Messana I. A cascade of 24 histatins (histatin 3 fragments) in human saliva. Suggestions for a pre-secretory sequential cleavage pathway. *J Biol Chem.* 2004; 279(40):41436–43. [PubMed: 15272024]

A) Crude Pscore 3.74-25**Graphical Fragment Mapper**

S P P G K P Q G P P P Q G G N Q P Q G P P P P G K P Q G P
 P P Q G G N K P Q G P P P G K P Q G P P P Q G G S K S R S
 A

B) Crude Pscore 7.98-43**Graphical Fragment Mapper**

S P P G K P Q G P P P Q G G N Q P Q G P P P P P G K P Q G P
 P P Q G G N K P Q G P P P P G K P Q G P P P Q G G S K S R S
 A

C) Crude Pscore 2.12-77**Graphical Fragment Mapper**

S P P G K P Q G P P P Q G G N Q P Q G P P P P P G K P Q G P
 P P Q G G N K P Q G P P P P G K P Q G P P P Q G G S K S R S
 A

Figure 1.

Peptide P-F protein (MW 5841.98 Da) MSMS spectra obtained from **A)** CID MSMS experiment in a Qstar XL instrument. RMS is 52 PPM for this experiment. **B)** IRMPD MSMS experiment in a LTQFT mass spectrometer; and **C)** ECD MSMS experiment using LTQFT mass spectrometer. CID and IRMPD MSMS experiments only allowed to unambiguously identifying the N-terminal part of the protein. The identification of protein Peptide P-F was solved with ECD MSMS data showing high mass c ions (underlined in Figure 1C) and unique ions (RMS on FTICR was less than 5 ppm). Additional supporting information is given in the Supplementary Material.

A) Form 1; Crude Pscore 1.33e-14

Graphical Fragment Mapper

N L N E D V Q E E S P S L I A G N P Q G P S P Q G G N K
 P Q G P P P P P G K P Q G P P Q G G N K P Q G P P P G
 K P Q G P P P Q G D K S R S P R

Form 2; Crude Pscore 7.48e-17

Graphical Fragment Mapper

N L N E D V Q E E S P S L I A G N P Q G P S P Q G G N K
 P Q G P P P P P G K P Q G P P Q G G N K P Q G P P P G
 K P Q G P P P Q G D K S R S P R

B) Form 2; Crude Pscore 2.35e-39

Graphical Fragment Mapper

N L N E D V Q E E S P S L I A G N P Q G P S P Q G G N K
 P Q G P P P P P G K P Q G P P Q G G N K P Q G P P P G
 P Q G P P P Q G D K S R S P R

C) Form 1; Crude Pscore 1.73e-149

Graphical Fragment Mapper

N L N E D V Q E E S P S L I A G N P Q G P S P Q G G N K
 P Q G P P P P P G K P Q G P P Q G G N K P Q G P P P G
 P Q G P P P Q G D K S R S P R

Form 2; Crude Pscore 1.74e-09

Graphical Fragment Mapper

N L N E D V Q E E S P S L I A G N P Q G P S P Q G G N K
 P Q G P P P P P G K P Q G P P Q G G N K P Q G P P P G
 K P Q G P P P Q G D K S R S P R

Figure 2.

A) protein II-2 sequence chart obtained from CID MSMS showing the presence of isoform 1. RMS was 30 ppm. **B)** IRMPD data allowed determining the presence of the second isoform of protein II-2 with unique b fragment ions which allowed us demonstrating the presence of a dehydroalanine residue in position 8. **C)** ECD MSMS data showing fragment ions matching both IMT's and sequence characteristics of the two protein II-2 isoforms. RMS was 1 ppm for LTQFT data. Characteristics of the two isoforms: isoform 1: pyroglutamic residue in N-terminus, proline 39 lacking and a phosphoserine in position 8; Isoform 2: pyroglutamic residue in N-terminus, presence of proline 39 and dehydroalanine in position 8. Additional supporting information is given in the Supplementary Material.

« Manual » PScore: 9.86e-109

Graphical Fragment Mapper

Figure 3.

Sequence chart showing the fragment ions and the modifications which fit with the IMT and MSMS data obtained for Ib8-a protein. The RMS on fragment ions is 2.7 ppm for this experiment. Modifications proposed in this case concerned the presence of a Q to H mutation, the presence of a glucose molecule, and the introduction of a methyl group that could either be on the glucose molecule or on the C-terminal carboxylic group, in the C-terminal portion of the protein. Similarly, unique fragment ions allowed assessing the sequence and proposed modifications. Additional supporting information is given in Supplementary Material.

Mono PKL Pscore: 1.99636×10^{-81}

Graphical Fragment Mapper

S P P G K P Q G P P P Q G G K P Q G P P P Q G G N K P Q G P
 P P P G K P Q G P P A Q G G S K S Q S A R A P P G K P Q G P
 P Q Q E G N N P Q G P P P A G G N P Q Q P Q A P P A G Q P
 Q G P P R P P Q G G R P S R P P Q

“Manual” Pscore: 1.9012×10^{-146}

Graphical Fragment Mapper

S P P G K P Q G P P P Q G G K P Q G P P P Q G G N K P Q G P
 P P P G K P Q G P P A Q G G S K S Q S A R A P P G K P Q G P
 P Q Q E G N N P Q G P P P A G G N P Q Q P Q A P P A G Q P
 Q G P P R P P Q G G R P S R P P Q

Figure 4.

Sequence chart of the proposed sequence for the 10.4 kDa protein sequence (P04280, AA 92–198). RMS on the fragment ions is 2.2 ppm for this experiment. The calculated “Manual” Pscore was 1.90×10^{-146} . Additional supporting information is given in the Supplementary Material.

Figure 5.

Processing workflow to achieve a full and reliable identification of proteins using top down MS and MSMS.

Step 1: This step is, except for the fact that genetic data and deduced sequences should be taken into account for automatic searching, identical to the solutions proposed in Prosight PTM and PC software. Completion of data processing is proposed with following steps. **Step 2:** corresponds to the calculation of masses of normal fragment ions as well as internal fragments or losses of ammonia and water from these ions. The goal is to differentiate normal fragment ions (b, y, c or z ions) from “unusual” ions and then to remove “useless” masses from the peaklist (PKL). Unique ions identified during this process are labeled in the PKL and are annotated (m/z peaks and charge states) in the MSMS spectrum. The following step (**Step 3**) is a validation of peaks identified as normal ions while checking for the real presence of the ion, based on the signal-to-noise ratio and the shape of the isotopic profile. To achieve this goal, the isotopic profile of ions of interest are calculated from the deduced chemical formula and overlaid with the experimental MSMS spectrum. At this stage a new spectrum is generated in which matched ions are removed. The purged spectrum is used for an iterative new search from step 1 to step 3. The last step (**Step 4**) is a manual analysis of the remaining peaks in the last purged spectrum. The goal is to achieve a full identification

of the ions present in the MSMS spectrum. The data analysis is complete when last purged MSMS spectrum displays only noise.

Table list of measured Intact Mass Tags of proteins cited in the article as well as related information such as sequence, accurate measured masses, theoretical molecular weights, accession numbers and amino-acid locations and observations.

Table 1

Protein ID	Measured Intact Mass Tag	Accurate Intact Mass Tag	Acc numbers	Sequences	Calculated MW (Da)	Observations
	MW av Da	Da	(swissprot)		mono/ AV	
Peptide P-C	4370.24 Da	4368.2220 Da	P02810 123–166	GRPQGPPQGGHQGGPPPP PGKPPQGGRRPGRRPGQGQSPQ	4368.17/4370.779	Regular sequence
		4367/2385 Da		GRPQGGPQGGHPRPRPPPPGKPKQ GPPQGGRRPGPPQGQSPQ	4367.23/4369.840	Replacement of QQGPPPP by PRPPR
		4369/2960 Da		GRPQGGPQGGHQGGPPPPPGKPKQ QGPPPGGRRPGPPQGQSPQ	4369.15/4371.763	Deamidation or SNP's at Q14
Histatin 1	4927.9 Da	N.D.	P34084 20–57	DSHEKRHHGYRRKHFKEHH SHRETFPYGDYGSNYLYDN	4845.226/4848.174	Phosphorylation possibly localised on Ser21 and Tyr30
Peptide D isoform	5268.73 Da	N.D.	P10161 169–222	SPPGKPPQPPQEEGNKPPGPPGKPPQ PPPGGNPQQAAPTAGKPPQGGPPPP	5268.886/5265.686	No PTM
Peptide H	5590.09 Da	N.D.	P04280 276–331	SPPGKPPQPPQEEGNPPGPPPAAGGNPQ QQQAAPTAGQPPRRPPQGGRRSRPQ	5586.775/5590.099	No PTM
Peptide B	5792.97 Da	N.D.	P02814 23–79	QRGPRGPPYPGLAPQPPGPGFVPPPPPPYPG PGRRPPPAFYGHFPPPTQP	5809.766/5806.656	Pyroglutamic acid in N-terminus
IB-8c (peptide F)	5842.52 Da	N.D.	P02812 265–325	SPPGKPPQPPGQGNQPPGPPPPGKPPGPPQ GGNKPQGGPPPGKPPGPPPGGSKSRSA	5838.984/5842.496	No PTM
IB-6 (Peptide-J)	5943.58 Da	N.D.	P02812 79–139	SPPGKPPQPPQGGNQPPGPPPGKPPGPPPP QGGNKPQGGPPPGKPPPGQGNKSRSS	5939.996/5943.558	No PTM
IB-9 (Peptide-E)	6023.7 Da	6020.0896 Da	P02811	SPPGKPPQPPGQGNQPPGPPPGKPPGPPPPQ GGNRPPGPPPPGKPPQGGIDKSRSPR	6020.081/6023.693	No PTM
		6023.0883 Da				
Peptide D	6949.73 Da	N.D.	P10161 169–238	SPPGKPPQPPQEEGNKPPGPPPGKPPGPPPG GNPQQAAPTAGKPPGPPPPQGGRRPPIAQQQPPQ	6945.546/6949.734	No PTM
II-2	7608.4 Da	7603.7188 Da/7607.7231 Da	C38355	QNLNEDVSQEESPSLIAGNPQGFSPQGGNKPQGPPPPPP GKPPGPPQGGNKPQGPPPGKPPQGGDKSRSPR	7637.818/7642.366	Pyroglutamic acid in N-terminus + Ser 8 replaced by a dehydroAlanine
		7604.7190 Da/7608.7218 Da		QNLNEDVSQEESPSLIAGNPQGFSPQGGNKPQGPPPPGKPKQ GPPPGGNNKPQGGPPPGKPPQGGDKSRSPR	7540.765/7545.249	Variant lacking a Proline residue in position 39 + pyroglutamic acid in N-ter + Phosphorylation in Ser 8
IB-1	9593.4 Da	N.D.	P04281	QNLNEDVSQEESPSLIAGNPQGAAPPQGGNKPQGPPSPGKPPGPPPPQ GGNQPPGPPPPGKPPGPPPGGNNK PQGPPPGKPPGPPPGDKSRSPR	9524.756/9530.439	Pyroglutamic acid in N-terminus + Phosphorylation in Ser 8
P-Ko	10433.39 Da	10428.2887 Da	P04280 92–198	SPPGKPPQPPGKPPQGGPPQGGNKPQGPPPPGKPPGPPPAQGGSKSQ SARAPPGKPPQGGNNPQGGPPPAAGNPQQP QAPPACQPQGPPPPQGGRRSPRPQ	10427.277/10433.503	No PTM
		10433.3097 Da	(K03205 mRNA data)			
PRP-3	11161.6 Da	N.D.	P02810 17–122	QDLDEDVSQEDVPLVISDGSDFSEDFERQGPPLGGQSQSPSAGDGN QNDGPPQPPQGGQGGPPPGKPPGKPPQGP PQGGHPPPPQGGRRPGPPQGGGHRPPR	11013.146/11018.627	Pyroglutamic acid in N-terminus + Phosphorylation on Ser 8 and 17 or 22 + mutation N4 to D