Extensive proteomics characterization of basic proline-rich proteins in human saliva and investigation on their properties as substrates of epithelial transglutaminase 2

Abstract

This PhD thesis describes the work carried out during three years upon the Department of "Scienze della Vita e dell'Ambiente" of the University of Cagliari, work that was also partly performed upon the "Istituto di Biochimica e Biochimica Clinica" of the Faculty of Medicine of the Catholic University of Rome.

Topics of the thesis were centered on the structural and functional characterization of some human salivary proteins.

In particular, the topics investigated were the following:

a) Extensive identification of the components of the family of basic (and glycosylated basic) proline-rich proteins (bPRPs and gPRPs), the most complex and heterogeneous family of the human salivary proteins, by a proteomic top-down platform with the aim to achieve the most complete knowledge possible of the main parent proteins present in human saliva and of their post-translational modifications. This study allowed the characterization of 55 new components of the family bringing the total number of naturally occurring components to 110.

b) Exploration of the reactivity *in vitro* of some bPRPs (P-C, P-H, P-D, P-D P_{32}⟶A, II-2, P-F and P-J) as substrate of the epithelial transglutaminase-2 (TG-2). The aim of this study was to establish whether these bPRPs can potentially contribute *in vivo* to the formation of the "oral mucosal protein pellicle", a protein network covering the oral mucosal epithelia devoted to the protection of the oral cavity. This study allowed establish that almost all the bPRPs are substrates of TG-2 and therefore are potential components of the "oral mucosal protein pellicle" and that, despite the great sequence similarity, their reactivity is significantly different.

Acknowledgement

During my PhD I was supported by different Professors and I would thank them for their valuable helps and expert advice. They are at first Prof. Tiziana Cabras, my kind supervisor, who continuously provoked me by her deep knowledge on the thesis topics, taught me the meaning of experimental precision and supported me with her significant insight on the topics of the thesis along all my PhD period; Prof. Enzo Tramontano, coordinator of the PhD course, for being always available and attempting to push me onward very friendly. It is a pleasure to convey my heartiest gratitude to Prof. Massimo Castagnola, my tutor in Rome, for his constant support while doing my thesis program; and I would like to take the opportunity to thank also Prof. Irene Messana for her advice and guidance in preparing this thesis.

I would like to acknowledge and thank also some other researchers, Alessandra Olianas, Barbara Manconi, Maria Teresa Sanna in Cagliari, Federica Iavarone, Federica Vincenzoni, Claudia Desiderio and Claudia Martelli in Roma which have given me some guideline for the successful completion of my thesis. Moreover, I deeply appreciate my friends Barbara Liori, Simone Serrao and Gianluigi Cabiddu for their friendships and wholehearted cooperation during my work.

This thesis is devoted to my parents, who always follow me. Their encouragement was at the end what made this dissertation possible. I send my deepest love for their dedication and for the many years of support during my postgraduate studies.

Contents **Page**

List of Tables **Page**

List of Figures **Page**

SECTION 1:

Characterization of the human salivary basic proline-rich proteins family by a proteomic top-down platform

1.1 Introduction

1.1.1 Proteomic platforms.

Proteomic platforms can be classified according to different criteria as depicted in Table 1.1. From the most general point of view, proteomic platforms can be divided in qualitative and quantitative (Nikolov M, et al. 2012). The goal of qualitative platforms is to define the complete set of proteins present in a sample, post-translational modifications (PTMs) comprised, without specific concern for the amount. However, qualitative proteomics has to face the unequal distribution of the concentration of distinct proteins present in the biological sample, because the highly abundant proteins can prevent the detection of that ones at low concentration. This problem is well known to researchers working in the field of plasma proteomics, where the low abundant proteins can be revealed only after the depletion of the most abundant ones. Qualitative information can be focused on distinct proteomic subsets, to define for instance either the phosphoproteome, or the components of a specific enzyme family, or the sub-proteome associated with the intracellular organelles. Such a systematic investigation strategy has been mostly pursued in the first decade of proteomics investigations in particular with large international initiatives fostered under the coordination of the Human Proteome Organization, e.g. the liver proteome initiative, the brain proteome initiative or the Plasma Proteome Project (Messana I, et al. 2013).

Nevertheless, it was soon clear that such a proteomic platform needs to be implemented for quantitative determination. In fact, if some proteins or some of their PTMs are uniquely associated with a particular disease, they are potentially eligible as biomarkers. However, this is a very rare condition, because the pathological status is generally associated by the modification of the concentration of some proteins or by a different relative abundance of PTMs. Quantitative approaches can be further divided in relative and absolute. The relative quantification allows establishing the differences in two (or more) proteomes, (i.e. healthy versus pathological subjects) evidencing statistically significant increases or decreases of proteins levels. For large proteomes, the relative quantification is the general approach (Messana I, et al. 2013).

Table 1.1 Proteomic classifications

	Qualitative	Quantitative
Bottom-up	Internal peptide sequencing Protein identification-Mass finger print Single PTMs sequencing	SRM/MRM proteotypic transitions Label free (XIC-SIM) Isotope labeling (metabolic and chemical)
Middle-down	Internal peptide sequencing Multiple PTMs sequencing	SRM/MRM proteotypic transitions Label free (XIC-SIM) Isotope labeling (metabolic and chemical)
Top-down	PTMs code Intact sequencing (dependent on molecular dimensions) N-terminus identification C-terminus identification	Label free XIC (relative or absolute; dependent on standard availability) Area of the ESI spectrum deconvolution

1.1.2 Top-down and bottom-up platforms

Top-down and bottom-up terms applied to proteomic platforms distinguish the strategy utilized in the sample treatment. Top-down proteomics investigates the intact sequence of the protein under examination, avoiding as much as possible any sample alteration. Bottom-up proteomics is centered on a sample pre-digestion (typically with trypsin) followed by the analysis of peptide fragments by high-throughput analytical methods. The presence of a protein in the sample is inferred by the detection of one or more of its specific fragments, implying bi-univocal correspondence between the intact protein and the tryptic fragments. The bottom-up approach derives its philosophy from the shot-gun strategies applied in the detection of DNA sequences in genomics, where the sequence of a long polynucleotide fragment often bi-univocally corresponds to a DNA sequence in a chromosome. The majority of proteins are submitted to extensive post-translational modifications, cleavages included, before reaching the mature functional structure. Furthermore, protein maturation can deeply vary as a function of cellular cycles, tissue and organ. Consequently, the minimalistic

approach of the bottom-up strategy, when transferred to a proteome, can result in the relevant loss of important molecular information (Castagnola M, et al. 2012a). In particular, PTMs are difficult to be highlighted in bottom-up shotgun experiments, where the vast majority of peptide sequences are often associated with a specific cDNA sequence, thus leveling out at a statistical level the presence of a PTM. Moreover, the association of molecular maturation events associated with the specific onset of a defined PTM will not be directly accessible by a bottom-up shotgun experiments (Messana I, et al. 2013).

1.1.3 The human salivary proteome

Most of the about 2400 different proteins of whole saliva (Ekström J, et al. 2017) characterized in recent years by proteomic studies are not of glandular origin but probably originate from exfoliating epithelial cells and oral microflora. Proteins of gland secretion origin should be not more than 200–300 and they represent more than 85% by weight of the salivary proteome (Fig. 1.1). They belong to the following major families: α -amylases, carbonic anhydrase, histatins, mucins, proline-rich proteins (PRPs), further divided in acidic (aPRPs), basic (bPRPs) and basic glycosylated (gPRPs), statherin, P-B peptide and S (salivary)-type, C, and D cystatins (Ekström J, et al. 2017) (Fig. 1.1). The function, origin and encoding genes of the major salivary proteins are reported in Table 1.2, together with the name of mature proteins and the main post-translational modifications occurring before, during and after secretion (Ekström J, et al. 2017). Histatins are a family of small peptides, the name referring to the high number of histidine residues in their structure. All the members of this family arise from histatin 1 and histatin 3, sharing very similar sequences and encoded by two genes (HTN1 and HTN3) located on chromosome 4q13 (Sabatini LM and Azen EA. 1989). Statherin is an unusual tyrosine-rich 43-residue phosphorylated peptide involved in oral cavity calcium ion homeostasis and teeth mineralization. Its gene (STATH) is localized on chromosome 4q13.3, near to the histatin genes (Schwartz SS, et al. 1992)

Salivary cystatins comprise cystatin S, SN and SA; they are inhibitors of cysteine proteinases and this property suggests their role in the protection of the oral cavity from pathogens and in the control of lysosomal cathepsins (Bobek LA and Levine MJ, 1992). Cystatin S1 and cystatin S2 correspond to mono- and di-phosphorylated cystatin S, respectively. The loci expressing all the S cystatins (CST1–5) are clustered on chromosome 20p11.21 together with the loci of cystatins C and D. While cystatin SA seems to be specifically expressed in the oral cavity, cystatin S and SN have also been detected in other body fluids and organs, such as tears, urine and seminal fluid (Dickinson DP, 2002; Ryan CM, et al. 2010). Salivary amylases consist of two families of isoenzymes, called A and B, each family comprising three isoforms whose differences are connected to different post-translational modifications (Scannapieco, FA, et al. 1993). Salivary mucins are

divided into two distinct classes: the large gel-forming mucins (MG1), and the small soluble mucins (MG2) (Offner GD, and Troxler RF, 2000; Thomsson KA, et al. 2002). MG1 represents a heterogeneous family of 20–40 mDa glycoproteins expressed by MUC5B, MUC4 and MUC19 genes. MG2, a much smaller mucin of 130–180 kDa, is the product of the MUC7 gene mapped to chromosome 4q13-q21 (Bobek LA, et al. 1996). Mucins are comprised of approximately 15–20% protein and up to 80% carbohydrate, present largely in the form of serine and threonine O-linked glycans (Strous GD, and Dekker J, 1992; Gendler SJ, and Spicer AP, 1995). The polypeptide backbone can be divided into three regions: the central region contains tandemly repeated sequences of 8 to 169 amino acids. This domain serves as the attachment site for the O-glycans, and each mucin has a unique, specific tandem-repeat sequence. Many mucins with monomeric molecular weights greater than 2 mDa form multimers more than ten times bigger than that size (Ekström J, et al. 2017).

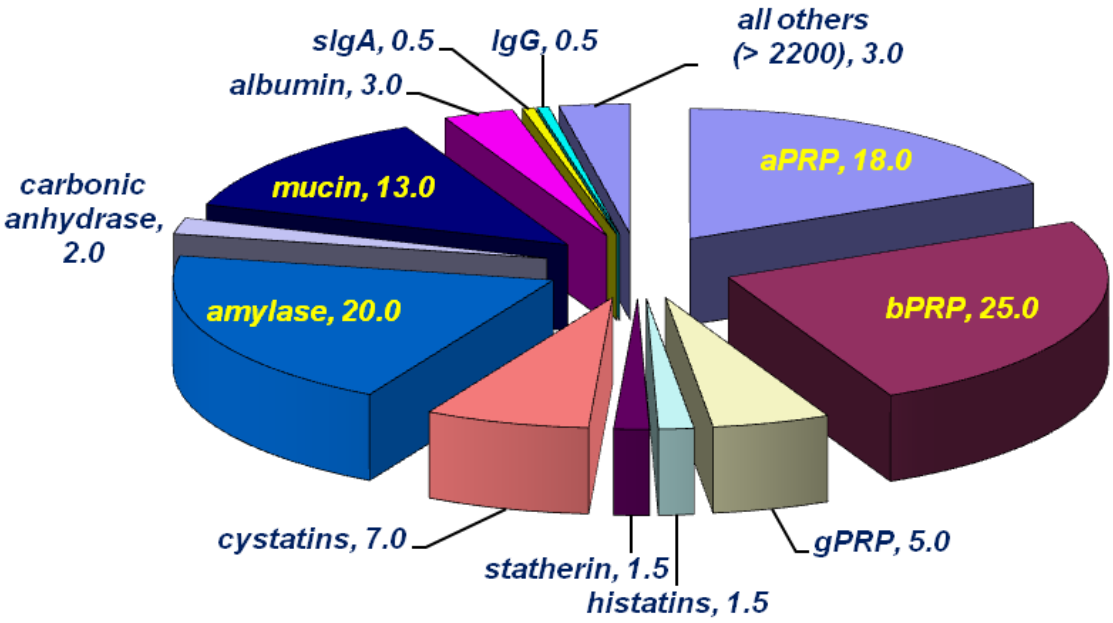


Figure 1.1 Approximate percentages (w/w) of the different protein families present in human adult whole saliva, assuming a comparable contribution of parotid and submandibular/sublingual glands (modified from Messana et al. 2008b).

Table 1.2

Family	Function	Origin	Gene	Mature proteins	Other PTMs
α-Amylases	Antibacterial, digestion, tissue coating	Pr Sm/Sl	*AMY1A*	α-Amylase 1	Disulfide bond, N-glycosylation, phosphorylation, proteolytic cleavages
Acidic PRPs	Lubrication, mineralization, tissue coating	Pr Sm/Sl	*PRH1, PRH2*	Db-s, pa, PIF-s, pa 2-mer, Db-f, PIF-f, PRP-1, PRP-2, PRP-3, PRP-4, P-C peptide	Disulfide bond, further proteolytic cleavages, phosphorylation, protein network
Basic PRPs Glycosylated PRPs	Binding of tannins, tissue coating Antiviral, lubrication	Pr	*PRB1, PRB2 PRB3, PRB4*	II-1, II-2, CD-IIg, IB-1, IB-6, IB-7, IB-8a (Con1−/+), P-D, P-E, P-F, P-J, P-H, proline-rich protein Gl 1–8, protein N1, salivary proline-rich protein Po	Disulfide bond (Gl 8), further proteolytic cleavages N- and O-glycosylation, phosphorylation, protein network
Carbonic anhydrase VI	Buffering, taste	Pr Sm	*CA6*	Carbonic anhydrase 6	Disulfide bond, glycosylation
Cystatins	Antibacterial, antiviral, mineralization, tissue coating	Pr Sm/Sl	*CST1,CST2 CST3, CST4 CST5*	Cystatin SN, cystatin SA, cystatin C, cystatin S and cystatin D	Disulfide bond, O-glycosylation, phosphorylation, sulfoxide, truncated forms
Histatins	Antifungal, antibacterial, mineralization, wound healing	Pr Sm/Sl	*HTN1, HTN3*	Histatin 1, histatin 2, histatin 3, histatin 5, histatin 6	Further proteolytic cleavages, phosphorylation, sulfation
Lactoferrin	Antibacterial, antifungal, antiviral, innate immune response	All salivary glands	*LTF*	Lactoferrin	Disulfide bond, glycosylation, phosphorylation
Lysozyme	Antibacterial	Pr Sm	*LYZ*	Lysozyme C	Disulfide bond
Mucins	Antibacterial, antiviral, digestion, lubrication, tissue coating	All salivary glands	*MUC5B, MUC19 MUC7*	Mucin-5B, mucin-19 Mucin-7	Disulfide bond, N- and O-glycosylation, phosphorylation
Peptide P-B	Not defined	Pr Sm/Sl	*SMR3B (PROL3)*	Proline-rich peptide P-B	Proteolytic cleavages
Statherins	Inhibits crystal formation, lubrication, mineralization, tissue coating	Pr Sm/Sl	*STATH*	Statherin, statherin SV2	Phosphorylation, proteolytic cleavages, protein network

Abbreviations: Pr: Parotid; Sm/Sl: Submandibular/sublingual; GCF: Gingival crevicular fluid.

1.1.4 Human salivary proline-rich proteins

Proline-rich proteins are the most composite family of salivary proteins (Oppenheim FG, et al. 2007; Messana I, et al. 2008a**)** coded by a cluster of 6 genes, strictly associated in a segment of around 4.0 Kb in length on chromosome 12 at band 13.2 (Azen EA, et al. 1985; Mamula PW, et al. 1985; Scherer S.E, et al. 2006) (Fig. 1.2)

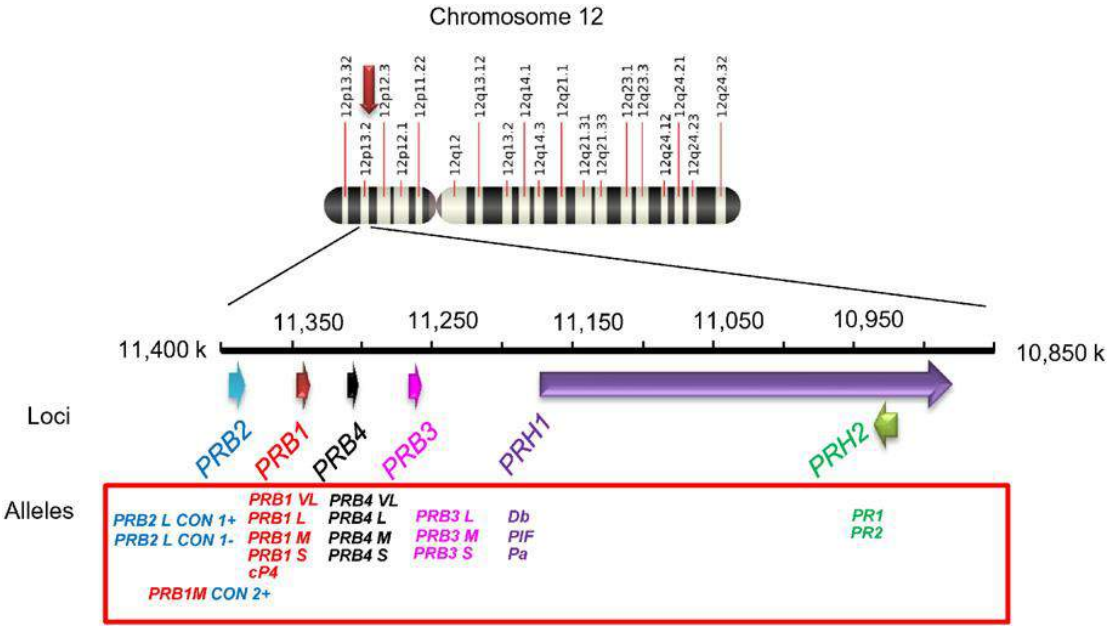


Figure 1.2 Schematic representation of the human PRP gene cluster. The six genes of PRPs (PRB2, ID: 653247; PRB1, ID: 5542; PRB4, ID: 5545; PRB3, ID: 5544; PRH1, ID: 5554; PRH2, ID: 5555) are contained within an ~0.5 Mb segment of the chromosome 12p13.2. The red box reports the main alleles found in Caucasian population (Manconi B, et al. 2016b).

Human salivary PRPs are unique among the PRP families for the complete absence of hydroxyproline, hydroxylysine, and aromatic amino acids. The major aPRPs are 150 residue-long and the acidic portion is restricted to the first 30 residues for the presence of many Asp and Glu residues. The remaining part of the sequence shows high similarities with bPRPs and is highly repetitive, although aPRP repeats differ slightly from bPRP repeats. Due to these structural features, bPRPs and aPRPs elute as distinct chromatographic clusters in RP-HPLC separations (Fig. 1.2A). While aPRPs are secreted by both parotid and submandibular/sublingual glands (in different percentages), bPRPs are secreted only by parotid glands. A further distinction between aPRPs and bPRPs is that while aPRPs can be found in saliva both as intact and truncated proteoforms, bPRPs encoded by *PRB1*, *PRB2* and *PRB4* genes are detectable in saliva only as fragments of the bigger proproteins (Manconi B, et al. 2016b).

Two loci, *PRH1* and *PRH2*, encode for aPRPs. In western population *PRH2* locus is commonly biallelic and gives rise to PRP-1 and PRP-2, two protein species of 150 amino acid residues differing only at position 50 (Asn/Asp, respectively). Three different alleles of *PRH1* locus give rise to the parotid isoelectric-focusing variant slow (PIF-s), the parotid acidic protein (Pa), both 150 residues long, and the double band isoform slow (Db-s), 171 amino acid residues long. The names, deriving from electrophoretic and isoelectric focusing separation of human parotid salivary proteins, are confusing because all the different aPRPs are secreted by both parotid and submandibular/sublingual glands, with a relative contribution of about 80% and 20%, respectively (Messana I, et al. 2008a). The three protein species encoded by *PRH1* locus differ: a) for the residue at position 26 that is Leu for Db-s and Pa, and Ile for PIF-s; b) for the insertion in Db-s of a 21 amino acid residues repeat after position 81; c) for the residue at position 103 that is Cys in Pa and Arg in PIF-s. In the Db-s isoform the Arg is shifted to the 124 position. Pa is commonly present in human saliva as a Pa-dimer, generated by the formation of a disulfide bond between the Cys_{103} residues of the monomers (Fig. 1.3) (Manconi B, et al. 2016b).

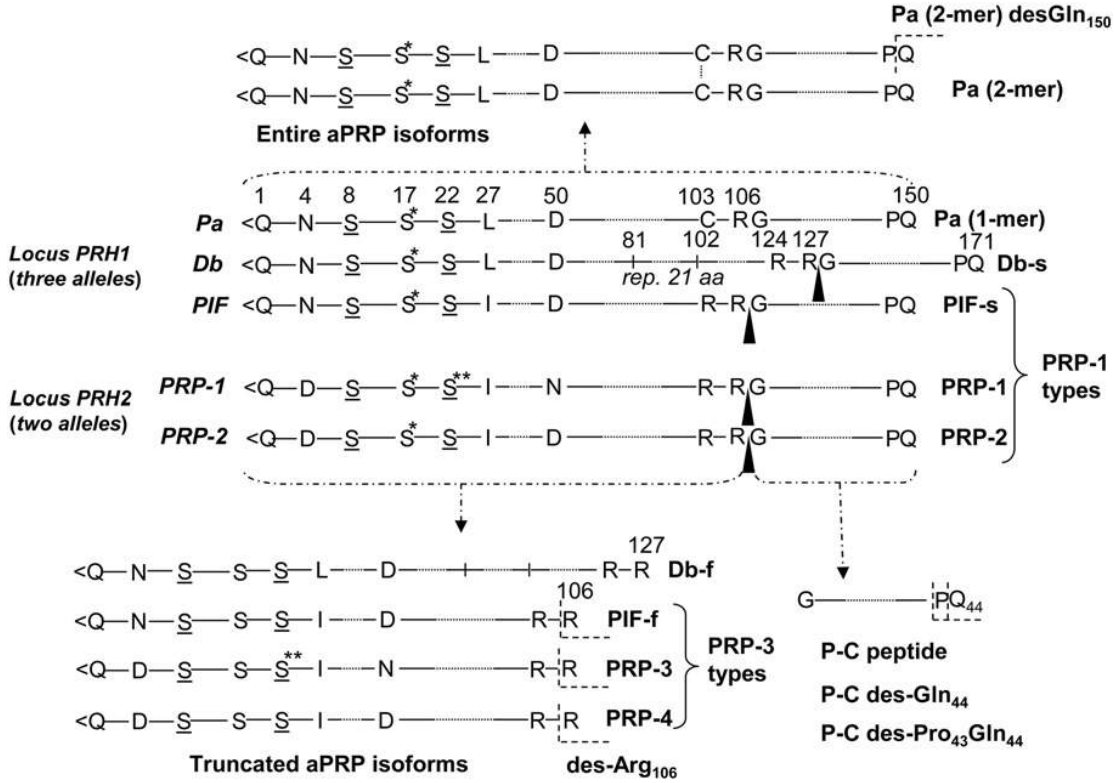


Figure 1.3 Schematic representation of the most common human salivary aPRP protein species detectable in adult saliva of western population (from Inzitari R, et al 2007). PRP-1, PRP-2, PIF-s, and Db-s are partially cleaved (bold arrows) at Arg106 (Arg127 in Db isoform) generating the four truncated protein species reported on the bottomleft of the figure and the P-C peptide. The Pa isoform, carrying the substitution Arg103 → Cys is not cleaved, and it is usually present in human saliva as a 2-mer. Some entire or truncated protein species can partially undergo carboxypeptidase removal of C-terminal residues. <Q: N-terminal pyroglutamic acid; S: phosphorylated Ser (Ser8 and Ser22); S*: minor site of phosphorylation (Ser17); S22**: pSer22 → Phe variation in PRP-1 (and PRP-3) Roma-Boston variant (Manconi B, et al. 2016b).

In adult human saliva both full-length and truncated aPRPs present a pyro-Glu at the N-terminus and are mainly diphosphorylated on Ser_8 and Ser_{22}, by the action of Fam20C, a physiological casein kinase that phosphorylates multiple secreted proteins within a SXE/pS consensus sequence (Tagliabracci VS, et al. 2012). More detailed information on aPRPs are available in a recent review on this topic (Manconi B, et al. 2016b).

The cluster of genes encoding for bPRPs and gPRPs includes four loci named *PRB1-PRB4*, each one existing in several allelic forms (Fig. 1.2). Each PRB gene covers four exons, the third of which is fully composed of 63-bp tandem repeats coding the proline-rich portion of the protein products. Variation in the numbers of these repeats is responsible for length differences in different alleles of the PRB genes (Lyons K.M, et al. 1988a; Maeda N, et al. 1985). At least four alleles (S, small; M, medium; L, large; and VL, very large) are present in the western population at *PRB1* and *PRB3* loci, and three (S, M, L) at *PRB2* and *PRB4* loci (Azen EA, et al. 1993). These alleles, in addition to tandem repeat length variations, show SNPs in the coding region, polymorphic cleavage sites and polymorphic stop codons. Moreover, alternative splicing generates multiple transcript variants encoding distinct protein species, and some alleles are still pending for their characterization (Azen EA, et al. 1996; Lyons KM, et al. 1988b; Stubbs M, et al. 1998). Genetic variability, PTMs implicated in the pre-secretory maturation processes and further transformations occurring in the oral environment give a contribution to the heterogeneity of bPRPs and gPRPs. The proteolytic cleavage is the occurring main post-translational event. Indeed, except for the protein encoded by the *PRB3* locus that originates several gPRPs, the pro-proteins encoded by each allele are completely cleaved by proprotein convertases before granule maturation, generating smaller peptides (Chan M, et al. 2001; Messana I, et al. 2008a). Moreover, after secretion these peptides are further cleaved by endogenous and exogenous (microbioma) proteinases generating numerous fragments (Vitorino R, et al. 2007; Siqueira WL, et al. 2009).

The function of this group of human salivary protein is not well established. As tannin-binding proteins, they have probably a protective role against the potential deleterious effects of these substances. Recently, an unidentified basic PRP was shown to inhibit HIV-I infectivity. If confirmed, this antiviral action can offer

adequate explanation for their abundance and molecular heterogeneity in the oral cavity. A preliminary integrated top-down/bottom up RP-HPLC-ESI-MS proteomic study of bPRPs was carried out more than fifteen years ago in the laboratories of the Istituto di Biochimica e Biochimica Clinica of the Catholic University and in the Dip. di Scienze della Vita, dell'Ambiente e del Farmaco of the University of Cagliari (Messana I, et al. 2004). In that period the used ion-trap MS with a resolution of about 1/5000 did not allow to characterize all the masses potentially attributable to bPRPs. In the last years, the availability of a high-resolution MS apparatus (Orbitrap MS) increased the analytical skills of these laboratories. This section of the thesis describes the results obtained in the characterization of the bPRPs and gPRPs family using this high resolution platform.

1.2 Materials and methods

1.2.1 Reagents.

Chemicals and reagents, all of LC − MS grade, were purchased from Merck/Sigma-Aldrich (Darmstadt, Germany), Waters Corporation (Milford, MA), ThermoFischer Scientific (Rockford, IL).

1.2.2 Ethics Statements and Subjects under Study.

The study protocol and written consent form were approved by the Ethical Committee of the Università Cattolica of Rome and has been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All rules were respected and written consent forms were obtained by the donors. Unstimulated whole saliva (WS) was collected from 86 adult healthy donors (40 ± 10 years old, males n = 42, females n = 44).

1.2.3 Salivary Sample Collection

Unstimulated whole saliva (WS) was collected according to a standardized protocol optimized to preserve saliva proteins from proteolytic degradation. Donors did not eat or drink at least 2 h before the collection, which was performed in the morning between 10:00 A.M. and 12:00 A.M. with a soft plastic aspirator. Saliva was transferred into a plastic tube in ice bath, and 0.2% 2,2,2-trifluoroacetic acid (TFA) was immediately added in 1:1 v/v ratio. The solution was then centrifuged at 10000g for 10 min at 4°C. The acidic supernatant was separated from the precipitate and either immediately analyzed by HPLC−ESI−MS or stored at −80 °C until the analysis.

1.2.4 HPLC Low-Resolution ESI−IT−MS Experiments

The acid soluble fractions (33 μL, corresponding to 16.5 μL of whole saliva) of salivary proteins/peptides have been analyzed by reversed-phase (RP)-HPLC-low-resolution ESI-IT-MS apparatus, constituted by a Surveyor HPLC system connected to an LCQ Advantage mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with an ESI source. The chromatographic column was a Vydac (Hesperia, CA) C8 column with 5 μm particle diameter (150 × 2.1 mm). The eluents were the following: (eluent A) 0.056% (v/v) aqueous TFA, and (eluent B) 0.05% (v/v) TFA in acetonitrile/water 80/20. The gradient applied was linear from 0 to 55% of B in 40

min, and from 55 to 100% of B in 10 min, at a flow rate of 0.10 mL/min toward the ESI source. During the first 5 min of separation, eluate was diverted to waste to avoid source contamination because of the high salt concentration. Mass spectra were collected every 3 ms in the m/z range 300−2000 in positive ion mode. The MS spray voltage was 5.0 kV, and the capillary temperature was 260°C. MS resolution was 6000. Deconvolution of averaged ESI-MS spectra was performed by MagTran 1.0 software (Zhang Z, et al. 1998). Average experimental mass values (Mav) were compared with the relative theoretical ones using PeptideMass program available on the Swiss-Prot data bank (https://www.expasy.org/proteomics).

1.2.5 nanoHPLC High-Resolution ESI−MSMS Experiments

For the structural characterization, 67 samples were analyzed by nanoHPLC-high resolution MSMS with an Ultimate 3000 RSLC Nano System HPLC apparatus (Thermo Fisher Scientific, Sunnyvale, CA) coupled to an LTQ-Orbitrap Elite apparatus (Thermo Fisher Scientific). The used chromatographic column was a Zorbax 300SB-C8 (3.5 µm particle diameter; 1.0 × 150 mm). Eluents were: (eluent A) 0.1% (v/v) aqueous formic acid (FA) and (eluent B) 0.1% (v/v) FA in acetonitrile/water 80/20. The gradient was: 0−2 min 5% B, 2−40 min from 5% to 55% B (linear), and 40−45 min from 70% to 99% B, at a flow rate of 50 µL/min. MS and MSMS spectra of intact proteins and peptides were collected in positive mode with the resolution of 60000. The acquisition range was from 350 to 2000 m/z, and the tuning parameters were: capillary temperature 300 °C, source voltage 4.0 kV, S-Lens RF level 60%. In data-dependent acquisition mode the five most intense ions were selected and fragmented by using collision induced dissociation (CID) or higher energy collision dissociation (HCD), with 35% normalized collision energy for 30 ms, isolation width of 5 m/z, and activation q of 0.25. The injected volume was 20 µL. HPLC-ESI−MS and MSMS data were generated by Xcalibur 2.2 SP1.48 (Thermo Fisher Scientific) using default parameters of the Xtract program for the deconvolution. MSMS data were analyzed by both manual inspection of the MSMS spectra recorded along the chromatogram and the Proteome Discoverer 1.4 software elaboration, based on SEQUEST HT cluster as a search engine (University of Washington, licensed to Thermo Electron Corporation, San Jose, CA) against the UniProtKB human data-bank (163,117 entries, release 2018_02). For peptide

matching, high-confidence filter settings were applied: the peptide score threshold was 2.3, and the limits were Xcorr scores greater than 1.2 for singly charged ions and 1.9 and 2.3 for doubly and triply charged ions, respectively. The false discovery rate (FDR) was set to 0.01 (strict) and 0.05 (relaxed), and the precursor and fragment mass tolerance was 10 ppm, and 0.5 Da, respectively. Pyroglutamination from E or Q residues, and serine phosphorylation were selected as dynamic modifications. Due to the difficulties of the automated software to detect with high confidence every bPRPs and their fragments, the structural information derives in part from manual inspections of the MSMS spectra, obtained either by CID or HCD fragmentation, against the theoretical ones generated by MS-Product software available at the Protein Prospector Web site (http://prospector.ucsf.edu/prospector/mshome.htm).

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://www.ebi.ac.uk/pride) via the PRIDE (Vizcaíno J.A, et al. 2016) partner repository with the data set identifier PXD009813.

1.3 Results

For their structural similarity, bPRPs and gPRPs are usually detected in the RP-HPLC-ESI-MS TIC (total ion current) profile as a characteristic cluster in the elution window comprised between 14.5-20.5 min under the experimental conditions used. Fig. 1.4A shows a typical TIC profile of the soluble acidic fraction of salivary proteins obtained by RP-HPLC-low resolution ESI-MS analysis. It is relevant to outline that many bPRPs elute closely, often in a single unresolved chromatographic peak and that some of them, as P-J and P-F, are difficult to separate with the common chromatographic methods due to their structural similarity. The main protein species detectable in the bPRP cluster are shown in the enlargement of Fig. 1.4B.

bPRPs (and gPRPs) encoded by *PRB1*, *PRB2* and *PRB4* loci are secreted as fragments of pro-proteins. The proteins and peptides identified in the different salivary samples are described in the following sections with the support of Fig. 1.5-1.8 and Tables 1.3-1.11, in which the elution times measured in HPLC-high resolution ESI-MS TIC profiles, under our experimental conditions, are also reported.

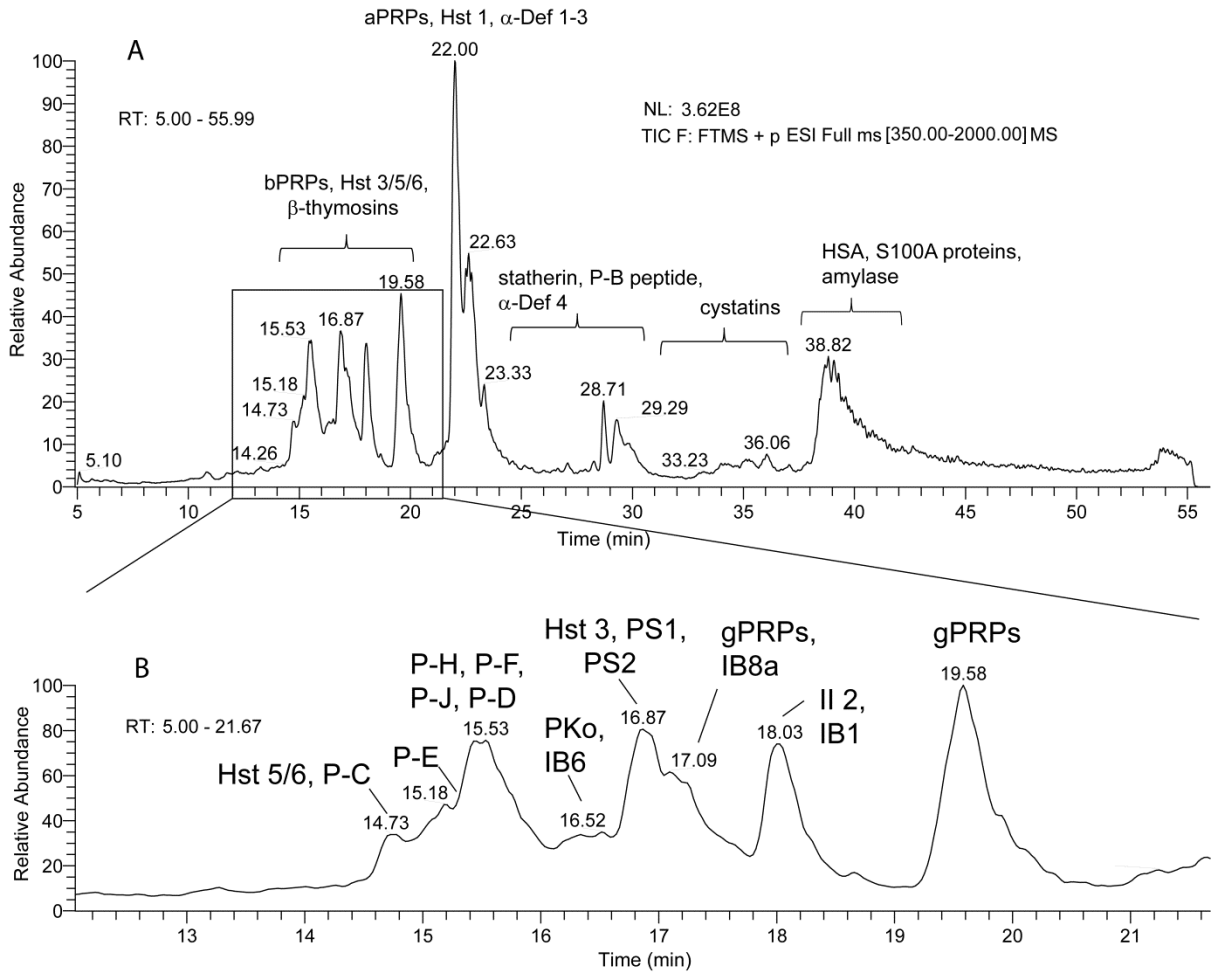


Figure 1.4 Panel A: typical HPLC-ESI-MS total ion current (TIC) profile of the acid soluble fraction of whole adult human saliva reporting the elution times of the principal families of salivary peptides. Basic (bPRPs) and glycosylated basic proline-rich proteins (gPRPs) for their structural similarity elute in a cluster comprised between 14.5-20.5 min under the experimental conditions used. Panel B: enlargement of the elution cluster of bPRPs and gPRPs. On the profile the elution time of the main parent bPRPs is reported (see text).

NL: normalization level; α-Def: α-defensins 1-4; Hst: histatin.

1.3.1Products of *PRB1* locus

Fig. 1.5 shows the parent peptides deriving from the different alleles detectable in the *PRB1* locus in the western population, which are P-E (also named IB-9), II-2, P-Ko, IB-6, Ps-1 and Ps-2. Table 1.3 reports their mass values, elution times, and sequences together with their frequencies within the cohort of 86 analyzed samples. The sequences of P-E, II-2 and IB-6 have been confirmed by high resolution MSMS analysis, which these bPRPs previously characterized with top-down proteomic platforms by our group (Messana I, et al. 2008a; Castagnola, M, et al. 2012b) and other research groups (Vitorino R, et al. 2009; Halgand F, et al. 2012). MS and MSMS data obtained on IB-6 did not correspond to the sequence reported on UniProtKB data bank (code P04280), for the presence of a serine instead of an alanine at position 63. Our top-down proteomic approach allowed to characterize for the first time by MSMS the intact Ps-1 protein (experimental monoisotopic $[M+H+]^{1+}$ value 23445.9 *m/z*), previously identified by our group only by a bottom-up approach (Cabras T, et al. 2012b).

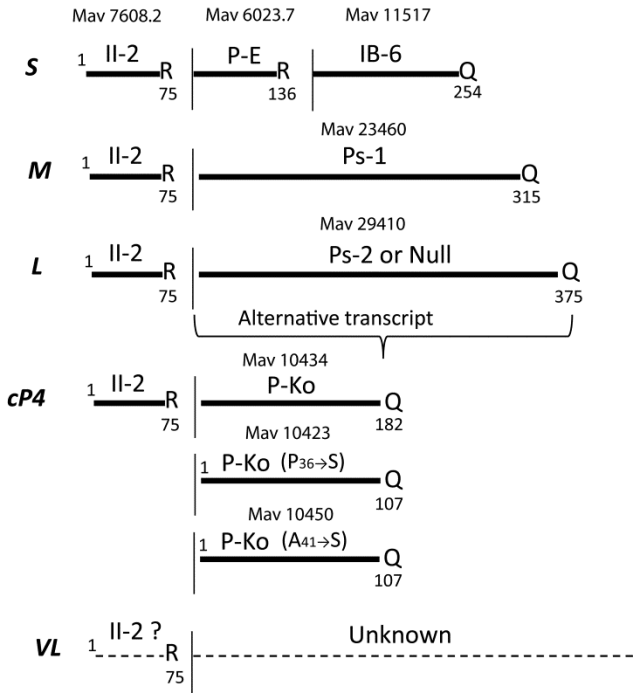


Figure 1.5 Schematic representation of the human salivary *PRB1* locus and its alleles, showing the coding regions for parent bPRPs.

Table 1.3 Average (Mav) and $[M+H]^{1+}$ monoisotopic masses, elution times, frequency, and sequence of the products of *PRB-1* locus (UniprotKB code P04280). The proteins characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Freq. (n = 86)	Sequence[a]
P-E (or IB-9)	6023.7 ± 0.7 (6023.69)	6021.09 ± 0.03 (6021.088)	14.9	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNRPQG PPPPGKPQGP PPQGDKSRSP R
II-2	7608.2 ± 0.8 (7608.19)	7604.69 ± 0.04 (7604.712)	19.2	86	<QNLNEDV**S**QE ESPSLIAGNP QGPSPQGGNK PQGPPPPPGK PQGPPPQGGN KPQGPPPPGK PQGPPPQGDK SRSPR
P-Ko	10434 ± 1.1 (10433.57)	10428.29 ± 0.05 (10428.285)	16.0	65	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKPQGPP AQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
P-Ko P_{36}→S	10423 ± 1.0 (10423.46)	10418.28 ± 0.05 (10418.264)	15.8	1	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKSQGPP AQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
P-Ko[b] A_{41}→S	10450 ± 1.1 (10449.57)	10444.30± 0.05 (10444.279)	16.0	11	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKPQGPP SQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
IB-6	11517 ± 1.2 (11516.67)	11510.80 ± 0.06 (11510.799)	16.7	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PAQGGSKSQS ARSPPGKPQG PPQQEGNNPQ GPPPPAGGNP QQPQAPPAGQ PQGPPRPPQG GRPSRPPQ
Ps-1[c]	23460 ± 3 (23459.07)	23445.9 ± 0.11 (23445.859)	17.6	52	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDKSQSP RSPPGKPQGP PPQGGNQPQG PPPPPGKPQG PPQQGGNRPQ GPPPPGKPQG PPPQGDKSRS PQSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGGNKP QGPPPPGKPQ GPPAQGGSKS QSARAPPGKP QGPPQQEGNN PQGPPPPAGG NPQQPQAPPA GQPQGPPRPP QGGRPSRPPQ
Ps-2	29410 ± 4 (29408.72)	29391.9 ± 0.14 (29391.881)	17.6	5	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDKSQSP RSPPGKPQGP PPQGGNQPQG PPPPPGKPQG PPPQGGNKPQ GPPPPGKPQG PPPQGDKSQS PRSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPQQGGNRP QGPPPPGKPQ GPPPQGDKSR SPQSPPGKPQ GPPPQGGNQP QGPPPPPGKP QGPPPQGGNK PQGPPPPGKP QGPPAQGGSK SQSARAPPGK PQGPPQQEGN NPQGPPPPAG GNPQQPQAPP AGQPQGPPRP PQGGRPSRPP Q

[a] <Q: pyro-glutamic acid; **S**: phosphorylated Ser. [b] UniprotKB code G5E9X6; [c] UniprotKB code Q86YA1.

On the contrary, it was not possible to confirm with confidence Ps-2 identity by MSMS sequencing, but the experimental monoisotopic $[M+H^+]^{1+}$ value of 29391.9 ± 0.14 *m/z* was in perfect agreement with the theoretical one ($[M+H^+]^{1+}$ 29391.881 *m/z*) reported in databases (*PRB1-L* allele, UniProtKB code P04280). Until now we were neither able to detect in whole saliva potential peptides or proteins deriving from the *PRB1-VL* nor the peptide with an average mass of 8391.2 Da that should originate from an allele *PRB1-L cP5*, a differentially spliced transcript of *PRB1-L* allele predicted on the basis of gene sequencing (Maeda, N, et al. 1985). Moreover, we were able to detect and characterize the P-Ko protein originated from the *PRB1-L cP4* (Table 1.3), and, during our survey on adult salivary samples, to identify two variants of P-Ko. The P_{36}→S variant, found in one sample (out of 86) was characterized from the inspection of the MSMS CID fragmentation spectra on the $[M+8H]^{8+}$ (1309.9 *m/z*), and $[M+9H]^{9+}$ (1159.1 *m/z*) multiply-charged ions. The attribution of the substitution to P_{36} among the multiple proline residues in the P-Ko sequence, was based on the detection of the b_{35} (exp. 3323.74; theor. 3323.740 *m/z*), b_{37} (exp. 3538.84; theor. 3538.830 *m/z*), y_{71} (exp. 7008.51; theor. 7008.500 *m/z*) and y_{72} (exp. 7095.54; theor. 7095.532 *m/z*) ions. The A_{41}→S variant of P-Ko (Table 1.3) was detected in 11 samples (out of 86). The b and y ions detected in the MSMS CID spectra performed on the $[M+8H]^{8+}$ (1307.2 *m/z*) ion, restricted the substitution to A_{41} and A_{50} residues, but some internal fragments were diagnostic for the A_{41} substitution: particularly the fragments $QGP_{30}PPPGKPQGPP_{40}SQ$ (exp. 1450.74; theor. 1450.744 *m/z*), and $PGKPQGPP_{40}SQGGSKSKSQ_{50}$-$NH_3$ (exp. 1501.76; theor. 1501.739 *m/z*) in agreement with a serine residue at position 41 and the fragments $GSKSQSA_{50}RAPPGKPQGP_{60}$-$H_2O$ (exp. 1613.82; theor 1613.850 *m/z*), and $GSKSQSA_{50}RAPPGKPQGP_{60}$-$NH_3$ (exp. 1614.81; theor. 1614.835 *m/z*) in agreement with an alanine residue at position 50. Table 1.4 reports the most common derivatives from the main proteoforms of *PRB1* locus, principally from II-2. Six out of 11 peptides of Table 1.4 were characterized in this study for the first time, while the others have been described also in previous studies (Messana I, et al. 2008a; Helmerhorst E.J, et al. 2008). A variant of II-2 peptide, lacking the proline residue at position 39 has been described (Halgand F, et al. 2012), but we were not able to detect it in any of the samples analyzed in the present study. II-2 was detected in all the samples analyzed, as expected since it originates from all the *PRB1* alleles (Fig. 1.5), P-Ko was highly frequent in our cohort of healthy adult population being detected in

68 subjects, of which 56 homozygous for the main P-Ko variant, 2 homozygous for the A_{41}→S variant and 1 for the P_{36}→S variant, and 9 heterozygous P-Ko/P-Ko A_{41}→S. Also the Ps-1 protein was frequently detected (56 out of 86 subjects) while the other *PRB1* products, P-E and IB-6 from *PRB1-S* allele, and Ps-2 form *PRB1-L* allele, were rarely observed.

Table 1.4 Average (Mav) and $[M+H]^{1+}$ monoisotopic masses, elution times of the main derivatives of the products of *PRB-1* locus (UniprotKB code P04280). Peptides characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)
II-2 (Fr. 18-32)[a]	1462.7 ± 0.2 (1462.54)	1462.71 ± 0.01 (1462.703)	8.9
II-2 (Fr. 1-23) non phosph. pyro-Gln	2406.3 ± 0.3 (2406.45)	2406.11 ± 0.02 (2406.106)	21.8
II-2 (Fr. 18-42)[a]	2415.2 ± 0.3 (2415.67)	2415.22 ± 0.02 (2415.216)	15.9
II-2 (Fr. 1-23) pyro-Gln, S_8(phosph)	2486.5 ± 0.3 (2486.43)	2486.07 ± 0.02 (2486.072)	20.5
II-2 (Fr. 18-53)	3474.4 ± 0.6 (3473.84)	3472.75 ± 0.03 (3472.747)	16.2
II-2 (Fr. 20-67)	4635.4 ± 0.8 (4635.18)	4633.41 ± 0.05 (4633.381)	17.0
II-2 (Fr. 18-75)	5690.9 ± 0.6 (5690.35)	5687.92 ± 0.03 (5687.783)	16.1
P-E Des R_{61}[b]	5867.5 ± 0.6 (5867.50)	5864.98 ± 0.03 (5864.987)	14.9
II-2 Des $R_{72}SPR_{75}$ pyro-Gln, S_8(phosph)	7111.7 ± 0.8 (7111.68)	7108.43 ± 0.04 (7108.425)	19.1
II-2 Des R_{75} pyro-Gln, S_8(phosph)[b]	7452.0 ± 0.8 (7452.01)	7448.61 ± 0.04 (7448.612)	19.2
II-2 non phosph. pyro-Gln[b]	7528.3 ± 0.8 (7528.21)	7524.75 ± 0.04 (7524.746)	19.7

Identified also in: [a], ref. (Cabras T, et al. 2012a); [b], ref. (Messana I, et al. 2004; Messana I, et al. 2008a).

1.3.2 Products of *PRB2* locus

Fig. 1.6 shows the bPRPs peptides deriving from the different alleles of the *PRB2* locus detected in the western population, namely P-H (or IB-4), P-F, P-J, IB-1, IB-8a Con 1^- and IB-8a Con 1^+.

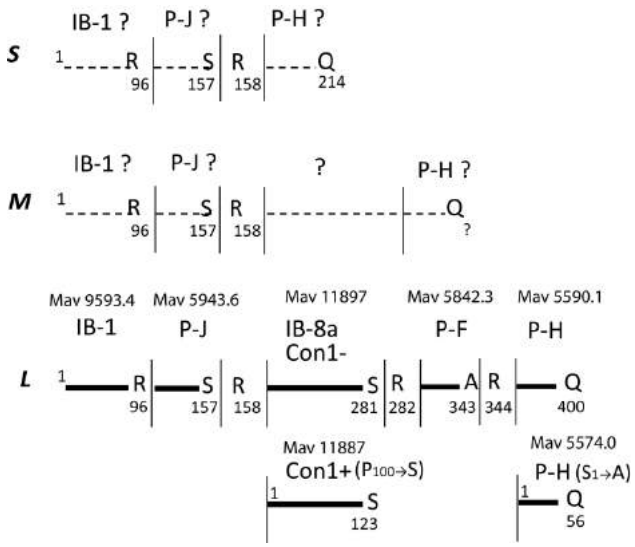


Figure 1.6 Schematic representation of the human salivary *PRB2* locus and its alleles, showing the coding regions for parent bPRPs. Even if the sequence of the S and M alleles is not known probably they encode IB-1, P-J and P-H peptides, because they are detectable in all the salivary samples investigated.

The sequences, mass values, elution times, and detection frequencies of the bPRPs encoded by *PRB2* locus are reported in Table 1.5. High resolution MSMS analysis confirmed the sequences previously characterized (Messana I, et al. 2008a; Castagnola M, et al. 2012b) and allowed to identify a new P-H $S_1 \rightarrow A$ variant not reported in UniProtKB database. The MSMS experiments performed by HCD fragmentation on the multiply-charged ion $[M+5H]^{5+}$ (1115.57 *m/z*) confirmed the S→A substitution at position 1.

Table 1.5 Average (Mav) and $[M+H]^{1+}$ monoisotopic masses, elution times, frequency, and sequence of the principal products of *PRB-2* locus (UniprotKB code P02812). Peptides characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Freq. (n = 86)	Sequence[a]
P-H S_1→A	5574.0 ± 0.6 (5574.14)	5571.79 ± 0.02 (5571.788)	15.2	9	APPGKPQGPP QQEGNNPQGP PPPAGGNPQQ PQAPPAGQPQ GPPRPPQGGR PSRPPQ
P-H (or IB-4)	5590.1 ± 0.6 (5590.10)	5587.77 ± 0.02 (5587.783)	15.2	85	SPPGKPQGPP QQEGNNPQGP PPPAGGNPQQ PQAPPAGQPQ GPPRPPQGGR PSRPPQ
P-F (or IB-8c)	5842.5 ± 0.7 (5842.49)	5840.00 ± 0.02 (5839.992)	14.7	83	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGGSKSRS A
P-J	5943.6 ± 0.7 (5943.56)	5941.00 ± 0.02 (5941.003)	14.5	86	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSRS S
IB-1	9593.4 ± 1.0 (9593.38)	9588.61 ± 0.04 (9588.703)	19.4	86	<QNLNEDV**S**QE ESPSLIAGNP QGAPPQGGNK PQGPPSPPGK PQGPPPQGGN QPQGPPPPPG KPQGPPPQGG NKPQGPPPPG KPQGPPPQGD KSRSPR
IB-8a $Con1^-$ P_{100}	11897 ± 2 (11896.16)	11890.05 ± 0.05 (11890.035)	16.7	42	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSQS ARSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGGNK**P** QGPPPPGKPQ GPPPQGGSKS RSS
IB-8a $Con1^+$ S_{100}	11887 ± 2 (11886.12)	11880.02 ± 0.05 (11880.014)	17.6	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSQS ARSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGG**NKS** QGPPPPGKPQ GPPPQGGSKS RSS
IB-8a $Con1^+$ S_{100} Glycoform-1	13291 ± 2 (13290.42)	ND (13283.521)	15.6	7	IB-8a $Con1^+$ S_{100} sequence + $dHex_1$+Hex_4+$HexNAc_3$
IB-8a $Con1^+$ S_{100} Glycoform-2	13656 ± 2 (13655.76)	ND (13648.653)	15.6	15	IB-8a $Con1^+$ S_{100} sequence + $dHex_1$+Hex_5+$HexNAc_4$
IB-8a $Con1^+$ S_{100} Glycoform-3	13802 ± 2 (13801.90)	ND (13794.711)	15.6	23	IB-8a $Con1^+$ S_{100} sequence + $dHex_2$+Hex_5+$HexNAc_4$
IB-8a $Con1^+$ S_{100} Glycoform-4	13948 ± 2 (13948.04)	ND (13940.769)	15.6	32	IB-8a $Con1^+$ S_{100} sequence + $dHex_3$+Hex_5+$HexNAc_4$
IB-8a $Con1^+$ S_{100} Glycoform-5	14093 ± 2 (14094.18)	ND (14086.827)	15.6	19	IB-8a $Con1^+$ S_{100} sequence + $dHex_4$+Hex_5+$HexNAc_4$
IB-8a $Con1^+$ S_{100} Glycoform-6	14239 ± 2 (14240.33)	ND (14232.885)	15.6	2	IB-8a $Con1^+$ S_{100} sequence + $dHex_5$+Hex_5+$HexNAc_4$

[a] <Q: pyroglutamic acid; **S**: phosphorylated Ser. dHex: deoxy-hexose, probably fucose; Hex: hexose, probably mannose or galactose; HexNAc: N-acetyl-hexosamine, probably N-acetyl-glucosamine. **NKS**: N-glycosylation consensus sequence; ND not determinable

The two proteoforms of IB-8a detected in human saliva derive from a SNP responsible for S_{100}→P substitution (Azen E.A, et al. 1996). IB-8a carrying P_{100} is not glycosylated and is named Con1$^-$, because it is not able to bind concanavalin A (Azen E.A, et al. 1996). IB-8a Con1$^+$ carries a serine at position 100, and it is glycosylated on N_{98}. The six different glycosylated protein species of IB-8a Con1$^+$ characterized by HPLC-ESI-MS in adult human saliva together with the non-glycosylated protein are reported in Table 1.5 (Cabras T, et al. 2012a). Five of the glycosylated species carry a biantennary *N*-linked glycan fucosylated in the innermost N-acetylglucosammine of the core, and show from zero to four additional fucoses in the antennal region. The sixth glycoform carries a monoantennary monofucosylated oligosaccharide. IB-8a was detected in 64 subjects (out of 86), 25 were homozygous for IB-8a Con 1$^-$ and 22 homozygous for IB-8a Con 1$^+$, while 17 subjects exhibited both the variants. Among the 39 subjects expressing IB-8a Con 1$^+$, 24 showed only the glycosylated proteoforms, while 15 showed also the apo-protein. While in the HPLC-ESI low resolution MS analyses it was possible to determine the Mav of the glycoforms of IB-8a-Con1$^+$, it was not possible to determine the monoisotopic mass value by deconvolution of the high-resolution ESI-spectra. IB-1, P-J, and P-H were detected in all the 86 samples analyzed, while P-F showed a frequency slightly lower (83 out of 86 subjects) (Table 1.5). The S_1→A variant of P-H peptide was detected in whole saliva of 9 adult subjects, with one of them homozygous for the S_1→A variant, and the other 8 heterozygous for P-H/P-H S_1→A. Several peptides characterized in this study derived specifically from the fragmentation of bPRPs expressed by *PRB2* locus, and they are reported in Table 1.6. Among them, five peptides were identified for the first time in this study, while the other three were also characterized in previous top-down investigations (Messana I, et al. 2008a; Helmerhorst EJ, et al. 2008; Manconi B, et al. 2016b).

Table 1.6 Average (Mav) and $[M+H]^{1+}$ monoisotopic masses, elution times of the main derivatives of the products of *PRB-2* locus (UniprotKB code P02812). The peptides characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)
P-H (Fr. 8-56)	4898.5 ± 0.5 (4898.34)	4896.42 ± 0.03 (4896.417)	17.7
P-H (Fr. 1-18)[a]	1856.9 ± 0.4 (1856.97)	1856.89 ± 0.02 (1856.89)	10.5
P-F Des $R_{59}SA_{61}$	5528.4 ± 0.6 (5528.19)	5525.81 ± 0.03 (5525.821)	16.3
P-J Des $R_{59}SS_{61}$	5613.4 ± 0.6 (5613.25)	5610.84 ± 0.03 (5610.838)	16.3
IB-1 (Fr. 33-42)[a]	961.1 ± 0.2 (961.09)	961.51 ± 0.01 (961.51)	13.4
IB-1 (Fr. 18-32)	1446.7 ± 0.2 (1446.54)	1446.71 ± 0.01 (1446.708)	8.1
IB-1 Des $R_{93}SPR_{96}$ pyro-Gln, S_8(phosph)	9097.0 ± 1.0 (9096.88)	9092.42 ± 0.05 (9092.416)	19.1
IB-1 Des R_{91} pyro-Gln (N-term) S_8(phosph)[b]	9437.0 ± 1.0 (9437.20)	9432.61 ± 0.05 (9432.602)	19.4

Identified also in: [a], ref. (Cabras T, et al. 2012a); [b], ref. (Messana I, et al. 2004; Messana I, et al. 2008a).

1.3.3 Products of the *PRB3* locus

Fig. 1.7 shows the asset of the *PRB3* locus. The sequence and the possible glycosylation sites of the most common glycoproteins codified by *PRB3* locus, namely Gl-1, Gl-2 and Gl-3 are reported in Table 1.7.

Each Gl protein carries a different number of putative *O*- and *N*-glycosylation sites depending on the length of the polypeptide backbone (Carpenter G.H, et al. 1999; Gillece-Castro B.L, et al. 1991). Due to the high heterogeneity of the glyco-moiety, ESI spectra of the intact glycoproteins are crowded of *m/z* signals and cannot be resolved by the deconvolution software. Therefore, until now we were unable to detect masses pertaining to these proteins by a top-down platform. Surprisingly, the Gl-2 (or PRP-3M) glycoforms were the only bPRPs detectable in significant amounts in newborns whole saliva (Manconi B, et al. 2016a).

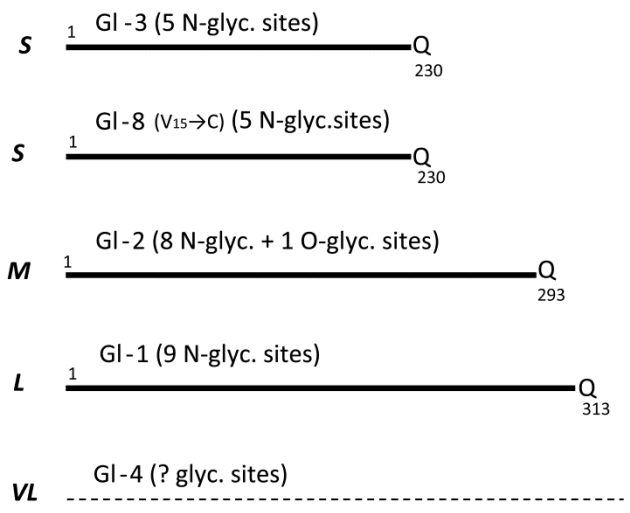


Figure 1.7 Schematic representation of the human salivary *PRB3* locus and its alleles, showing the coding regions for parent gPRPs.

Table 1.7 Sequence and potential glycosylation sites of the products of *PRB3* locus (UniprotKB code Q04118).

Name	Sequence[a]
Gl-3 or PRP-3S (5 N-glycosyl. sites)	<QSLNEDV**S**QE ESPSVISGKP EGRRPQGGNQ PQR**T**PPPPGK PEGRPPQGG**N** **QS**QGPPPRPG KPEGPPPQGG **NQS**QGPPPRP GKPEGQPPQG G**NQS**QGPPPR PGKPEGPPPQ GG**NQS**QGPPP RPGKPEGPPP QGG**NQS**QGPP PRPGKPEGSP SQGGNKPQGP PPHPGKPQGP PPQEGNKPQR PPPPGRPQGP PPPGGNPQQP LPPPAGKPQG PPPPPQGGRP HRPPQGQPPQ
Gl-2 or PRP-3M (8 N-glycosyl. sites)	<QSLNEDV**S**QE ESPSVI**S**GKP EGRPPQGGNQ PQR**T**PPPPGK PEGPPPQGG**N** **QS**QGPPPRPG KPEGQPPQGG **NQS**QGPPPRP GKPEGPPPQG G**NQS**QGPPPR PGEPEGPPPQ GG**NQS**QGPPP HPGKPEGPPP QGG**NQS**QGPP PRPGKPEGPP PQGG**NQS**QGP PPRPGKPEGP PPQGG**NQS**QG PPPRPGKPEG PPPQGG**NQS**Q GPPPRPGKPE GSPSQGGNKP RGPPPHPGKP QGPPPQEGNK PQRPPPPRRP QGPPPPGGNP QQPLPPPAGK PQGPPPPPQG GRPHRPPQGQ PPQ
Gl-1 or PRP-3L (9 N-glycosyl. sites)	<QSLNEDV**S**QE ESPSVISGKP EGRRPQGGNQ PQR**T**PPPLGK PEGRPPQGG**N** **QS**QGPPPRPG KPEGPPPQGG **NQS**QGPPPRP GKPEGQPPQG G**NQS**QGPPPR PGKPEGPPPQ GG**NQS**QGPPP RPGEPEGPPP QGG**NQS**QGPP PHPGKPEGPP PQGG**NQS**QGP PPRPGKPEGP PPQGG**NQS**QG PPPRPGKPEG PPPQGG**NQS**Q GPPPRPGKPE GPPPQGG**NQS** QGPPPRPGKP EGSPSQGGNK PRGPPPHPGK PQGPPPQEGN KPQRPPPPRR PQGPPPPGGN PQQPLPPPAG KPQGPPPPPQ GGRPHRPPQG QPPQ

[a] <Q: pyroglutamic acid; **S**: phosphorylated S; **NQS**: N-glycosylation site; T_{34} of Gl-2 is the O-glycosylation site (Kauffman D.L, et al. 1993). The sequence and the glycosylation sites of Gl-2 have been confirmed experimentally (Kauffman D.L, et al. 1993). The PTMs of Gl-3 and Gl-1 are supposed by similarity.

Characterization of Gl-2 glycoforms was performed by RP-HPLC high-resolution ESI-MS before and after *N*-deglycosylation with PNGase F of an enriched fraction isolated from newborns saliva. Furthermore, peptides obtained by Glu-C digestion were submitted to MSMS sequencing. In this way it was possible to characterize the peptide backbone and to identify the *N*- and *O*-glycosylation sites. The heterogeneous mixture of the glycoforms derived from the combination of eight different neutral and sialylated glycans *O*-linked to T_{34}, and thirty-three different glycans *N*-linked to N_{50}, N_{71}, N_{92}, N_{113}, N_{134}, N_{155}, N_{176} and N_{197} residues. It is plausible that similar glycoforms are present on Gl-1 and Gl-3, by similarity.

1.3.4 Products of the *PRB4* locus

Fig. 1.8 shows the asset of the *PRB4* locus. Among the products of this locus, the two variants of P-D peptide, carrying either P or A at position 32 were the unique detectable under our experimental conditions. The other products of *PRB4* locus are highly glycosylated proteins not completely characterized until now and their sequences (reported in Table 1.8) derive from gene sequencing (Stubbs M, et al. 1998; Kauffman D.L, et al. 1993).

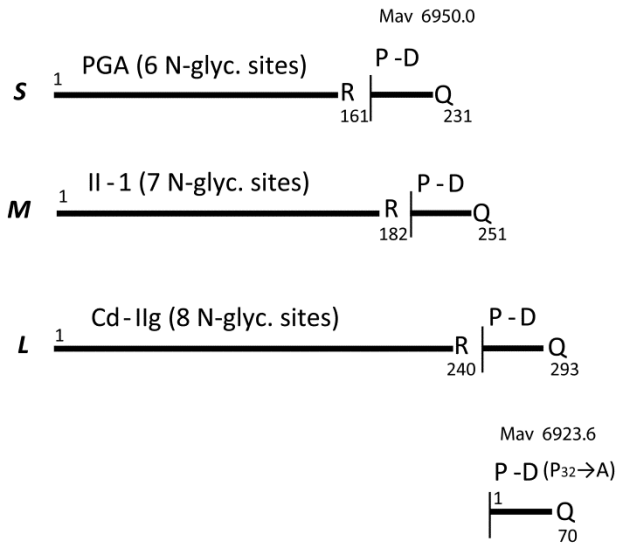


Figure 1.8 Schematic representation of the human salivary *PRB4* locus and its alleles, showing the coding regions for parent bPRPs and gPRPs.

Table 1.8 Average (Mav) and $[M+H]^{1+}$ monoisotopic mass values, elution times, frequency, and sequence of the products of *PRB-4* locus and their derivatives (UniprotKB code P10163). The peptides characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Frequency (n = 86)	Sequence[a]
P-D (Fr. 49-60)	1153.6 ± 0.2 (1153.32)	1153.61± 0.01 (1153.611)	15.7	31	GPPPPPQGGRPP
P-D (Fr. 1-18)[b]	1871.0 ± 0.3 (1871.05)	1870.94 ± 0.01 (1870.941)	13.5	4	SPPGKPQGPP QQEGNKPQ
P-D (Fr. 59-70)[c]	2242.2 ± 0.3 (2241.52)	2241.16 ± 0.02 (2241.164)	15.9	38	GPPPPPQGGRPPRPAQGQQPPQ
P-D (Fr. 1-27)[b,c]	2727.3 ± 0.4 (2727.06)	2726.42 ± 0.01 (2726.401)	15.4	23	SPPGKPQGPP QQEGNKPQGP PPPGKPQ
P-D (Fr. 1-60)	5861.5 ± 0.7 (5861.58)	5859.00 ± 0.03 (5859.002)	15.2	3	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP PPGGNPQQPQ APPAGKPQGP PPPPQGGRPP
P-D (P_{32}→A)	6923.6 ± 0.8 (6923.69)	6920.54 ± 0.04 (6920.538)	15.9	18	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP P**A**GGNPQQPQ APPAGKPQGP PPPPQGGRPP RPAQGQQPPQ
P-D (or IB-5)	6950.0 ± 0.8 (6949.73)	6946.55 ± 0.04 (6946.554)	16.7	70	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP P**P**GGNPQQPQ APPAGKPQGP PPPPQGGRPP RPAQGQQPPQ
Glycos. Protein A	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGG**N QS**QGPPPPPG KPEGRPPQGG **NQS**QGPPPHP GKPERPPPQG G**NQS**QGTPPP PGKPERPPPQ GG**NQS**HRPPP PPGKPERPPP QGG**NQS**QGPP PHPGKPEGPP PQEG**NKS**RSA R
II-1	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGG**N QS**QGPPPPPG KPEGRPPQGG **NQS**QGPPPHP GKPERPPPQG G**NQS**QGTPPP PGKPEGRPPQ GG**NQS**QGPPP HPGKPERPPP QGG**NQS**HRPP PPPGKPERPP PQGG**NQS**QGP PPHPGKPEGP PPQEG**NKS**RS AR
Cd-IIg	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGG**N QS**QGPPPPPG KPEGRPPQGG **NQS**QGPPPHP GKPERPPPQG G**NQS**QGPPPH PGKPESRPPQ GGHQSQGPPP TPGKPEGPPP QGG**NQS**QGTP PPPGKPEGRP PQGG**NQS**QGP PPHPGKPERP PPQGG**NQS**HR PPPPPGKPER PPPQGG**NQS**Q GPPPHPGKPE GPPPQEG**NKS** RSAR

[a] <E: pyroglutamic acid; S: phosphorylated Ser (hypothetical, by similarity); [b], identified also in ref. (Helmerhorst EJ, et al. 2008); [c], ref. (Vitorino R, et al. 2009)

As for the other glycosylated bPRPs, their ESI spectra were crowded for the heterogeneous glycan moieties and it was not possible to establish their molecular masses by our mass spectrometry. Table 1.6 reports the mass values, the sequences and the elution times of the bPRPs encoded by *PRB4* locus and the frequencies determined in healthy adults. Top-down high resolution MSMS experiments allowed to confirm the sequences of the two P-D variants, which we have already characterized (Castagnola M, et al. 2012b; Messana I, et al. 2008a) and to identify some P-D fragments, two of them described for the first time in this study. P-D peptide was detected in 75 subjects, 57 were homozygous for the main P-D variant, 5 for the P_{32}→A variant, and 13 were heterozygous P-D/P-D P_{32}→A (Table 1.8). In the Table 1.8 the sequence reported in the literature for the glycosylated proteins expressed by the three *PRB4* alleles is also shown.

1.3.5 Other fragments of bPRPs.

The sequences, mass values, elution times and the possible origin of 34 bPRP fragments eluting in the bPRPs chromatographic cluster are reported in Table 1.9. Among them, 21 were never detected in previous investigations, while the others have already been described (Messana I, et al. 2008a; Vitorino R,et al. 2009; Helmerhorst EJ, et al. 2008; Vitorino R, et al. 2010; Huq N.L, et al. 2007). Furthermore, fragmentation of bPRPs generates a large number of very small and polar fragments that elute before the bPRPs cluster, namely between 4 and 14 minutes. A list of 36 naturally occurring peptides detected in adult human saliva by HPLC-ESI-MS is reported in Table 1.10. Among them, 17 were never detected in previous investigations while 19 were previously characterized (Huq N.L, et al. 2007, Helmerhorst EJ, et al. 2008; Messana I, et al. 2008a; Vitorino R, et al. 2009).

Table 1.9 List of the most common fragments from bPRPs eluting in the bPRPs cluster (14.0-20.0 min) that cannot be attributed to a specific parent bPRP. The peptides characterized for the first time in this study are reported in bold.

Sequence	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	El. time (min± 0.4)	Possible origin
QPLPPPAGKPQ[a]	1129.6 ± 0.2 (1129.34)	1129.64 ± 0.01 (1129.636)	15.7	Gl-3
GPPPPAGGNPQQPQ[a,b]	1341.7 ± 0.2 (1341.45)	1341.66 ± 0.01 (1341.655)	14.3	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1,
GPPPPGKPQGPPPQ	1350.7 ± 0.2 (1350.56)	1350.72 ± 0.02 (1350.716)	14.8	P-E, II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GGNQPQGPPPPPGKPQ[a]	1552.8 ± 0.2 (1552.73)	1552.79 ± 0.02 (1552.787)	15.2	IB-1, IB-6, P-F, P-J, P-E, Ps-1, Ps-2, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPRPPQGGRPSRPPQ	1680.9 ± 0.2 (1680.91)	1680.90 ± 0.02 (1680.904)	14.7	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPPGKPQGPPPQGDKS	1737.9 ± 0.3 (1737.95)	1737.89 ± 0.02 (1737.892)	14.0	II-2, P-E, Ps-1, Ps-2, IB-1
SPPGKPQGPPPQGGNQPQ[a,b,c,d,e]	1767.9 ± 0.3 (1767.92)	1767.89 ± 0.02 (1767.877)	14.3	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGKPQGPPAQGGSKSQ	1869.1 ± 0.3 (1869.08)	1868.96 ± 0.02 (1868.961)	17.2	IB-6, P-Ko, Ps-1, Ps-2
GPPPQGGNKPQGPPPPGKPQ[a,e]	1932.2 ± 0.4 (1932.17)	1932.01 ± 0.03 (1932.009)	14.7	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$
PQGGNKPQGPPPPGKPQGPP	1932.0 ± 0.4 (1932.17)	1932.01 ± 0.03 (1932.009)	14.5	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGGNPQQPLPPPAGKPQGPP	2028.3 ± 0.3 (2028.31)	2028.08 ± 0.02 (2028.066)	18.2	Gl-1, Gl-2, Gl-3
GPPPPGGNPQQPLPPPAGKPQ[a,d]	2028.3 ± 0.3 (2028.31)	2028.07 ± 0.02 (2028.066)	18.2	Gl-1, Gl-2, Gl-3
GPPPQGGNQPQGPPPPPGKPQ[a,e]	2029.2 ± 0.4 (2029.24)	2029.03 ± 0.03 (2029.025)	16.0	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PQGGNQPQGPPPPPGKPQGPP	2029.4 ± 0.3 (2029.26)	2029.03 ± 0.02 (2029.025)	16.0	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPPGKPQGPPPQGGNKPQ	2029.4 ± 0.3 (2029.30)	2029.06 ± 0.02 (2029.061)	14.8	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGKPQGPPPQGGNKPQGPPP	2029.4 ± 0.3 (2029.30)	2029.06 ± 0.02 (2029.061)	14.9	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGKPQGPPPQGDKSRSP	2078.3 ± 0.3 (2078.33)	2078.08 ± 0.02 (2078.078)	14.5	II-2, P-E, Ps-1, Ps-2, IB-1

Sequence	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	El. time (min± 0.4)	Possible origin
GPPPQEGNKPQRPPPPGRPQ	2132.1 ± 0.3 (2131.40)	2131.12 ± 0.02 (2131.115)	14.6	GL-3
GPPPPPQGGRPHRPPQGQPPQ[d]	2180.1 ± 0.3 (2179.45)	2179.13 ± 0.02 (2179.127)	14.9	Gl-1, Gl-2, Gl-3
GPPPPAGGNPQQPQAPPAGQPQGPP[a]	2339.6 ± 0.3 (2339.57)	2339.15 ± 0.02 (2339.153)	18.1	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPPQGGNQPQGPPPPP GKPQ[c]	2721.0 ± 0.4 (2721.05)	2720.40 ± 0.03 (2720.390)	16.3	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGKPQGPPPQGGNKPQGPPPPGK PQGPPP	2885.7 ± 0.5 (2885.31)	2884.52 ± 0.03 (2884.522)	16.1	II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPPGKPQGPPPQGGNKPQGPPPPG KPQGPP	2885.7 ± 0.5 (2885.31)	2884.54 ± 0.03 (2884.522)	16.1	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPQGGNKPQGPPPPGKPQGPPP QGDKSRSP[c]	3136.5 ± 0.6 (3136.50)	3135.61 ± 0.03 (3135.608)	15.2	II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGGNPQQPLPPPAGKPQGPP PPPQGGRPH	3203.7 ± 0.6 (3203.64)	3202.67 ± 0.03 (3202.666)	18.1	Gl-1, Gl-2, Gl-3
GPPQQEGNNPQGPPPPAGGNPQQP QAPPAGQPQGPP	3486.7 ± 0.6 (3486.75)	3485.66 ± 0.03 (3485.658)	18.4	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPPQGGNQPQGPPPPP GKPQGPPPQGGNKPQ	3779.2 ± 0.6 (3779.22)	3777.92 ± 0.04 (3777.921)	16.4	IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGGNPQQPLPPPAGKPQGPPPP PQGGRPHRPPQGQPPQ[d]	4190.2 ± 0.7 (4189.71)	4188.17 ± 0.04 (4188.175)	18.1	Gl-1, Gl-2, Gl-3
SPPGKPQGPPPQGGNQPQGPPPPP GKPQGPPPQGGNKPQGPPPPGKP Q	4635.2 ± 0.8 (4635.22)	4633.38 ± 0.05 (4633.381)	16.9	IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPQQEGNNPQGPPPPAGGNPQQP QAPPAGQPQGPPRPPQGGRPSRPP Q	4898.4 ± 0.8 (4898.35)	4896.44 ± 0.05 (4896.417)	17.9	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
QGPPQQEGNNPQGPPPPAGGNPQ QPQAPPAGQPQGPPRPPQGGRPSR PP	4898.4 ± 0.8 (4898.35)	4896.44 ± 0.05 (4896.417)	17.9	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPQQEGNNPQGPPPPA GGNPQQPQAPPAGQPQGPPRPPQ GGRPSRPP	5462.7 ± 0.9 (5462.01)	5459.73 ± 0.06 (5459.724)	18.1	IB-6, P-H S_1
PPPGKPQGPPPQGGNKPQGPPPPG KPQGPPAQGGSKSQSARAPPGKPQ GPPQQEGNNPQGPPPPAGGNPQQP QAPPAGQ	7611.7 ± 1.3 (7611.42)	7607.75 ± 0.07 (7607.820)	20.5	Ps-1, Ps-2
PQGGNKPQGPPPPGKPQGPPAQG GSKSQSARAPPGKPQGPPQQEGNN PQGPPPPAGGNPQQPQAPPAGQPQ GPPRPPQ	7613.7 ± 1.3 (7613.39)	7609.74 ± 0.07 (7609.811)	20.4	P-Ko, Ps-1, Ps-2,

Identified also in: [a], ref. (Helmerhorst EJ, et al. 2008); [b], ref. (Huq NL, et al. 2007); [c], ref. (Vitorino R, et al. 2010); [d], ref (Vitorino R, et al. 2009); [e], ref. (Messana I, et al. 2008a).

Table 1.10 List of the most abundant naturally occurring fragments of bPRPs eluting before the bPRPs cluster. The peptides characterized for the first time in this study are reported in bold.

Sequence	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Possible origin
PQGPPPQ[a]	719.8 ± 0.2 (719.80)	720.37 ± 0.01 (720.368)	8.1	II-2, IB-1, Gl-1, II-1, CDII-g, Glycosyl. Pr. A
PPPPGKPQ[d]	816.8 ± 0.2 (816.96)	817.46 ± 0.01 (817.466)	4.9	P-E, IB-6, II-2, P-Ko, Ps-1, Ps-2, P-F, P-J, P-D P_{32}, P-D A_{32}, IB-1, IB-8a $Con1^-$, IB-8a $Con1^+$, Glycosyl. Pr. A, II-1, CD-IIg
PPPPGRPQ	844.7 ± 0.2 (844.97)	845.46 ± 0.01 (845.426)	10.9	Gl-3
GPPPPGKPQ[a,d]	874.3 ± 0.2 (874.01)	874.48 ± 0.01 (874.478)	5.5	II-2, P-E, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$, P-D P_{32}, P-D A_{32}
PPPPPGKPQ[d]	914.2 ± 0.2 (914.07)	914.51 ± 0.01 (914.509)	8.5	II-2, P-E, IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$, Glycosyl. Pr. A, II-1, CD-IIg
GPPPPGGNPQ[b]	917.0 ± 0.2 (916.99)	917.45 ± 0.01 (917.448)	7.5	P-D, Gl-3, Gl-2, Gl-1
GGRPSRPPQ	951.1 ± 0.2 (951.05)	951.51 ± 0.01 (951.512)	4.3	P-Ko, IB-6, Ps-1, Ps-2, P-H S_1, P-H A_1,
GPPPPPGKPQ[a,c,d]	971.3 ± 0.2 (971.12)	971.53 ± 0.01 (971.531)	12.0	II-2, P-E, IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPPGKPE	972.0 ± 0.2 (972.11)	972.52 ± 0.01 (972.515)	12.8	Glycosyl. Pr. A, II-1, CD-IIg
GPPPHPGKPQ[b,c]	1011.3 ± 0.2 (1011.15)	1011.38 ± 0.01 (1011.537)	5.8	Gl-1, Gl-2, Gl-3,
GPPPHPGKPE[b]	1012.4 ± 0.2 (1012.13)	1012.52 ± 0.01 (1012.521)	7.3	Gl-1, Gl-2, Glycosyl. Pr. A, II-1, CD-IIg
SPQSPPGKPQ	1022.0 ± 0.2 (1022.13)	1022.53 ± 0.01 (1022.526)	6.5	Ps-1, Ps-2
GPPPRPGKPE	1031.3 ± 0.2 (1031.18)	1031.56 ± 0.01 (1031.563)	8.1	Gl-1, Gl-2, Gl-3
SPRSPPGKPQ	1050.1 ± 0.2 (1050.18)	1050.57 ± 0.01 (1050.569)	4.7	Ps-1, Ps-2
RPPPPPGKPQ[a]	1070.6 ± 0.2 (1070.26)	1070.61 ± 0.01 (1070.611)	9.5	Glycosyl. Pr. A, II-1, CDII-g

Sequence	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Possible origin
GPPPQGGNQPQ[a,b,d]	1076.4 ± 0.2 (1076.13)	1076.51 ± 0.01 (1076.512)	4.6	P-E, IB-6, Ps-1, Ps-2, IB-1, P-J,
GPPPQGGNKPQ[a,b,c]	1076.3 ± 0.2 (1076.18)	1076.55 ± 0.01 (1076.548)	4.7	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
RPAQGQQPPQ	1106.5 ± 0.2 (1106.21)	1106.57 ± 0.01 (1106.570)	5.0	P-D P_{32}, P-D A_{32}
GPPQQGGNRPQ	1135.3 ± 0.2 (1135.20)	1135.56 ± 0.01 (1135.560)	4.5	Ps-1, Ps-2
GPPPQEGNKPQ	1148.0 ± 0.2 (1148.24)	1148.57 ± 0.01 (1148.569)	4.5	Gl-1, Gl-2, Gl-3
GPPQQEGNNPQ[a,b]	1165.5 ± 0.2 (1165.18)	1165.52 ± 0.01 (1165.523)	5.6	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPQQEGNKPQ	1179.5 ± 0.2 (1179.25)	1179.58 ± 0.01 (1179.575)	4.3	P-D P_{32}, P-D A_{32}
SQGTPPPPGKPE[b]	1191.1 ± 0.2 (1191.31)	1191.60 ± 0.01 (1191.600)	13.1	Glycosyl. Pr. A, II-1, CDII-g
GPPPPPQGGRPH[c]	1193.4 ± 0.2 (1193.34)	1193.62 ± 0.01 (1193.617)	9.4	Gl-1, Gl-2, Gl-3
PQGPPPPPGKPQ	1196.5 ± 0.2 (1196.37)	1196.65 ± 0.01 (1196.642)	13.9	II-1, P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$,
GPPRPPQGGRPS	1202.6 ± 0.2 (1202.34)	1202.64 ± 0.01 (1202.639)	13.4	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPQGDKSRSP[a]	1222.6 ± 0.2 (1222.32)	1222.62 ± 0.01 (1222.617)	4.3	II-2, P-E, Ps-1, Ps-2, IB-1
SQGPPPHPGKPE[b]	1227.4 ± 0.2 (1227.34)	1227.61 ± 0.01 (1227.612)	11.9	Gl-1, Gl-2, Gl-3, Glycosyl. Pr. A, II-1, CDII-g
SQGPPPRPGKPE	1246.7 ± 0.2 (1246.39)	1246.65 ± 0.01 (1246.654)	12.3	Gl-1, Gl-2, Gl-3
PPQGGRPSRPPQ[d]	1273.1 ± 0.2 (1273.42)	1273.68 ± 0.01 (1273.676)	11.0	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SHRPPPPPGKPE	1295.6 ± 0.2 (1295.46)	1295.69 ± 0.01 (1295.685)	8.9	Glycosyl. Pr. A, II-1, CDII-g
GGNKPQGPPPPGKPQ	1455.8 ± 0.2 (1455.64)	1455.77 ± 0.01 (1455.770)	12.9	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$

Sequence	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Possible origin
GPPPPGKPQGPPPQGGSKS	1766.9 ± 0.3 (1766.97)	1766.92 ± 0.02 (1766.918)	13.8	P-F, IB-8a Con1^{-}, IB-8a Con1^{+}
SPPGKPQGPPQQEGNKPQ[c]	1870.9 ± 0.3 (1871.04)	1870.94 ± 0.03 (1870.941)	13.8	P-D P_{32}, P-D A_{32}
GPPPQGDKSQSPRSPPGKPQ[b]	2042.1 ± 0.4 (2042.24)	2042.05 ± 0.03 (2042.041)	13.1	Ps-1, Ps-2
GPPPQGDKSRSPQSPPGKPQ	2042.1 ± 0.4 (2042.24)	2042.04 ± 0.03 (2042.041)	13.0	Ps-1, Ps-2

Identified also in: [a], ref. (Messana I, et al. 2008a); [b], ref. (Helmerhorst EJ, et al. 2008); [c], ref (Vitorino R, et al. 2009); [d], ref. (Hug NL, et al. 2007).

1.3.6 Fragments of other salivary proteins that can be confused with anomalous bPRPs.

Several masses were often detected in the chromatographic cluster of bPRPs, and characterized by our group as naturally occurring fragments deriving from other salivary proteins, mainly P-B peptide and aPRPs. These fragments usually detected in human adult saliva are listed in Table 1.11, and comprise 15 fragments never detected in previous investigations, and 6 fragments already characterized in human saliva by other research groups (Vitorino R, et al. 2009; Helmerhorst EJ, et al. 2008; Hardt M, et al. 2005).

Table 1.11 List of the most common peptides or fragments of proteins, which elute in the bPRPs cluster and might be confused with anomalous bPRPs (UniprotKB code is P02814 for P-B fragments, P02810 for P-C and aPRP fragments). The peptides/proteins characterized for the first time in this study are reported in bold.

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Sequence[a]
P-B Fr. 37-45[b,c]	948.1 ± 0.1 (948.14)	948.52 ± 0.01 (948.519)	20.4	IPPPPPAPY
P-B Fr. 24-32[c,d]	960.1 ± 0.2 (960.15)	960.52 ± 0.01 (960.519)	19.7	VPPPPPPPY
P-B Fr. 23-32[d]	1107.2 ± 0.2 (1107.25)	1107.57 ± 0.01 (1107.569)	17.2	FVPPPPPPPY
P-B Fr. 33-45[b,c,d]	1315.6 ± 0.2 (1315.55)	1315.72 ± 0.01 (1315.716)	20.0	GPGRIPPPPP APY
aPRP Fr. 31-44[d]	1436.6 ± 0.2 (1436.56)	1436.72 ± 0.01 (1436.724)	14.7	RQGPPLGGQQ SQPS
P-C Fr. 15-35	2040.3 ± 0.3 (2040.33)	2040.08 ± 0.01 (2040.077)	15.5	GPPPPPPGKP QGPPPQGGRP Q
aPRP Fr. 77-105[c]	2938.3 ± 0.4 (2938.24)	2937.47 ± 0.01 (2937.473)	15.7	GPPQQGGHPP PPQGRPQGPP QQGGHPRPP
aPRP Fr. 67-105	3922.4 ± 0.5 (3922.371)	3921.00 ± 0.02 (3920.992)	16.2	GPPPPQGKPQ GPPQQGGHPP PPQGRPQGPP QQGGHPRPP
aPRP Fr.50-106	5852.4 ± 1.1 (5852.367)	5849.87 ± 0.03 (5849.875)	16.5	D DGPQQGPPQQ GGQQQQGPPP QGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr.44-105	6238.7 ± 1.2 (6239.69)	6236.97 ± 0.03 (6236.967)	16.3	AGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr.29-93	6580.0 ± 1.2 (6580.95)	6577.12 ± 0.03 (6577.126)	17.2	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQ
aPRP Fr.40-105	6638.1 ± 1.2 (6638.09)	6635.16 ± 0.03 (6636.158)	16.7	S QPSAGDGNQD DGPQQGPPQQ GQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr.31-105	7501.0 ± 1.4 (7501.04)	7498.60 ± 0.04 (7498.572)	17.4	QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr.29-105	7786.3 ± 1.6 (7786.35)	7782.73 ± 0.04 (7782.732)	18.2	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr.18-93	7853.1 ± 1.6 (7853.14)	7849.55 ± 0.04 (7849.545)	19.8	DGG D**S**EQFIDEER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQ

Name	Exp Mav (theor.)	Exp $[M+H]^{1+}$ (theor.)	Elution time (min± 0.4)	Sequence[a]
aPRP Fr.29-106	7942.9 ± 1.6 (7943.53)	7938.8 ± 0.04 (7938.833)	16.9	ER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr.26-106	8300.9 ± 1.7 (8300.88)	8296.96 ± 0.04 (8297.011)	17.5	IDEER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr.18-105	9060.2 ± 1.8 (9061.47)	9055.20 ± 0.05 (9056.134)	19.6	DGG D<u>S</u>EQFIDEER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr.18-106	9216.3 ± 1.8 (9216.67)	9211.30 ± 0.05 (9211.251)	19.1	DGG D<u>**S**</u>EQFIDEER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr. 37-150	11460.576 (11461.397)	11454.56 ± 0.06 (11454.563)	19.8	GQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPRGRPQ GPPQQGGHQQ GPPPPPPGKP QGPPPQGGRP QGPPQGQSP
aPRP Fr. 29-150	12296.0 ± 2 (12296.33)	12289.03 ± 0.06 (12288.998)	19.0	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPRGRPQ GPPQQGGHQQ GPPPPPPGKP QGPPPQGGRP QGPPQGQSPQ

[a], <u>S</u>: phosphorylated Ser; identified also in: [b],ref. (Hug NL, et al. 2007); [c], ref. (Vitorino R, et al. 2009); [d], ref. (Tagliabracci VS, et al. 2012)

1.4 Discussion

The top-down approach applied to the proteomic characterization of human saliva allowed to highlight the great heterogeneity of bPRP family, which on the basis of current results includes 55 new components, detected for the first time in this study, bringing the total number of bPRPs to 110. The heterogeneity of the parent bPRPs is really amazing, but the great similarity among some of them, evident by looking at the sequences reported in Tables 1.1, 1.3, 1.5, 1.6, suggested the division of the bPRPs in two main groups and a third minor hybrid group (Figure 1.9).

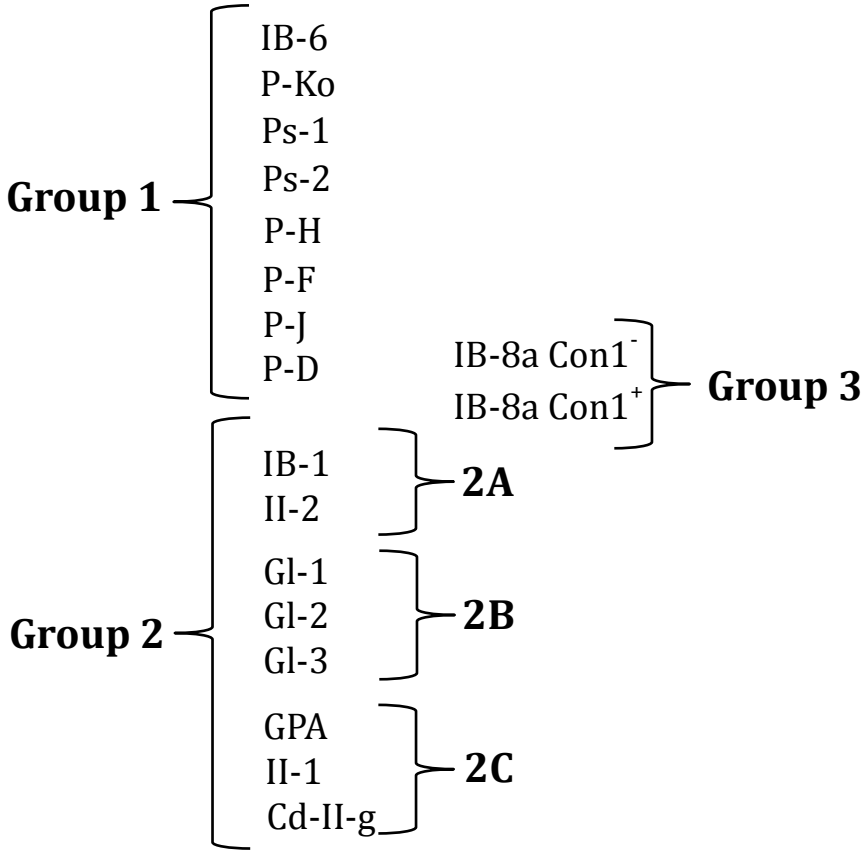


Figure 1.9 Schematic classification of the parent bPRPs, performed on the basis of their sequence similarity.

The first group, named Group 1, includes P-E, P-Ko, IB-6, Ps-1, Ps-2, P-H, P-F, P-J, and P-D. The sequence of all these bPRPs starts with the same SPPGKPQGPP motif followed by sequences somewhat similar, but showing small variations among the different components. The central part of the sequences shows similar repeats. P-

E, IB-6, Ps-1 and Ps-2 sequences originate from DNA length polymorphisms in exon 3 of *PRB1* locus, thus they exhibit high similarity (Azen E.A, et al. 1993; Azen E.A, et al. 1996; Stubbs M, et al. 1998). While PRB1-S proprotein contains two convertase cleavage sites that generate II-2 (first cleavage) and P-E and IB-6 (second cleavage) (Fig. 1.5), PRB1-M and L proproteins, due to the substitution Arg_{131}→Gln that abolishes the second cleavage site, undergo only one convertase cleavage, that generates II-2 together with Ps-1, and Ps-2, respectively, as already suggested by Azen and co-workers (Azen E.A, et al. 1993). The bPRP with a Mav of 10433.5 Da, detected in whole saliva and in parotid secretory granules (Messana I, et al. 2004; Messana I, et al. 2008a) and named P-Ko by Halgand et al. (Halgand F, et al. 2012), is encoded by *cP4*, a differentially spliced transcript of *PRB1L* allele (Maeda N, et al. 1985). cP4 pro-protein lacks the sequence 106-299 of PRB1-L (P04280), and its cleavage generates II-2 peptide, and P-Ko protein (Fig. 1.5).

Group 2 includes IB-1, II-2 and the glycosylated bPRPs codified by *PRB3* and *PRB4* genes, namely Gl-1, Gl-2, Gl-3, GPA, II-1 and Cd-IIg. Their sequences start with the similar motif (E/Q)XXXEDVSQEES, where XXX is LNE in IB-1, II-2, Gl-1, Gl-2 and Gl-3, and SSS in GPA, II-1 and Cd-IIg. The central part of the sequences comprise similar repeats with differences from the repeats of the members belonging to Group 1. The N-terminal glutamine of IB-1 and II-2 is converted to a pyro-glutamic acid moiety and the serine at 8 position is phosphorylated for the presence of the SXE consensus sequence recognized by the Golgi casein kinase Fam20C (Tagliabracci VS, et al. 2012), responsible for the phosphorylation of all the salivary peptides (aPRPs, histatin 1, statherin and cystatin S). In a previous work we demonstrated that, in a resemblance with IB-1 and II-2, the N-terminal glutamine of Gl-2 is converted to a pyro-glutamic acid moiety and that serine at position 8 is phosphorylated (Manconi B, et al. 2016a). Phosphorylation is an almost complete event, since less than 1% of the non-phosphorylated forms can be detected in parotid granules, parotid and whole saliva, and probably occurs after the cleavage of the pro-protein (Messana I, et al. 2008a). It can be supposed, by sequence similarity, that also Gl-1 and Gl-3 undergo the same post-translational modifications (PTMs), reported in Table 1.5 as hypothetical. The presence of a glutamic acid residue at the N-terminus of GPA, II-1 and Cd-IIg and the SQE consensus sequence (for Serine-8) suggests similar PTMs for these bPRPs too, namely the N-terminal pyro-E and phosphorylation of S_8. These PTMs are reported as hypothetical in Table 1.6. A

second potential phosphorylation site at S_3 is present in the sequence of GPA, II-1 and Cd-IIg, but, due to the absence of experimental evidence of this modification in this study and in literature, the phosphorylation of S_3 is not reported in Table 1.8. All the glycosylated proteins of Group 2, after the initial sequence similar to IB-1 and II-2, contain a variable number of similar repeats characterized by the presence of the N-glycosylation consensus sequence NQS. Moreover, all these glycosylated proteins show potential O-glycosylation sites. On the basis of structural differences, members of Group 2 can be divided in three subgroups: Group 2A, including IB-1 and II-2, without glycosylation sequons, Group 2B, including the Gl proteins codified by the alleles of *PRB3* locus, and Group 2C including the glycosylated proteins codified by the alleles of *PRB4* locus. Differently from the other *bPRP* loci, the proproteins expressed by the *PRB3* locus are not submitted to a proteolytic cleavage before secretion. Gl proteins can be found at least in nine size variants in different populations (Azen E.A, et al. 1979; Minaguchi K, et al. 1981; Lyons, KM, et al. 1988b; Azen E.A, et al. 1990). In black and white populations the four allelic size variants *S, M, L* and *VL* encode for the corresponding Gl protein size variants Gl-4/*PRB3-VL* > Gl-1/*PRB3-L* > Gl-2/*PRB3-M* > Gl-3/*PRB3-S* (Lyons, K.M, et al. 1988). The Gl-8 glycoprotein derives from a single nucleotide insertion in the *PRB3-S*^{Cys} allele, which converts R_{15} to C. Gl-8 protein is electrophoretically distinct from the other Gl protein variants because it forms a disulfide-bond heterodimer under the action of the salivary peroxidase (Azen E.A, et al. 1990). In Table 1.7 only the three most common variants described in the Caucasian population are reported. The small Group 3 is a hybrid group, which includes the two proteoforms of IB-8a, Con1$^-$ and Con1$^+$. The initial sequence of these two proteins resembles that of Group 1, while the terminal sequence is similar to the repeat responsible for the glycosylation of the bPRPs of Groups 2B and 2C.

We never detected a putative PRB2-like Con2$^+$ protein, neither in the non-glycosylated nor in the glycosylated form. Indeed, it was reported that this protein, 60 residue long and encoded by a hybrid *PRB1-M CON2*$^+$ allele, had a single potential N-glycosylation site (Azen E.A, et al. 1996).

We were able to characterize by MSMS the structure of some variants of bPRPs. Particularly, the characterization of P-H $S_1 \rightarrow A$ variant, previously detected by Kaufmann and colleagues (Robinson R, et al. 1989) and attributed to the fragment 337-392 of the *PRB1-S* allele (corresponding to the fragment 63-118 of IB-6),

resulted not correct from our data for two reasons: a) IB-6 has a serine residue at position 63 instead of the alanine reported by Kaufmann; b) in saliva of 9 subjects (out of 86) carrying this variant we never detected the complementary IB-6 1-62 fragment. Moreover, we characterized two variants of P-Ko: the P_{36}→S variant identified for the first time in this study, detected in only one subject, and the A_{41}→S variant, detected in 11 out of 86 subjects, which corresponded to the fragment 92-198 of the sequence deposited at the UniprotKB human data bank with the code G5E9X6. This sequence, obtained from a large scale genomic DNA investigation, is attributed to a polymorphism of *PRB1* locus that encodes for a proprotein with a single convertase cleavage site from which II-2 and the P-Ko A_{41}→S variant are generated. The parent bPRPs reported in Tables 1.1, 1.3, 1.5 and 1.6 were submitted to naturally occurring fragmentations and the peptide products were shown in Tables 1.4, 1.6, 1.9 and 1.10. The fragmentations observed on bPRPs can be divided in two types, those occurring before secretion and those occurring after secretion (Messana I, et al. 2008a). The first type commonly occurs at the C-terminal residues and it is a widespread event observed in many secretory processes ascribed to specific carboxy-exopeptidases acting after the convertase cleavage. The post-secretory cleavage is mainly due to exogenous proteinases deriving from the oral microbiota and generates numerous small fragments recurrently found in whole saliva. Because of the great sequence similarities of bPRPs, it is impossible to establish the parent protein of the fragments reported in Table 1.9 and 10. Many of these peptides terminate with a KPQ sequence, and this finding allowed to the research group of Oppenheim F. to characterize a glutamine endoproteinase from Rothia species bacteria as the responsible for this cleavage (Zamakhchari M, et al. 2011).

Twenty-one peptides/proteins eluting in the chromatographic range of the bPRP cluster were identified as fragments of other salivary proteins (Table 1.11). Indeed, almost all the human secreted salivary proteins are submitted to proteolysis by various proteinases acting before, during and after glandular secretion (Messana I, et al. 2008a; Castagnola M, et al. 2012b). The fragments shown in Table 1.11 derived mainly from aPRPs, P-C and P-B salivary peptides. It is important to remind that P-C and P-B peptides were sometimes ascribed to the bPRP family. However, P-C is a peptide of 44 amino acid residues resulting from the cleavage of PRP-1, PRP-2, Pif-s and Db-s proteoforms of aPRPs, and therefore it must be considered a member of the aPRPs family. P-B peptide is the product of *PROL3* gene (PBI;

http://www.ensembl.org/Homo_sapiens/ ENSG00000171201) localized on chromosome 4q13.3, very close to the statherin gene. It shows high sequence homology with statherin, and, as statherin, displays some tyrosine residues in its sequence (completely absent in bPRPs family). As statherin, it is secreted both from parotid and submandibular/sublingual glands, and it does not derive from the cleavage of a bigger pro-protein. For all these reasons, it has to be considered a member of the statherin family.

Polymorphisms and PTMs of bPRPs generate a high number of proteins/peptides with rather similar structures. Being the naturally occurring proteolytic cleavage the most represented event, top-down proteomics represents a key tool for the characterization of the bPRP complexity. The meaning of this amazing complexity is still largely obscure. Salivary proline-rich proteins are highly conserved in mammalian saliva, although significant structural differences are present in different animals, suggesting they play a crucial role in the oral protection. Some bPRPs exhibit ability to bind harmful tannins (Carlson D.M, et al. 1988), other to modulate the oral flora (Ruhl S, et al. 2004), some others are involved in bitter taste perception (Cabras T, et al. 2012b). Some bPRP fragments are involved in the enamel pellicle formation (Vitorino R, et al. 2007) and others act as antagonists of the progesterone induced cytosolic Ca^{2+} mobilization (Palmerini CA, et al. 2016). The intrinsic propensity of some fragments to adopt a polyproline-II helix arrangement joined to PxxP motifs was suggestive for the interaction with the SH3 domain family (Macias M.J, et al. 2002). Interestingly, interaction were highlighted (Palmerini CA, et al. 2016) with Fyn, Hck, and c-Src SH3 domains, which are included in the Src kinases family, suggesting that some basic bPRPs can be involved in the signal transduction pathways modulated by these kinases. In human, bPRPs are secreted only by parotid glands, and this regio-selectivity is puzzling. Moreover, their expression appears to be related to the growth with different trends among the several bPRPs (Cabras T, et al. 2009).

1.5 Conclusions of section 1

Although various aspects of bPRPs have still to be defined, the survey described in this thesis and recently published (Padiglia A, et al. 2018) may be considered an updated reference for the peptides included in this family. The increased information obtained on human salivary bPRPs might facilitate future studies devoted to establish the specific biological roles of the different components of this complex family of proteins.

SECTION 2:

Mass spectrometry mapping of transglutaminase 2 active sites of several human salivary small basic proline-rich proteins, P-C peptide and statherin.

2.1 Introduction

Transglutaminases (TGs) are monomeric globular proteins (Hannig C, et al. 2005) generating a cross-link between two peptide chains, typically between a lysine residue (which acts as a lone-pair donor) and a glutamine residue (which acts as acceptor) (Ahvazi B, et al. 2003). There are nine different genes for TGs in humans (Table 2.1) (Eckert RL, et al. 2014). TG-2 is the enzyme of major interest for this thesis because it is the principal TG active in oral cavity (together with TG-1 and TG-3, in minor amounts).

Table 2.1 Properties of transglutaminases

Gene	Protein	Chromosomal Location	Molecular Mass, kDa	Main Function	Tissue Distribution	Alternate Names
TGM1	TG1	14q11.2	90	Cell envelope formation during keratinocyte differentiation	Membrane-bound keratinocytes	TG_k, keratinocyte TG, particulate TG
TGM2	TG2	20q11-12	80	Apoptosis, cell adhesion, matrix stabilization, signal transduction	Many tissues: cytosolic, nuclear, membrane, and extracellular	Tissue TG, TG_c, liver TG, endothelial TG, erythrocyte TG, Ghα
TGM3	TG3	20q11-12	77	Cell envelope formation during keratinocyte differentiation	Hair follicle, epidermis, brain	TG_E, callus TG, hair follicle TG, bovine snout TG
TGM4	TG4	3q21-22	77	Reproduction, especially in rodents as a result of semen coagulation	Prostate	TG_p androgen-regulated major secretory protein, vesiculase, dorsal prostate protein 1
TGM5	TG5	15q15.2	81	Cell envelope formation in keratinocytes	Foreskin keratinocytes, epithelial barrier lining, skeletal muscular striatum	TG_x
TGM6	TG6	20q11	78	Not known	Testis and lung	TG_y
TGM7	TG7	15q15.2	81	Not known	Ubiquitous but predominately in testis and lung	TG_z
F13A1	FXIIIa	6q24-25	83	Blood clotting, wound healing, bone synthesis	Platelets, placenta, synovial fluid, chondrocytes, astrocytes, macrophages, osteoclasts and osteoblasts	Fibrin-stabilizing factor, fibrinoligase, plasma TG, Laki-Lorand factor
EPB42	Band4.2	15q15.2	72	Membrane integrity, cell attachment, signal transduction	Erythrocyte membranes, cone marrow, spleen	B4.2, ATP-binding erythrocyte membrane protein band 4.2

TG-2 is released by the epithelial cells, and plays a principal role in the formation of the “oral protein pellicle” covering the oral epithelia (Esposito C, et al. 2005; Wang Z, et al. 2012). The in vivo pellicle is thought to be an insoluble network of proteins generated by transglutaminase cross-linking. As mentioned above, TGs generate a cross-link between two peptide chains typically between a lysine residue and a glutamine residue. The reaction is accomplished by the loss of an ammonia molecule (Figure 2.1). Cross-links in oral cavity were demonstrated at first by

Bradway et al. for the formation of oral mucosal pellicle, a network of proteins produced by components of saliva adsorbed onto buccal epithelial cells that cover the oral mucosal surface (Bradway S D, et al. 1989; Bradway S D, et al. 1992). This protein molecular network could interact with the oral epithelial-cell plasma membrane and its associate cytoskeleton and might contribute to the mucosal epithelial flexibility and turnover. It was demonstrated that acidic-proline-rich proteins, statherin, and the major histatins are substrates of oral transglutaminase 2 and they participate in cross-linking reactions (Yao Y, et al. 1999; Yao Y et al. 2000; Cabras T, et al. 2006) as putative pellicle precursor proteins. It has been already known that TG-2 can generate cyclo-statherin *in vitro* and *in vivo* (Cabras T, et al. 2006), involving the unique Lys-6 residue and almost specifically Gln-37 (~95%), with the percentage of Gln-39 implicated in the cross-linking being less than 5%.

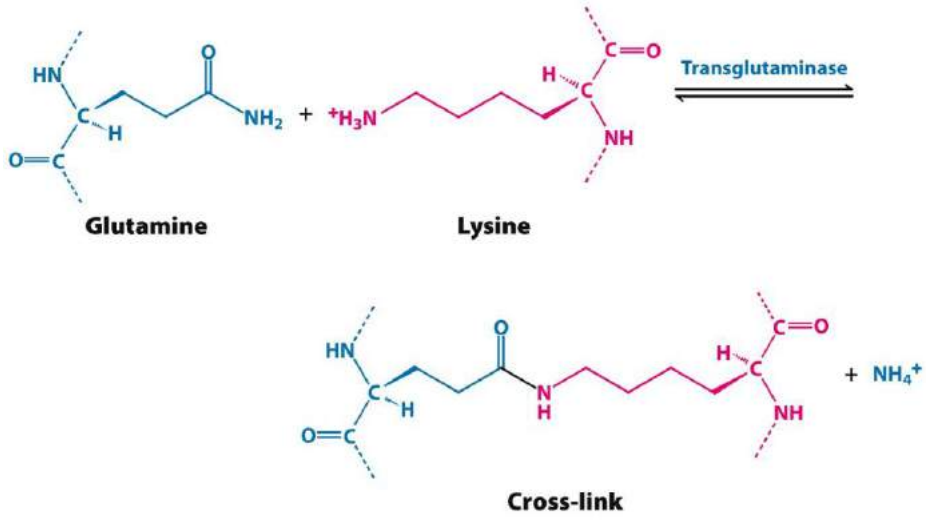


Figure 2.1 Transglutaminase generates a cross-link between two peptide chains typically between a Lys residue and a Gln residue

Mechanism of the reaction catalyzed by TG-2 is represented in Fig. 2.2. Typically, TGs use a cysteine protease-like catalytic mechanism to release ammonia from protein-bound glutamine. In the presence of Ca2+, the active-site cysteine residue of TG-2 reacts with the γ-carboxamide group of an incoming glutaminyl residue of a protein/peptide substrate to yield a thioacyl-enzyme intermediate and ammonia (Figure 2.2 **Step 1).** The thioacyl-enzyme intermediate reacts with a nucleophilic primary amine substrate, resulting in the covalent attachment of the amine-containing donor to the substrate glutaminyl acceptor and regeneration of the cysteinyl residue at the active site. If the primary amine is donated by the ε-amino group of a lysyl residue in a protein/polypeptide, a Nε-(γ-L-glutamyl)-L-lysine (GGEL) isopeptide bond is formed (Figure 2.2 **Step 2)** (Ahvazi B, et al. 2003; Gatta NG, et al. 2016; Gatta NG, et al. 2017).

Subsequently, TGs transfer the γ-glutamyl moiety to:

(i) another amine (transamidation),

(ii) an aliphatic alcohol (esterification), or

(iii) water (glutamine hydrolysis; deamidation), like shown in Fig. 2.2.

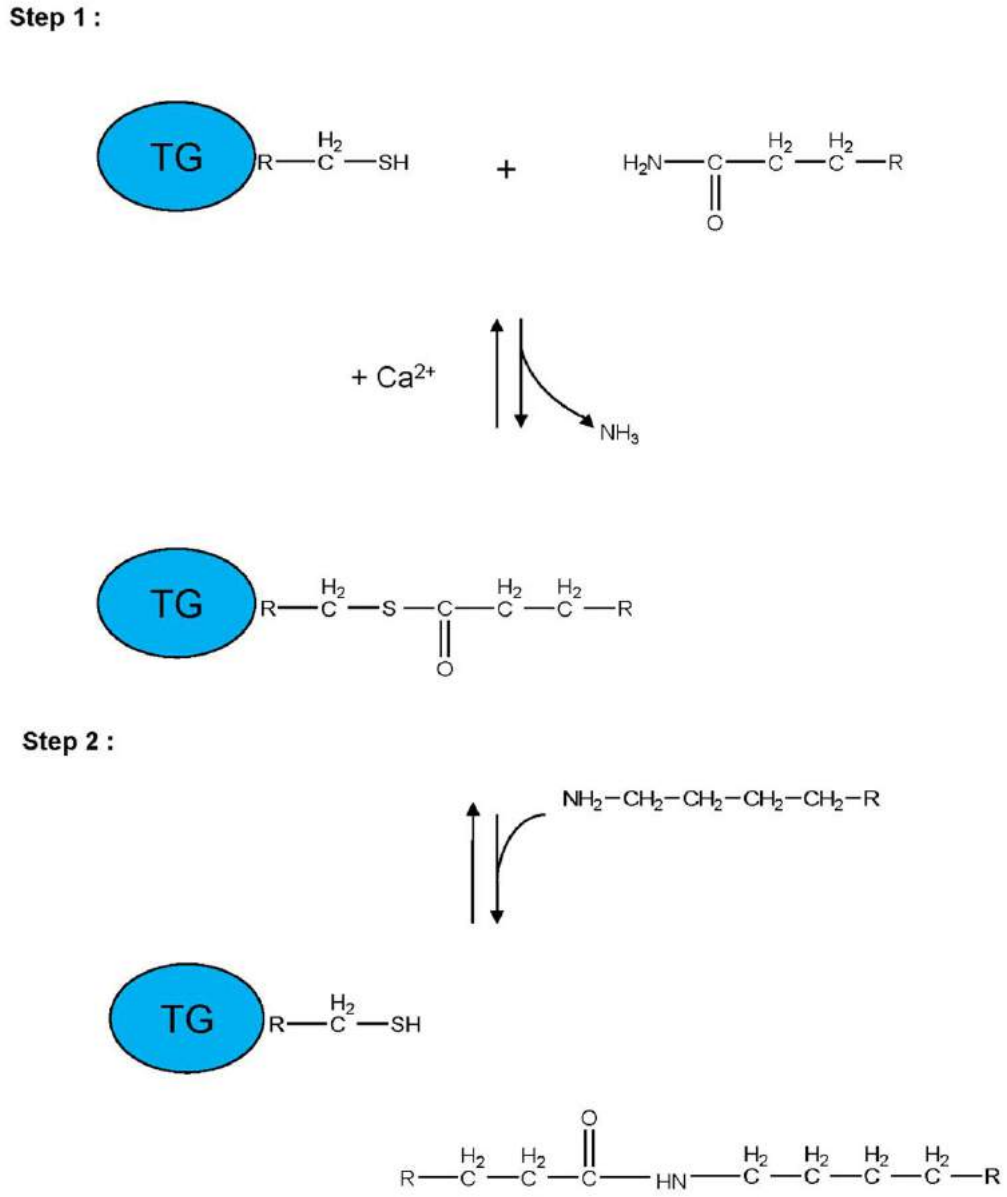


Figure 2.2 Schematic representation of a two-step transglutaminase reaction.

Transglutaminases catalyze various post-translational reactions. Transamidation can cause protein crosslinking by forming a Nε(γ-glutamyl) lysine isopeptide bridge between the deprotonated lysine (Lys) donor residue of one protein (purple ellipse, Figure 2.3) and the acceptor glutamine (Gln) residue of another (blue rectangle) (Figure 2.3, **a**). In addition it can make the incorporation of an amine (H_2NR) into the Gln residue of the acceptor protein (diamines and polyamines might act as a tether in a bis-glutaminyl adduct between two acceptor molecules) (Figure 2.3, **b**) and the acylation of a Lys side chain of the donor protein (Figure 2.3, **c**). Reactions **b** and **c** compete against the crosslinking that is shown in **a**. Transglutaminases (TGs) react only with the γ-amides of select endo-Gln residues in some proteins and peptides. TGs show specificities both for their Gln and Lys substrates. Transamidations proceed probably with little change in free energy; in the absence of phase separation (clotting, precipitation), the reactions might be reversible. The same applies to esterification (Figure 2.3, **d**), but not to deamidation (Figure 2.3, **e**), and isopeptide cleavage (Figure 2.3, **f**). Electron movements (curved arrows) are shown for the nucleophilic displacement reactions in the absence of enzyme. In the presence of TGs, however, the pathway of catalysis is more complicated and, as with papain, involves the formation of a thiolester acylenzyme intermediate (Lorand L and Graham RM. 2003).

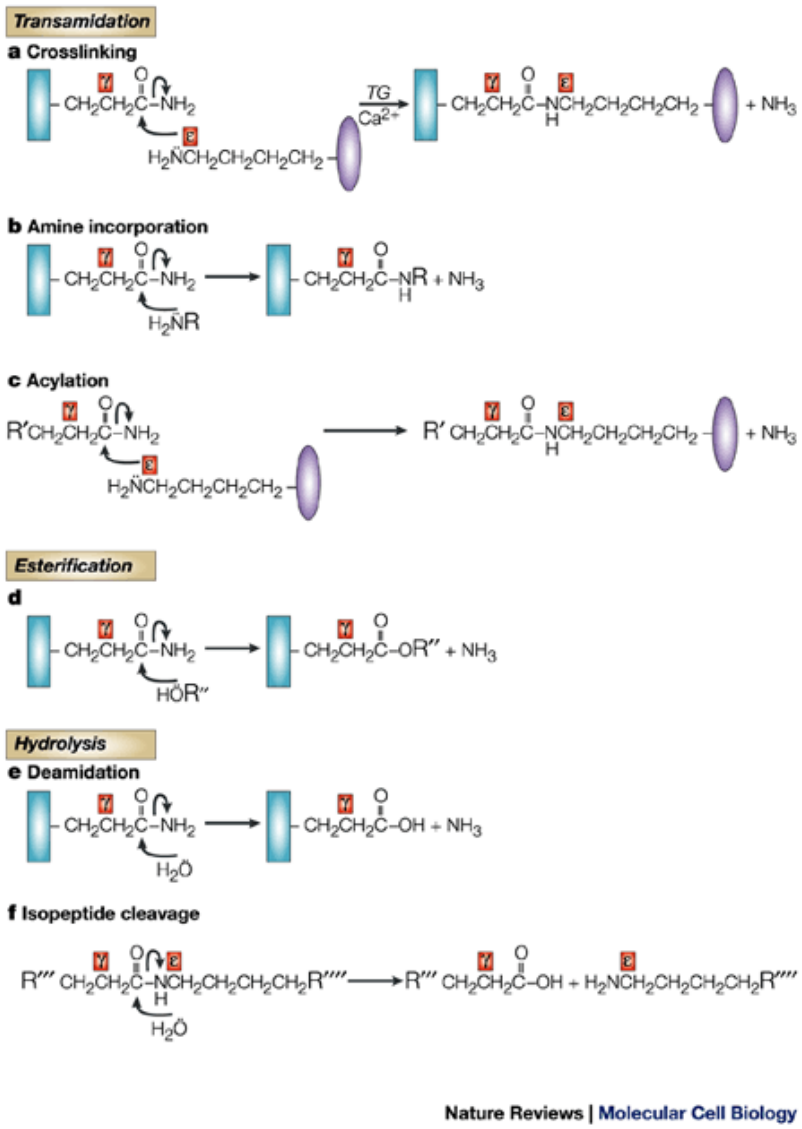


Figure 2.3 Transglutaminases catalyze various post-translational reactions. R represents the side chain in a primary amine; R′, a Gln-containing peptide; R″, a ceramide; R‴ and R⁗, the side chains in branched isopeptides (Lorand L and Graham RM. 2003).

In vitro, many amines, diamines, polyamines, and alcohols are capable of interaction with the protein- γ-glutamyl-enzyme intermediate. However, *in vivo*, only lysine ε-amino groups and polyamines are abundantly available amine substrates. The transfer of protein-γ-glutamyl residue to these amines yields γ-glutamyl-ε-lysine (GGEL), or γ-glutamyl-polyamines, respectively (Pastor MT, et al. 1999; Lorand L, et al. 2003; Nemes Z, et al. 2005).

The tissue transglutaminase 2 (TG-2) is an enzyme requiring Ca^{2+} ions. Multiple Ca^{2+} can bind to a single TG-2 molecule. In contrast, the binding of one molecule of GTP or GDP inhibits the crosslinking activity of the enzyme (Jin X, et al. 2011; Clouthier CM, et al. 2012; Klöck C, et al. 2012; Keillor JW, et al. 2015; Akbar A, et al. 2017)

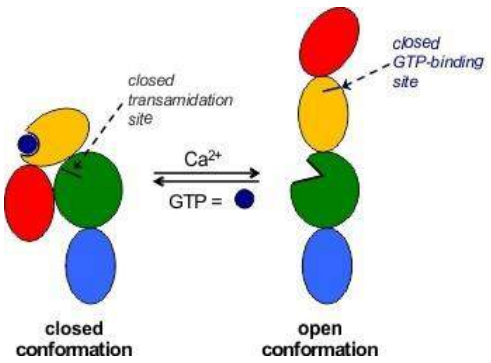


Figure 2.4 Allosteric model of irreversible inhibition of both transamidation and GTP-binding activity. Ca^{2+} and guanine nucleotide binding inversely regulate the transamidating activity of TG-2. GTP bound TG-2 has a closed conformation and it is catalytically inactive. Binding of Ca^{2+} is essential to acquire a catalytically active 'open' or 'extended' conformation (Keillor JW, et al. 2015; Akbar A, et al. 2017).

TG-2 is a relatively non-specific crosslinking enzyme, and its activity in and outside the cell is also regulated by the redox potential. Binding of Ca^{2+} (dissociation constant of approximately 60 μM) is essential for TG-2 to acquire a catalytically active 'open' or 'extended' conformation. In contrast, binding of GTP/GDP (dissociation constant of approximately 1.6 μM) renders TG-2 in a catalytically inactive 'closed' or 'compact' conformation. Under physiological conditions, high levels of GTP, low redox potential, and low free Ca^{2+} level keep TG-2 in its catalytically inactive compact state. However, a calcium ion influx due to extreme stress or cell damage can induce the catalytically active or 'extended' conformation. In comparison with the intracellular environment, the extracellular matrix (ECM) has a considerably lower GTP level and relatively high Ca^{2+} level. Therefore, the newly secreted TG-2 can be expected to be in a catalytically active state (Jin X, et al. 2011; Diraimondo TR, et al. 2012; Klöck C, et al. 2012; Agnihotri N, et al. 2013; Huelsz-Prince G, et al. 2013). However, a large fraction of the extracellular TG-2 in most organs is in an inactive form because of disulfide bonding, between two surface cysteine residues, C370 and C371. In the compact or catalytically inactive state, TG-2 can act as a scaffold protein and result in the activation of various signaling pathways. In its extended and catalytically active state, TG-2 catalyzes highly stable protein crosslinking, resulting in apoptotic death if inside the cell or stabilization of the matrix if outside the cell (Figure 2.5) (Pinkas DM, et al. 2010; Agnihotri N, et al. 2013; Huelsz-Prince G, et al. 2013).

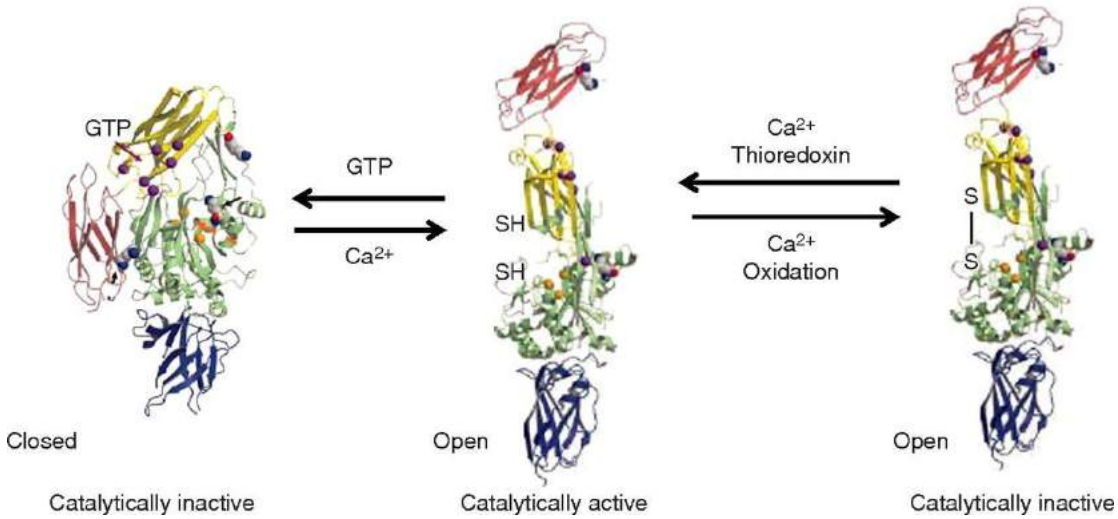


Figure 2.5 Allosteric regulation of tissue transglutaminase (TG-2) activity and functions.

As mentioned above, TG-2 transamidation reactions require Ca^{2+}, in both *in vitro* and *in vivo* assays (Ahvazi B, et al. 2003). Due to the drastic and potentially disruptive effects of uncontrolled protein cross-linking in living cells, the enzymatic activity of vertebrate TGs is tightly controlled also by the availability of calcium ions, which are essential cofactors for the attainment and maintenance of catalytically active conformation states (Nemes Z, et al. 2005). In the few cases measured, the Ca^{2+} concentration required to activate an enzyme isoform (>500 μM) is far higher than net intracellular Ca^{2+} ion concentrations (about 100 nM). Thus manipulation of intracellular Ca^{2+} concentrations could afford an effective way to control TG functions, including cross-linking (Ahvazi B, et al. 2003).

The consensus sequences recognized by TG-2 are not well known (Esposito and Caputo 2005), but it has been reported that the enzyme is much less selective towards the lysine donor than towards acceptor glutamine residue. Indeed, while the recognition of specific lysine residues seems to be governed only by their steric hindrance, the spacing and structure of neighboring residues seems to be a crucial factor for the TG-2 specificity towards targeted glutamine residues. In particular, proline residues seem to be relevant for glutamine recognition: a glutamine residue is not recognized as a substrate if it occurs between tow proline residues (Pastor et al. 1999). Moreover, while the enzyme is able to recognize **Q**XP residues, a +1 or +3 flanking proline residue seems to completely abolish TG-2 recognition (Piper et al. 2002) and two adjacent glutamine residues may act as amine acceptors in a consecutive reaction (Parameswaran et al. 1990; Esposito and Caputo, 2005).

The advent of automated proteome analysis has generated increasing demand for the analysis of post-translational modifications. However, unlike phosphate, lipid, or sugar attachments, a covalent cross-linking of proteins forms branching in the sequence of involved proteins and thus renders an extra dimension to the complexity of such structures. Identification of GGEL (γ-glutamyl-ε-lysine) cross-links in biological samples is therefore difficult and of low throughput. Nevertheless, the demonstration, quantitation, and sequence localization of the cross-links is indispensable for postulating, determining, or characterizing their biological importance. The broadening availability, improving performance, and simplified operation of mass spectrometric techniques should overcome methodical drawbacks which have hitherto compromised sensitivity and specificity of cross-link

identification in proteins. Future technical development may also provide new potent methods for the qualitative analysis of glutamine deamidation and poly-amination, alternative to cross-linking (Nemes Z, et al. 2005).

The analysis of TG products is beset with numerous methodical problems, the most important of which are the small differences in mass and physico-chemical properties associated with poly-amination and deamidation and, at the other extreme, the large size and poor solubility of GGEL cross-linked protein aggregates (Nemes Z, et al. 2005).

Glandular secretions do not contain TG. In the oral cavity, this enzyme derives from epithelial cells and from crevicular fluids (Bradway SD, et al. 1992; Hannig C, et al. 2005). TG activity is present on the surfaces of oral epithelial cells (Bradway SD, et al. 1989; Bradway SD, et al. 1992). TG-2 is responsible for the formation of the epithelial cell envelope of mucosal cells by the cross-linking of salivary components to each other or to the epithelial cytoskeleton (Bradway SD, et al. 1989). Transglutaminase acting on these peptides generates a network of proteins, covalently linked, covering and protecting the oral epithelia that are different from the mucosal surface of any other human mucosa. The ability of TG-2 to cross-link salivary proteins is evidenced by the weakness of the oral epithelial surface in patients with Sjögren syndrome, a common rheumatic disease, which is characterized by low or absent secretion of salivary and lachrymal glands (Mathews SA, et al. 2008).

2.1.1 Aims of this study

As mentioned in the introduction, the aim of section 2 of this thesis was to verify if some bPRPs are substrates of TG-2 and if they are therefore potential candidates in the formation of the oral mucosal protein pellicle. As shown in Section 1 human bPRPs comprise more than 11 parent peptides/proteins with similar sequence secreted only by parotid glands. Because some of them (as Ps-1 and Ps-2) are too difficult to separate and to study as substrates for TG-2 for their dimensions, in this thesis only the properties of P-H, P-D, II-2, P-F and P-J were studied. Also P-C peptide was investigated as a potential substrate of TG-2. Moreover, in order to verify if the experimental conditions applied were comparable to that used previously for the statherin, the action of TG-2 on purified statherin was also investigated.

2.2 Materials and Methods

2.2.1 Reagents.

Chemicals and reagents, all of LC−MS grade, were purchased from Merck/Sigma-Aldrich (Darmstadt, Germany), Waters Corporation (Milford, MA), ThermoFischer Scientific (Rockford, IL).

2.2.2 Salivary Sample Collection

Whole saliva was collected, according to a standardized protocol optimized to preserve saliva proteins from proteolytic degradation, from normal adult volunteers between 2:00 and 4:00 p.m. when the secretion of the parotid gland is at a maximum, by a soft plastic aspirator. The samples were immediately added to aqueous TFA (0.5%), 80:20, in an ice bath. After centrifugation, at 10000g for 10 min at 4°C, the acidic supernatant was diluted 1:1 with 0.5 mM zinc chloride, the solution was brought up to pH 9.0 by adding 0.1 M NaOH, stored at ice for 20 min in order to precipitate statherins and histatins. After second centrifugation, at 10000g for 10 min at 4°C, the supernatant was lyophilized; and the precipitate, which containing mainly histatins and statherin, was dissolved in 5% FA.

2.2.3 Peptide Purification

The freeze-dried sample was dissolved in water and submitted to purification by gel filtration chromatography on a Sephadex-G75 column (2cm×80cm). Absorbance of the fractions was measured by spectrophotometer at 214 nm and 278 nm.

The fractions corresponding to each peak were unified and submitted to a second purification by RP-HPLC on a preparative C8 column (Vydac Revers Phase C8, 5 μm particle, diameter 250 × 10 mm. Moreover, the dissolved precipitate was submitted to a purification by RP-HPLC on a semi-preparative C8 column in order to separate statherin and histatins.

The concentration of each purified peptide (Statherin, P-C, P-H, P-D, II-2, P-F and P-J) was measured by Bicinchoinic Acid (BCA) assay.

2.2.4 TG-2 reactions

50 µl of each purified peptide at a concentration ≈ 1.4 µg/ml was mixed by 38.75 µl of buffer (160 mM Tris-HCl, 2 mM DTT, 2 mM EDTA buffer, pH 7.55). Then 30 µl of 0.55µM (0.01 EC unit/ml) guinea pig TG-2 (purchased from Zedira GmbH, Germany) was added to this mixture. Finally, after the addition of 6.25µl of 20 mM $CaCl_2$, the reaction solution (final volume 125µl) was ready to incubate at 37°C.

20µl of the reaction solution was picked up after 5 minutes and reaction was stopped by addition 3.2 µl of 0.2 M EDTA (final concentration 33 mM). The sample was centrifuged, at 10000g for 10 min at 4°C, and the supernatant was immediately analyzed by high resolution mass spectrometry or stored at -80°C.

This step was repeated after 1, 2, 3 and 4 hours.

Three kind of experiments were performed:

1) reaction with TG-2 and the peptide alone

2) DC reaction: reaction with peptide and 2 mM monodansylcadaverin (DC, purchased from Sigma Aldrich, Switzerland) devoted to label reactive Gln residues in salivary peptides (Fig. 2.6a).

3) BQG reaction: reaction with peptide and 1mM benzoyl-glutamine-glycine (BQG; γ-glutamyl donor substrate purchased from Zedira GmbH, Germany) devoted to label reactive Lys residues in salivary peptides (Fig. 2.6b).

a **b**

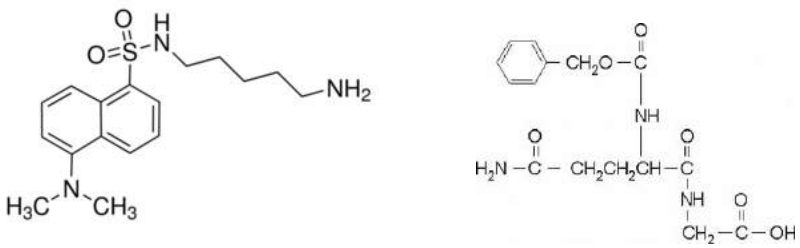


Figure 2.6 **a**) dansylcadaverine **b**) benzoyl-glutamine-glycine

Some reactions (1-3 types on statherin, P-C and II-2 peptides) were also performed at 25° C and 45° C, in order to investigate how temperature influences the enzyme activity and in order to see if bPRPs react with BQG in these conditions.

Some reactions (1-3 types on statherin, P-C and II-2 peptides) were also performed using human TG-2 (purchased from Zedira GmbH, Germany) under the

same conditions used for experiments with guinea pig TG-2, in order to see if the results were comparable with those obtained using guinea pig TG-2.

2.2.5 HPLC Low- and High-Resolution ESI−IT−MS Experiments

The conditions used for Low- and High-Resolution HPLC-ESI-MS and MSMS experiments were the same reported in the first part of the thesis in Sections **1.2.4** and **1.2.5.**

2.3 Results

2.3.1 Purification of peptides

Peptides utilized in this study were purified from resting whole human saliva (usually 80 mL) in various steps (see material and methods). The first step consisted in a precipitation with zinc chloride (Gusman H, et la. 2004). The precipitate contained mainly histatins and statherins, while the supernatant contained mainly bPRPs, aPRPs and amylase. The supernatant was submitted to a gel filtration (Sephadex G75, Figure 2.7). In the gel filtration profile, the peaks without absorbance at 278 nm were characteristic for bPRPs and aPRPs, which completely miss aromatic amino acids as tyrosine and tryptophan. The peaks were pooled and freeze-dried. The freeze-dried powders were submitted to a further reversed-phase separation on a preparative C8 column. Figure 2.8a shows for example the separation of the pool containing IB-1 and II-2. The first peak of the chromatogram corresponded to II-2 Des R_{75}. Instead, the zinc chloride precipitate was submitted directly to a purification by reversed-phase separation on a preparative C8 column, as shown in Figure 2.8b.

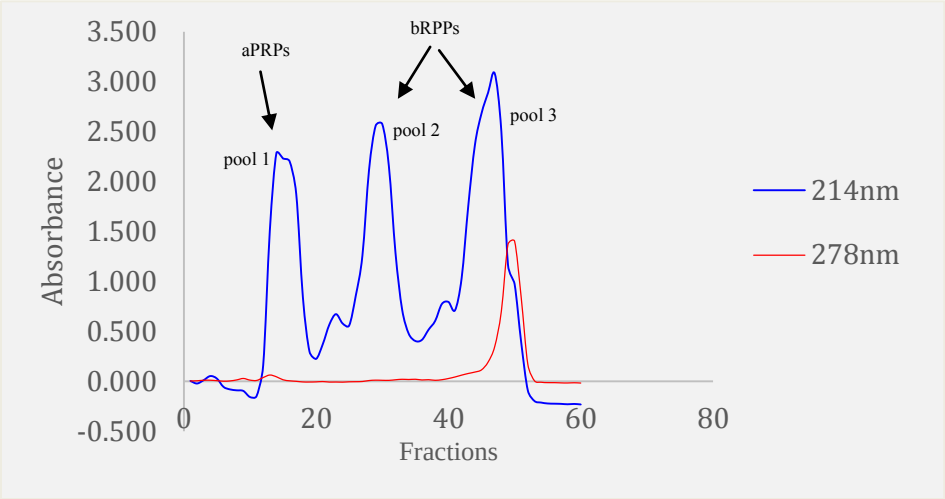
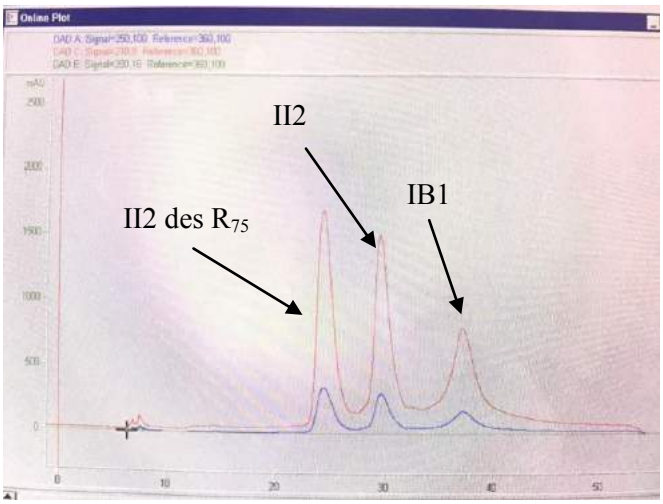


Figure 2.7 Gel filtration chromatography of the supernatant obtained from whole human saliva (80 ml) after zinc-chloride precipitation on a Sephadex-G75 column (2cm×80cm).

a)



b)

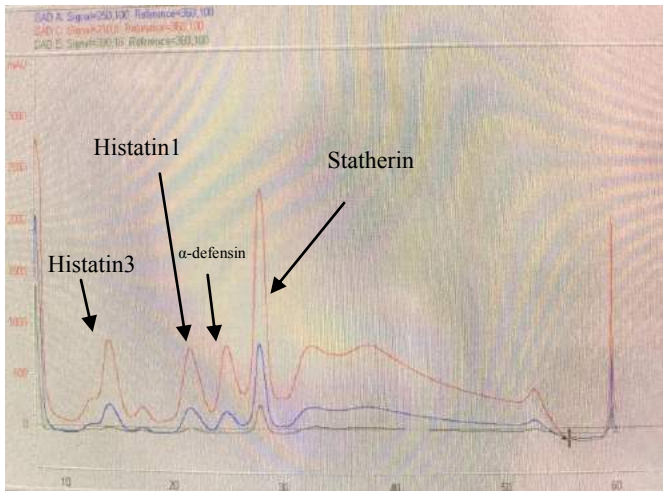


Figure 2.8a RP-HPLC on a semi-preparative C8 column of one of the freeze-dried fractions (pool 2 in Fig. 2.7) obtained from the gel filtration of Figure 2.7. The first peak corresponded to II-2 Des R_{75}, the second peak corresponded to II-2 and the third peak corresponded to IB-1.

Figure 2.8b. RP-HPLC of the zinc chloride precipitate on a semi-preparative C8 column.

2.3.2 Reaction of purified peptides with TG-2: Mapping the reactive residues (glutamines and lysines) by MSMS analyses.

The experiments carried out submitting the purified peptides to the action of TG-2 (guinea pig) evidenced that the unique product of the reactions was a cyclo-derivative. All the investigated bPRPs, P-C peptide, and statherin were able to generate a cyclo-peptide by an intrachain Gln-Lys isopeptide bound. The formation of the cycle was evinced by the appearance in the chromatographic TIC profile of a new peptide with a ΔM_{av} = -17 Da, corresponding to the loss of an ammonia molecule, as reported for example in Fig. 2.9 and Fig. 2.10 for the P-C peptide. The average and monoisotopic mass values (experimental and theoretical) of every peptide and of cyclo-derivatives, the multiply-charged ions used for MSMS fragmentation experiments, the elution times and the multiply-charged ions used for the quantifications by XIC procedure are reported in Table 2.2. The high-resolution MSMS analysis performed on the TG-2 reaction products allowed us to characterize their cyclo-derivatives and to individuate the Gln residues involved in the intra-chain crosslink. High-resolution MSMS carried out on the cyclic-derivative were sometimes indicative of the Lys residues involved (Fig. S1 – Fig. S5 supplemental files). However, P-C and P-H have only one lysine in their sequence and therefore no doubts arise about the residue involved in the formation of the cyclo-derivative. From the MSMS of the cyclo-derivative the detection of the Lys involved was possible only for P-D P_{32} and P-D A_{32}, while the detection of the Lys involved in the formation of the II-2 cyclo-derivative was not possible. The low reactivity of P-F and P-J and therefore the very low amounts of cyclo-derivatives did not generate MSMS spectra enough good for the detection of glutamine and lysine recognized by the enzyme.

Also statherin has only one lysine residue, clearly responsible for the formation of cyclo-derivative, and the high-resolution MSMS analysis allowed to confirm Q_{37} as the first residue recognized by TG-2 (Cabras T, et al. 2006).

Table 2.2 Monoisotopic and average mass values (exper. and theoretic.), elution time (with respect to the Orbitrap raw files), *m/z* ions used for MSMS fragmentation, and *m/z* ions for XIC.

Peptide Name	***Exp Monois. $[M+H]^{1+}$ (theor.)***	***Exp Mav (theor.)***	***El. time (min± 0.4)***	***m/z ions used for MSMS***	***m/z ions used for XIC***
P-D	6946.55 ± 0.04 (6946.554)	6950.0 ± 0.8 (6949.73)	16.5-17.8	1738.33 (+4)	1738.33 (+4), 1391.00 (+5), 1159.47 (+6)
cyclo-P-D	6929.55 ± 0.04 (6929.5272)	6933.20 ± 0.8 (6933.7604)	17.4-17.9	1156.27 (+6) 990.23 (+7)	1733.87 (+4), 1387.20 (+5), 1156.27 (+6)
P-D P_{32}→A	6920.54 ± 0.04 (6920.538)	6923.60 ± 0.1 (6923.69)	16.6-17.8	1731.93 (+4)	1731.93 (+4), 1384.60 (+5), 1155.00 (+6)
cyclo-P- D P_{32}→A	6903.54 ± 0.04 (6903.5115)	6907.6 ± 0.2 (6907.7222)	17.6-18.3	1727.53 (+4) 1152.10 (+6)	1727.53 (+4), 1382.23 (+5), 1152.10 (+6)
P-H	5587.77 ± 0.02 (5587.783)	5591.40 ± 0.6 (5591.14)	17.8-18.8	1398.93 (+4)	1865.4 (+3), 1398.93 (+4), 1119.06 (+5)
cyclo-P-H	5570.76 ± 0.02 (5570.7561)	5574.40 ± 0.4 (5574.1156)	18.5-19	1115.76 (+5)	1858.26 (+3), 1394.4 (+4), 1115.76 (+5)
II-2	7604.69 ± 0.04 (7604.712)	7608.20 ± 0.02 (7608.19)	20.5-21.8	1268.96 (+6)	1522.55 (+5), 1268.96 (+6), 1087.82 (+7)
cyclo-II-2	7587.71 ± 0.04 (7587.685)	7591.70 ± 0.8 (7592.2316)	22.2-23	1266.12 (+6)	1519.15 (+5), 1266.12 (+6), 1085.53 (+7)
P-F	5840.00 ± 0.02 (5839.992)	5842.50 ± 0.02 (5842.49)	17.1-18.1	835.58 (+7)	1169.41 (+5), 974.67 (+6), 835.58 (+7)
cyclo-P-F	5822.96 ± 0.02 (5822.965)	5825.99 ± 0.7 (5826.515)	18-19	-	1166.00 (+5), 971.84 (+6), 833.43 (+7)
P-J	5941.02 ± 0.02 (5941.003)	5944.02 ± 0.7 (5944.6)	17.2-18.2	850.01 (+7)	1189.61 (+5), 991.54 (+6), 850.01 (+7)
cyclo-P-J	5923.98 ± 0.02 (5923.9763)	5926.9± 0.7 (5927.5775)	18.1-19.1	-	1189.61 (+5), 991.54 (+6), 850.01 (+7)
P-C	4369.188 ± 0.2 (4369.183)	4371.9 ± 0.1 (4371.814)	16.1-16.8	1093.9 (+4) 875.4 (+5)	1458.3 (+3), 1093.9 (+4) , 875.4 (+5)
cyclo-P-C	4352.166 ± 0.2 (4352.156)	4354.70 ± 0.2 (4354.7877)	17.1-17.7	1452.06 (+3) 1089.29 (+4) 871.64 (+5)	1452.06 (+3), 1089.29 (+4), 871.64 (+5)
Statherin	5378.444 ± 0.8 (5377.4496)	5381.426 ± 0.8 (5380.7652)	30.4-31.1	1346.36 (+4)	1794.45 (+3), 1346.36 (+4), 1077.91 (+5)
cyclo-statherin	5360.436 ± 0.04 (5360.4230)	5363.435 ± 0.8 (5363.7386)	31.2-32-2	-	1788.48 (+3), 1341.61 (+4), 1073.44 (+5)

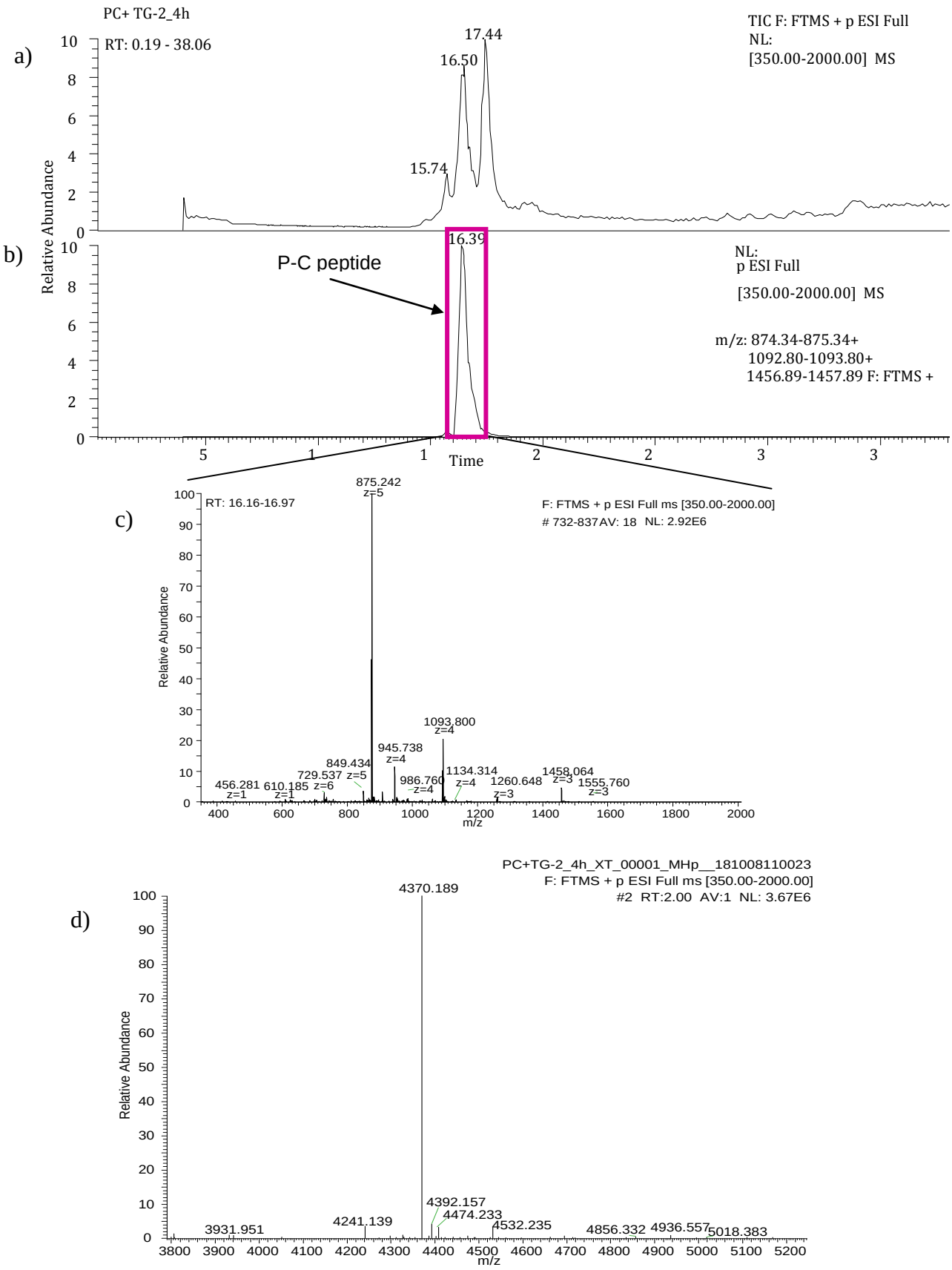


Figure 2.9 - a) TIC profile obtained from high-resolution MS analysis of P-C peptide treated for 4 hours with TG-2; **b**) XIC peak of the P-C peptide (Area of XIC peak used for quantitative analysis); **c**) *m/z* spectrum and **d**) deconvoluted mass spectrum.

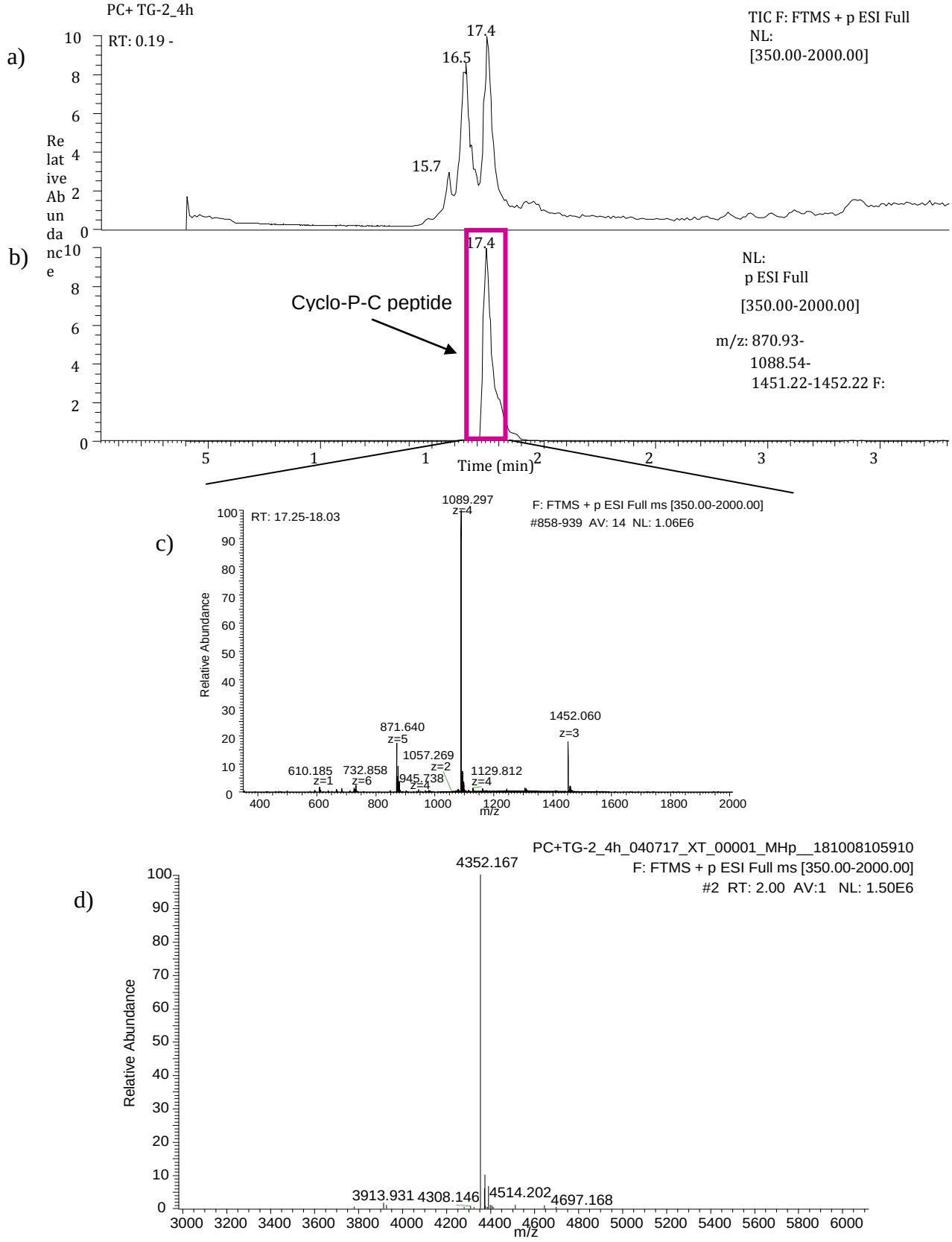


Figure 2.10 - a) TIC profile obtained from high-resolution MS analysis of P-C peptide treated for 4 hours with TG-2; **b**) XIC peak of the cyclo-P-C peptide (Area of XIC peak used for quantitative analysis); **c**) *m/z* spectrum and **d**) deconvoluted mass spectrum.

The annotated sequences reported in the following section highlight only the fragment ions that were significant to map the Q residues. The complete annotated spectra are showed in the supplemental section.

Cyclo-P-D.

The residues recognized by TG-2 on P-D P_{32} and P-D A_{32} were **Q_{37}** and **K_{25}** (with **Q_{40}** as minor site**)**. The fragments b_{19}, b_{25}, y_{32} and y_{33} after MSMS analysis of the ion $[M+7H]^{7+}$ 990.23 *m/z* (CID) of cyclo-P-D, showed in **Fig. S1**, and of the ion $[M+6H]^{6+}$ 1156.6 *m/z* (CID) discriminated for **Q_{37}** and **K_{25}**. The same result was obtained for the cyclo-P-D A_{32}, the MSMS analysis on the ion $[M+6H]^{6+}$ 1152.09 *m/z* (CID) is reported in **Fig. S2**.

b_{19} b_{25}

SPPGKPQGPP QQEGNKPQG⌉P PPPG⌊K_{25}⌉PQGPP PP$_{32}$(/A)GGNPQ_{37}⌊Q⌊P

y_{55} y_{33} y_{32}

Q_{40}APPAGKPQG P PPPPQGGRP P RPAQGQQPP Q

Cyclo-P-H.

The residues recognized by TG-2 on P-H were **Q_{29}** and **K_5**, **(**with **Q_{11}** probably as minor site**).** As previously described, **K_5** is the unique lysine residue in the sequence. The fragments b_{29}, y_{27} and y_{54} after MSMS analysis of the ion $[M+5H]^{5+}$ 1115.76 *m/z* (CID) of cyclo P-H discriminated for **Q_{29}** (**Fig. S3)**

b_{29}

SP⌊PGK_5PQGPP Q_{11}QEGNNPQGP PPPAGGNPQ_{29}⌉⌊Q PQAPPAGQPQ

y_{54} y_{27}

GPPRPPQGGR PSRPPQ

Cyclo-II-2.

The residue recognized by TG-2 on II-2 was **Q_{21}**, **(**the **K** involved was not established until now**).** The fragments y_{35}, y_{56}, y_{59} and y_{63} after MSMS analysis of the ion $[M+6H]^{6+}$ 1266.12 *m/z* (CID) of cyclo II-2 discriminated for **Q_{21}**. (**<Q:** pyro-glutamic; **S** phosphorylated serine) (**Fig. S4)**

<QNLNEDVSQE ES⌊PSLI⌊AGN⌊P Q_{21}GPSPQGGNK PQGPPPPPGK
y_{63} y_{59} y_{56}

⌊PQGPPPQGGN KPQGPPPPGK PQGPPPQGDK SRSPR
y_{35}

Cyclo-P-F and cyclo-P-J.

The very low reactivity of **P-F** and **P-J** did not allow to establish the residues recognized by TG-2 until now.

P-F

SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGGSKSRS A

P-J

SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSRS S

Cyclo-P-C.

The residues recognized by TG-2 on P-C were Q_{41} and K_{23}, **(**with Q_{39} probably as minor site**).** As previously described K_{23} is the unique lysine residue in the P-C sequence. The fragments b_{41}, b_{42} and y_{25} after MSMS analysis of the ion $[M+4H]^{4+}$ 1089.29 *m/z* (CID) of cyclo P-C discriminated for Q_{41} (**Fig. S5).**

GRPQGPPQQG GHQQGPPPP⌊P PGK_{23}PQGPPPQ GGRPQGPPQ_{39}G
b_{41} b_{42} y_{25}

Q_{41}⌉S⌉PQ

Cyclo-statherin.

It was not possible to confirm by MSMS analysis that Q_{37} is the residue involved in the formation of the cyclic derivative, as demonstrated in a previous study (Cabras T, et al. 2006) by enzymatic digestion of statherin and cyclo statherin with carboxypeptidase. K_6 is the unique lysine residue present in the statherin sequence and surely involved in the formation of cycle.

DSSEEK_6FLRR IGRFGYGYGP YQPVPEQPLY PQPYQPQ_{37}YQQ YTF

2.3.3 Reaction of purified peptides with TG-2: Quantitative considerations about the reaction products

The results demonstrated that bPRPs were therefore substrates of TG-2, but with very different reactivity. The area of the extracted ion current (XIC) peaks of the linear and cyclic peptides were measured in the TIC profile at different incubation times.

Results clearly indicated that the different peptides under study displayed a very different reactivity for the formation of the cyclo-derivative under the action of TG-2 (guinea pig). Among bPRPs, P-D and P-H were more reactive with respect to P-F, P-J and II-2, whose reactivity was almost negligible. The P-C peptide, which has to be considered a peptide pertaining to the family of aPRPs, generated a high percentage of cyclo-derivative, but with the lowest reactivity. Finally, high percentages of cyclo-statherin were generated by TG-2 in a fast time. Interestingly, the P-D P_{32}→A variant showed a reactivity significantly lower than P-D, suggesting that a more rigid peptide conformation facilitated the formation of the cyclo-derivative.

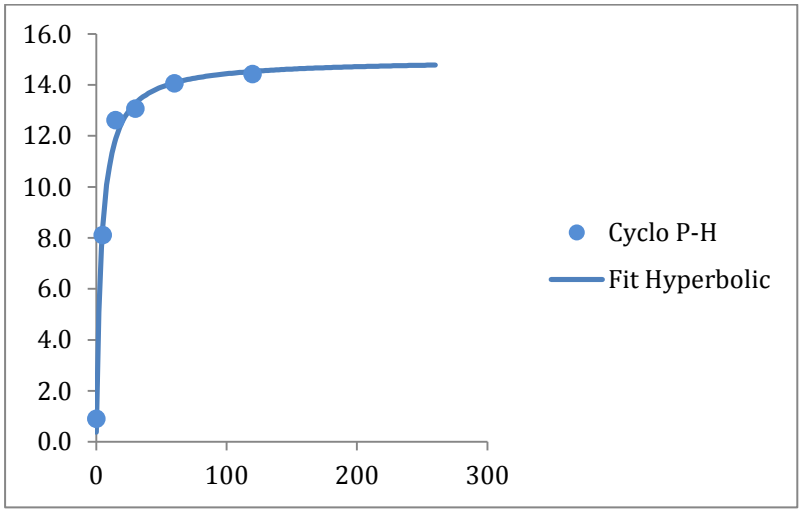


Fig. 2.11 Percentages of cyclo-P-H peptide generated TG-2 at 37°C and measured from the area of XIC peaks at different times of incubation. %*max* = 15; *t(1/2)* = 4 min; recovery = 78%.

Fig. 2.11 shows for example the results obtained for the P-H peptide. In the y axis the percentages of the reaction product (at different times of incubation) obtained computing the ratio between the area of the XIC peak of the cyclo-derivative and the area of the XIC peak of the peptide under study at zero time (x100) are reported. This value was indicated as %(*t*). On the x axis the time of incubation (min) are reported.

The %(*t*) was fitted as a function of time of incubation according to the empirical hyperbolic equation:

Equation 1

$$\%(t) = \frac{\%max \times t}{t(\frac{1}{2}) + t}$$

Where the %*max* is the percentage obtained by the fit for $t \rightarrow \infty$ and *t(1/2)* is the time necessary the reach half of %*max*. Indeed, for $t >> t(1/2)$

$$\%(t) = \frac{\%max \times t}{t(\frac{1}{2}) + t} \cong \frac{\%max \times t}{t} \cong \%max$$

And for $t = t(1/2)$

$$\%(t) = \frac{\%max \times t}{t(\frac{1}{2}) + t} = \frac{\%max \times t(\frac{1}{2})}{2 \times t(\frac{1}{2})} = \frac{\%max}{2}$$

Obviously, comparisons between %*max* and *t(1/2)* obtained using different peptides and different incubation conditions (i.e. temperature) can be made only if similar concentration of peptide and enzyme in the incubation solution are used (conditions employed in all the experiments performed).

In Fig.2.11 the values of %*max* and *t(1/2)* obtained by the best fitting procedure are also reported.

Peptides submitted to the action of TG-2, further than generate a cyclo-derivative, can cross react generating an insoluble network. In order to have a rough information of this potential reaction, a recovery percentage was computed as the ratio between the sum of the area of the XIC peaks of un-reacted peptide and cyclo-peptide and the area of the XIC peak of the peptide under study at zero time (x100). Low value of this recovery suggests that the peptide was involved in cross-reactions generating products undetectable in the HPLC-ESI-MS profile.

In Table 2.3 the values of %*max*, *t(1/2)* and recovery obtained at 37 °C for all the peptides under study are reported.

Table 2.3 values of %max, t(1/2) and recovery at 37 °C for all the peptides under study

Peptide Name	*t(1/2)(min)*	*%max*	Corr. coeff.	Recovery
cyclo-P-D	2	25	0.96	60
cyclo-P- D P_{32}→A	10	12	0.99	55
cyclo P-H	4	15	0.99	70
cyclo II-2	2	5	0.80	87
cyclo P-F	2	5	0.75	88
cyclo P-J	4	4	0.91	88
cyclo-P-C	16	49	0.99	78
cyclo-statherin	3	38	0.85	42

2.3.4 Reaction of purified peptides with TG-2 in the presence of dansyl-cadaverine: Mapping the reactive residues (glutamines and lysines) by MSMS analyses.

In the experiments performed in the presence of dansyl-cadaverine in order to map the glutamine residues reactive to TG-2, we observed that the cyclo-derivative was produced despite DC competed with lysine as lone pair donor to glutamine. However, multiple DC-adduct were detected for all the peptides under study and the amount of the cyclo-derivatives generated by TG-2 was reduced with respect to the experiments performed with TG-2 alone. Moreover, DC-adducts of the cyclo-peptides were detected suggesting that specific glutamine residues were preferably involved in the intra-chain cross-link with the lysine residue.

The average and monoisotopic mass values (experimental and theoretical) of every DC-adduct generated by TG-2 action, the multiply-charged ions used for MSMS fragmentation experiments, the elution times and the multiply-charged ions used for the quantifications by XIC procedure are reported in **Table 2.4**.

Table 2.4 Monoisotopic and average mass values (exper. and theoretic.), elution time (with respect to the Orbitrap raw files), *m/z* ions used for MSMS fragmentation, and *m/z* ions for XIC.

Peptide Name	Exp. Monois. $[M+H]^{1+}$ (theor.)	Exp Mav (theor.)	El. time (min± 0.4)	m/z ions (MSMS)	m/z ions (XIC)
P-D + 1 DC	7264.714 ± 0.04 (7264.6940)	7268.711 ± 04 (7268.9272)	20.8-21.2	1212.29 (+6) 1039.53 (+7)	1212.29 (+6) 1039.53 (+7) 909.47 (+8)
cyclo P-D + 1 DC	7247.693 ± 0.04 (7247.6674)	7251.690 ± 0.4 (7251.9006)	21.1-21.5	-	1209.452 (+6) 1036.962 (+7) 907.468 (+8)
P-D + 2 DC	7582.839 ± 0.04 (7582.8342)	7586.841± 0.8 (7587.0674)	24.2-25.2	1265.313 (+6) 1084.698 (+7)	1265.313 (+6) 1084.698 (+7) 949.237 (+8)
P-D P_{32}→A + 1 DC	7238.689 ± 0.04 (7238.6783)	7242.692 ± 0.4 (7242.8890)	20.7-21.3	1207.95 (+6) 1035.82 (+7)	1207.95 (+6) 1035.53 (+7) 906.22 (+8)
P-D P_{32}→A + 2 DC	7556.815 ± 0.04 (7556.8185)	7560.826± 0.8 (7561.0292)	24.1-24.5	-	1260.976(+6) 1080.982(+7) 945.985 (+8)
P-H + 1 DC	5905.945 ± 0.04 (5905.9229)	5908.945 ± 0.4 (5909.2824)	22.2-23.2	1182.59 (+5) 985.66 (+6) 844.99 (+7)	1182.59 (+5) 985.66 (+6) 844.99 (+7)
cyclo P-H + 1 DC	5888.9± 0.04 (5888.8963)	5892.9± 0.8 (5892.2558)	22.5-23.5	-	1179.39 (+5) 982.99 (+6) 842.71 (+7)
P-H + 2 DC	6224.073 ± 0.02 (6224.0631)	6227.079 ± 0.4 (6227.4226)	27.1-28.1	1246.42 (+5) 1038.85 (+6)	1246.42 (+5) 1038.85 (+6) 890.45 (+7)
II-2 + 1 DC	7922.8 ± 0.06 (7922.8521)	7926.87 ± 0.8 (7927.3984)	26.1-27.5	1133.27 (+7)	1321.99 (+6) 1133.27 (+7) 991.74 (+8)
cyclo II-2 + 1 DC	7905.9± 0.1 (7905.8255)	7909.98± 0.7 (7910.3718)	26.6-28	-	1319.98 (+6) 1131.56 (+7) 990.11 (+8)
P-F + 1 DC	6158.14 ± 0.02 (6158.1318)	6161.14 ± 0.7 (6161.6822)	26.2-26.8	-	1027.53 (+6) 881.55 (+7) 771.13 (+8)
P-J + 1 DC	6259.16 ± 0.02 (6259.1431)	6262.15 ± 0.7 (6262.7443)	26.2-26.8	-	895.46 (+7) 783.65 (+8)
P-C + 1 DC	4687.33± 0.2 (4687.3236)	4689.33 ± 0.6 (4689.9545)	21.3-21.6	1173.09 (+4) 938.67 (+5)	1173.09 (+4) 938.67 (+5) 782.39 (+6)
cyclo P-C + 1 DC	4670.3 ± 0.04 (4670.2970)	4673.31 ± 0.8 (4672.9279)	22.3-22.9	935.47 (+5)	1168.83 (+4) 935.47 (+5) 779.72 (+6)

P-C + 2 DC	5005.46 ± 0.04 (5005.4638)	5008.47 ± 0.8 (5008.0947)	25.1-26.2	1002.5 (+5) 835.56 (+6)	1002.5 (+5) 835.56 (+6)
statherin + 1 DC	5695.589 ± 0.02 (5695.5898)	5698.599 ± 0.4 (5698.9054)	32.9-33.5	1140.33 (+5)	1425.40 (+4) 1140.33 (+5)
cyclo statherin + 1 DC	5678.55 ± 0.04 (5678.5632)	5682.55 ± 0.8 (5681.8788)	33-33.2	-	1421.4 (+4) 1137.2 (+5)
statherin + 2 DC	6013.710 ± 0.04 (6013.7300)	6016.723 ± 0.8 (6017.0456)	34.2-34.5	1504.93 (+4) 1204.55 (+5)	1504.935 (+4) 1204.351 (+5) 1007.616 (+6)
cyclo statherin + 2 DC	5996.696 ± 0.04 (5996.7034)	5999.703 ± 0.8 (6000.0190)	34.6-35	-	1500.681 (+4) 1200.947 (+5) 1003.628 (+6)
statherin + 3 DC	6331.838 ± 0.06 (6331.8702)	6334.870 ± 0.4 (6335.1858)	35.9-36.5	1584.73 (+4) 1267.98 (+5)	1584.727 (+4) 1267.780 (+5) 1056.651 (+6)

MSMS fragmentation spectra carried out on the DC-derivatives showed that only specific glutamine residues were recognized by TG-2 enzyme. Also in this case, the low reactivity of P-F and P-J and therefore the very low amounts of DC-derivatives did not generate MSMS spectra enough good for the detection of glutamines recognized by the enzyme. MSMS characterization was not possible, in addition, for the DC-adducts of the cyclo-peptides. The annotated sequences reported in the following section highlight only the fragment ions that were significant to map the Q residues. The complete annotated spectra are shown in the supplemental section.

P-D + dansyl-cadaverine.

The MSMS analysis of the ion $[M+7H]^{7+}$ 1039.53 *m/z* (CID) of P-D + 1DC The fragments b_{37}, y_{32}, y_{33} and y_{34} discriminated for **Q_{37}** **(Fig. S6).** The same results were obtained for P-D A_{32} variant, the MSMS analysis performed on the ion $[M+7H]^{7+}$ 1039.53 *m/z* is reported in **Fig. S7).**

SPPGKPQGPP QQEGNKPQGP PPPGK$_{25}$PQGPP PP$_{32}$(/A)GGNP⌊Q$_{37}$⌉⌊Q⌊P
QAPPAGKPQG P PPPPQGGRP P RPAQGQQPP Q

b_{37} y_{34} y_{33} y_{32}

The fragments b_{36}, b_{37}, b_{38}, b_{41}, y_{30}, y_{32} $y_{33,}$ and y_{35} after MSMS analysis of the ion $[M+7H]^{7+}$ 1084.7 *m/z* (CID) of P-D + 2DC discriminated for **Q_{37}** and **Q_{40}** **(Fig. S8).** The MSMS fragmentation for P-D A_{32} + 2 DC was not good enough, thus until now it was not possible to confirm **Q_{40}** as second acceptor for DC.

SPPGKPQGPP QQEGNKPQGP PPPGK$_{25}$PQGPP PP$_{32}$GGN⌊P⌉Q$_{37}$⌉⌊Q⌉⌊P
Q$_{40}$A⌉⌊PPAGKPQG P PPPPQGGRP P RPAQGQQPP Q

b_{36} b_{37} b_{38} y_{35} y_{33} y_{32}

b_{41} y_{30}

P-H + dansyl-cadaverine.

The MSMS analysis of the ion $[M+6H]^{+6}$ 985.66 *m/z* (CID) of P-H + 1DC highlighted the presence of the fragment ions b_{29}, b_{30}, y_{27} and y_{29}, which allowed to

individuate the **Q29** as the site recognized by TG-2 to link dansyl-cadaverine (**Fig. S9).**

b_{29} b_{30}

SPPGK$_5$PQGPP QQEGNNPQGP PPPAGGNPQ$_{29}$QPQAPPAGQPQ GPPRPPQGGR PSRPPQ

y_{29} y_{27}

The fragments b_{13}, b_{16}, y_{27}, y_{29}, y_{40} and y_{48} generated by MSMS analysis of the ion $[M+5H]^{5+}$ 1246.42 *m/z* (CID) of P-H + 2DC allowed to individuate as reactive sites for DC **Q$_{29}$** and **Q$_{11}$** (**Fig. S10).**

b_{13} b_{16}

SPPGK$_5$PQGPP Q$_{11}$QEGNNPQGP PPPAGGNPQ$_{29}$QPQAPPAGQPQ GPPRPPQGGR PSRPPQ

y_{48} y_{40} y_{29} y_{27}

II-2 + dansyl-cadaverine.

II-2 peptide was able to bound only one DC, the MSMS sequencing of the DC-adduct, performed on the ion $[M+7H]^{7+}$ 1133.42 *m/z* (CID), was in agreement with the DC linking on **Q$_{21}$**, in particular the ion fragments y_{54} and y_{56} were discriminating for this Q residue (**Fig. S11).**

<QNLNEDVSQE ESPSLIAGNP Q$_{21}$GPSPQGGNK PQGPPPPPGK PQGPPPQGGN KPQGPPPPGK PQGPPPQGDK SRSPR

y_{56} y_{54}

P-C + dansyl-cadaverine.

P-C peptide was able to react with two DC moieties and MSMS analysis allowed to individuate the specific Q residues involved. The MSMS analysis of the ion $[M+5H]^{5+}$ 938.8 *m/z* (CID) of P-C + 1DC was in accordance with the presence of the dansyl-cadaverine moiety linked to **Q$_{41}$**, the detection of the ion fragments b_{40}, b_{41}, b_{42} and y_5 was critical for this attribution (**Fig. S12)**.

GRPQGPPQQG GHQQGPPPPP PGK$_{23}$PQGPPPQ GGRPQGPPQ$_{39}$G

b_{40} y_5

b_{41} b_{42}

Q$_{41}$S PQ

In the case of P-C + 2DC, the MSMS analysis of the ion $[M+6H]^{6+}$ 835.58 *m/z* (CID) was in accordance for the DC linking to **Q$_{41}$** and **Q$_{39}$**, as shown in the annotated fragmentation spectrum reported in **Fig. S13,** where the detection of the ion fragments b_{39}, b_{40}, b_{41} and b_{42} was discriminating for the attribution.

GRPQGPPQQG GHQQGPPPPP PGK$_{23}$PQGPPPQ GGRPQGPPQ$_{39}$G

b_{39} b_{40}

b_{41} b_{42}

Q$_{41}$S PQ

It was possible to characterize also the DC adduct of cyclo-P-C (**Fig. S14)**. The ion fragments b_{40}, b_{41}, b_{42} and y_5 detected in the MSMS fragmentation spectrum obtained from the ion $[M+5H]^{5+}$ 935.478 *m/z* (CID) of cyclo P-C + 1DC discriminated for the **Q$_{41}$**.

GRPQGPPQQG GHQQGPPPP P PGK$_{23}$PQGPPPQ GGRPQGPPQ$_{39}$G

b_{19} y_{24}

b_{41} b_{42}

Q$_{41}$S PQ

Statherin + dansyl-cadaverine.

Staherin was able to link until 3 DC moieties, and cyclo-statherin linked until 2 DC moieties. The MSMS analysis of DC-derivatives of statherin confirmed that **Q$_{37}$** is the main residue involved but also evidenced that **Q$_{39}$** and **Q$_{40}$** are secondary sites of TG-2 recognition (**Fig. S15-S17).**

The detection of the ion fragments b_{35}, b_{37}, b_{38} and y_8 in the fragmentation MSMS spectrum obtained from the ion $[M+5H]^{5+}$ 1140.33 *m/z* (CID) of the statherin + 1DC discriminated for **Q$_{37}$**. **(S** phosphorylated serine)

DSSEEK$_6$FLRR IGRFGYGYGP YQPVPEQPLY PQPYQ$^{b_{35}}$ $_{y_8}$**PQ$_{37}$**$^{b_{37}}$**Y**$^{b_{38}}$**Q$_{39}$Q$_{40}$**

YTF

The MSMS analysis of the ion $[M+5H]^{5+}$ 1204.55 *m/z* (CID) of the statherin + 2DC was in accordance with the Dc linking to **Q$_{37}$** and **Q$_{39}$** residues, and the detection of the ion fragments b_{37}, b_{38}, b_{39}, b_{40}, y_5, y_6, y_8 and y_9 was discriminating for this attribution.

DSSEEK$_6$FLRRIGRFGYGYGP YQPVPEQPLY PQPY$_{y_9}$**Q**$_{y_8}$**PQ$_{37}$**$^{b_{37}}$ $_{y_6}$**Y**$^{b_{38}}$ $_{y_5}$**Q$_{39}$**$^{b_{39}}$**Q$_{40}$**$^{b_{40}}$

YTF

The MSMS analysis of the ion $[M+5H]^{5+}$ 1267.98 *m/z* (CID) of the statherin + 3DC revealed that the **Q$_{37}$**, **Q$_{39}$** and **Q$_{40}$** residues were the linking sites for DC moieties, and the detection of the ion fragments b_{37}, b_{38}, b_{39}, b_{40}, y_5, y_6, y_8 and y_9 was discriminating for this attribution.

DSSEEK$_6$FLRRIGRFGYGYGP YQPVPEQPLY PQPYQ$_{y_8}$**PQ$_{37}$**$^{b_{37}}$ $_{y_6}$**Y**$_{y_5}$**Q$_{39}$**$^{b_{39}}$ $_{y_4}$**Q$_{40}$**$^{b_{40}}$

$^{b_{41}}$**Y TF**

2.3.5 Reaction of purified peptides with TG-2 in the presence of dansyl-cadaverine: Quantitative considerations about the reaction products

Incubations performed in the presence of dansyl-cadaverine as a function of time provided results sometimes complex for the following reasons:

a) DC competed with lysine as lone pair donor to glutamine. Consequently, all the peptides under study reduced the amount of the cyclo-derivative generated by TG-2 with respect to the experiments performed with TG-2 alone.

b) Some of the peptides under study had more than one reactive glutamine residue as evident by the multiple derivatives detectable in the HPLC-ESI-MS profile (Fig 2.12). The percentages of the intermediate derivative (i.e. 1-DC-derivative and cyclo-derivative when the peptide had two reactive glutamines or 1- and 2-DC-derivatives, cyclo- and cyclo-1-DC-derivatives when the peptide had three reactive glutamines), when plotted as a function of time, showed a biphasic trend with an initial increase of the first reaction product, followed by its decrease due to further transformations by the action of TG-2.

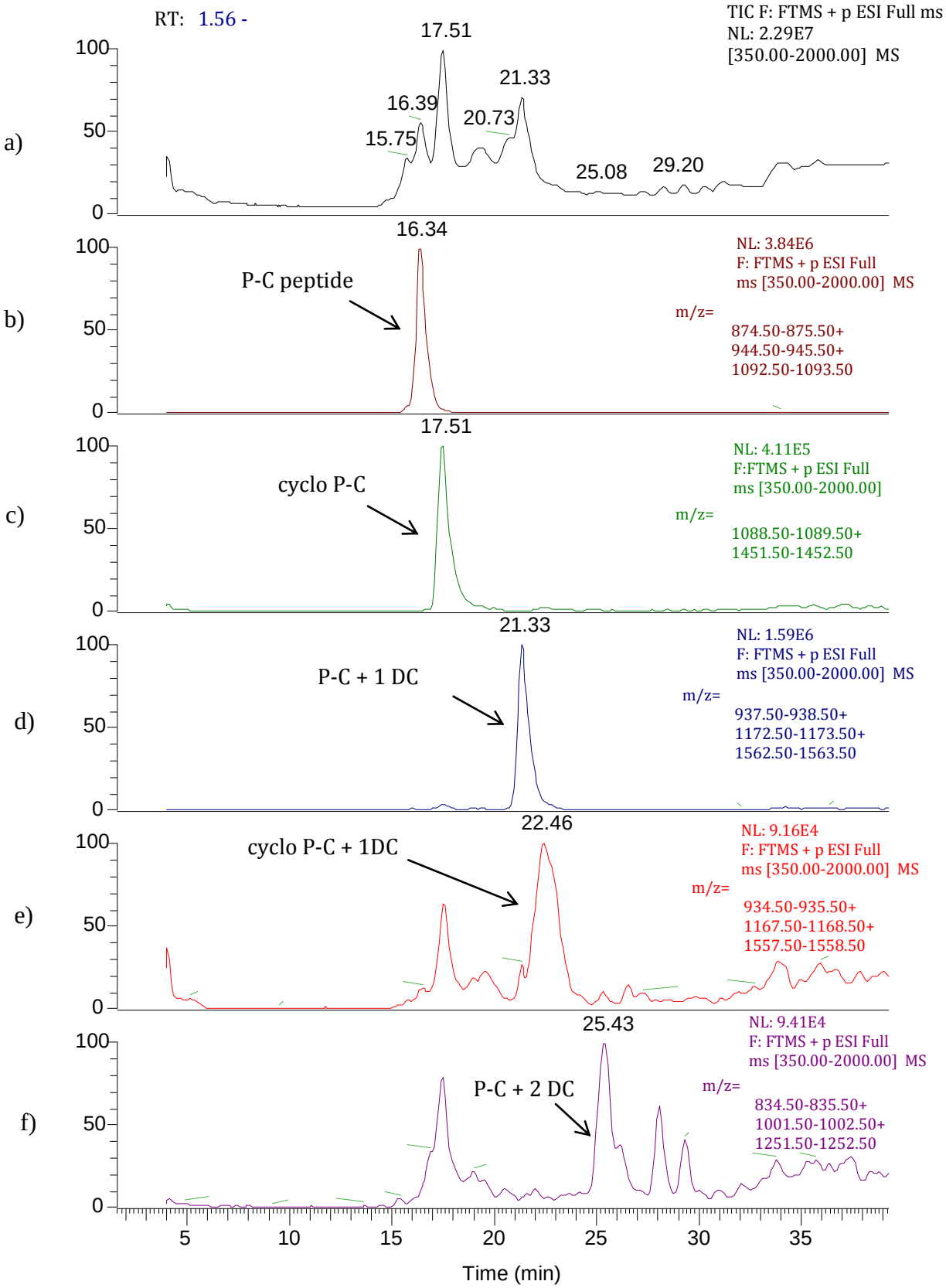


Figure 2.12 - **a**) TIC profile obtained by high resolution MS analysis of P-C peptide treated for 4 hours with TG-2 in the presence of dansyl-cadaverin.; **b**) XIC peak of P-C peptide (Area of XIC peak used for quantitative analysis); **c**) XIC peak of cyclo P-C ; **d**) XIC peak of P-C + 1DC ; **e**) XIC peak of cyclo P-C + 1DC; **f**) XIC peak of P-C + 2DC.

For the above reasons, it was impossible to fit the percentages of the intermediate derivatives using the empirical hyperbolic equation 1 described in the previous paragraph.

Even though the results obtained cannot be quantified by simple equations, the results reported in Figures 2.13-2.20 provided information in agreement with the experiments in the presence of TG-2 alone. In the Figures, on the y axis the percentages were plotted with a logarithmic scale in order to amplify at a glance the lowest values.

All the bPRPs investigated reacted with DC, but the most reactive were, P-H, P-D P_{32} and P-D A_{32}, while the reactivity of II-2, P-F and P-J was negligible. P-H, had two DC reactive sites, II-2, P-F and P-J one reactive site. The comparison between P-D and P-D P_{32}→A variant confirm that P-D P_{32} is more reactive than the P-D A_{32} variant because the 2-DC derivatives and of the cyclo-1-DC-derivative were not detected in the latter. P-C has two DC reactive sites while statherin was the most reactive, and it had three DC reactive sites.

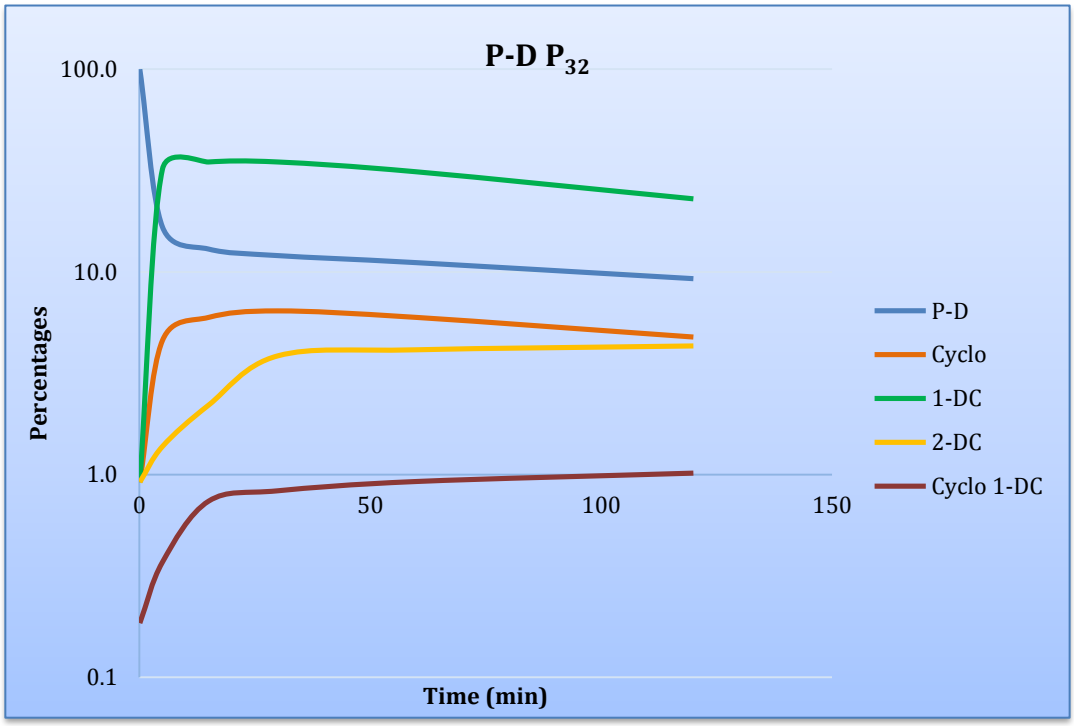


Figure 2.13 Variations of the percentages of the derivatives of P-D P_{32} as a function of time (min) during the incubation with TG-2 and DC (37 °C).

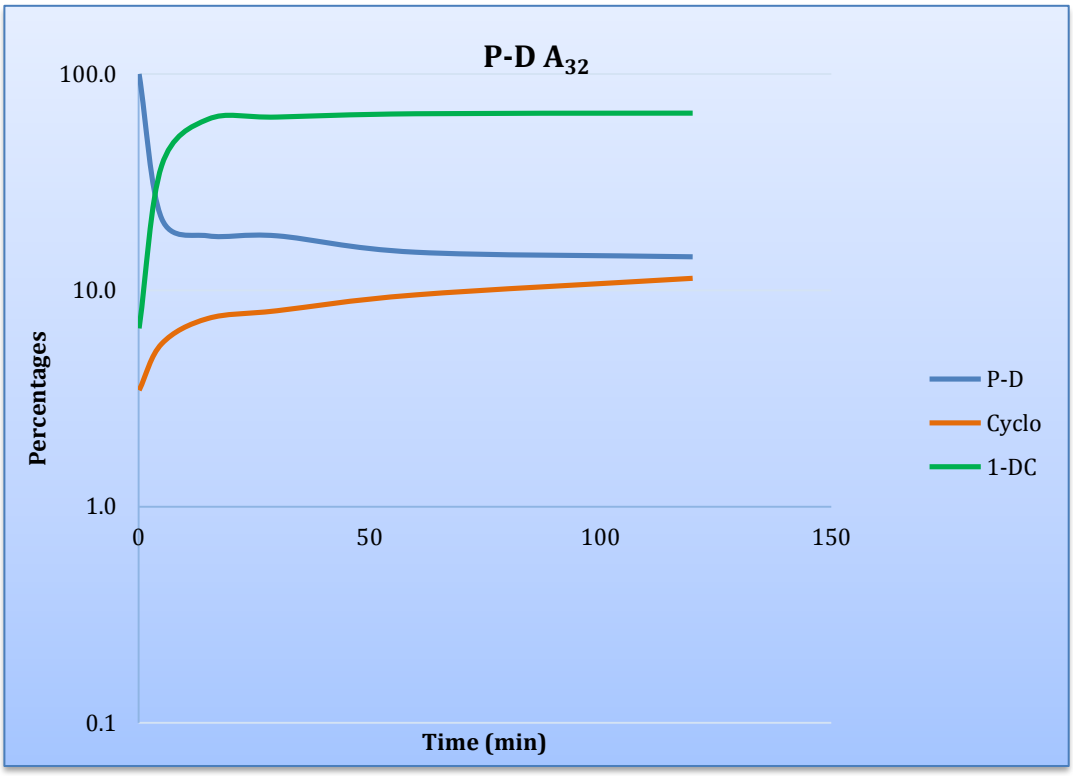


Figure 2.14 Variations of the percentages of the derivatives of P-D A_{32} as a function of time (min) during the incubation with TG-2 and DC (37 °C).

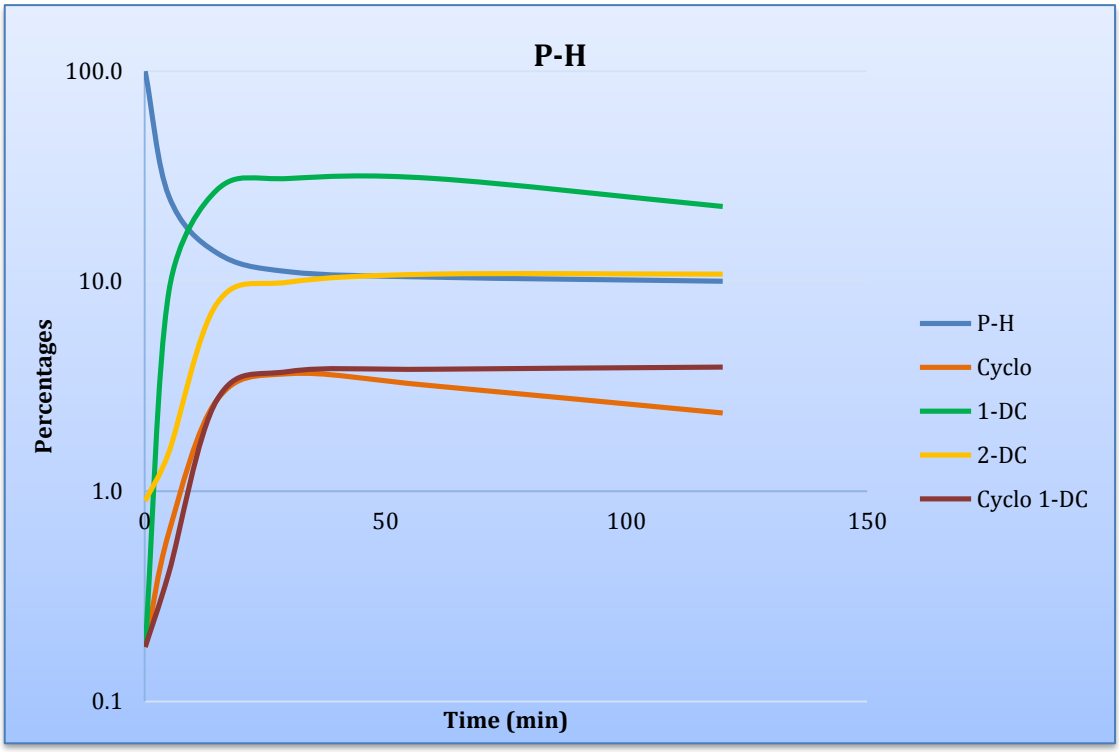


Figure 2.15 Variations of the percentages of the derivatives of P-H as a function of time (min) during the incubation with TG-2 and DC (37 °C).

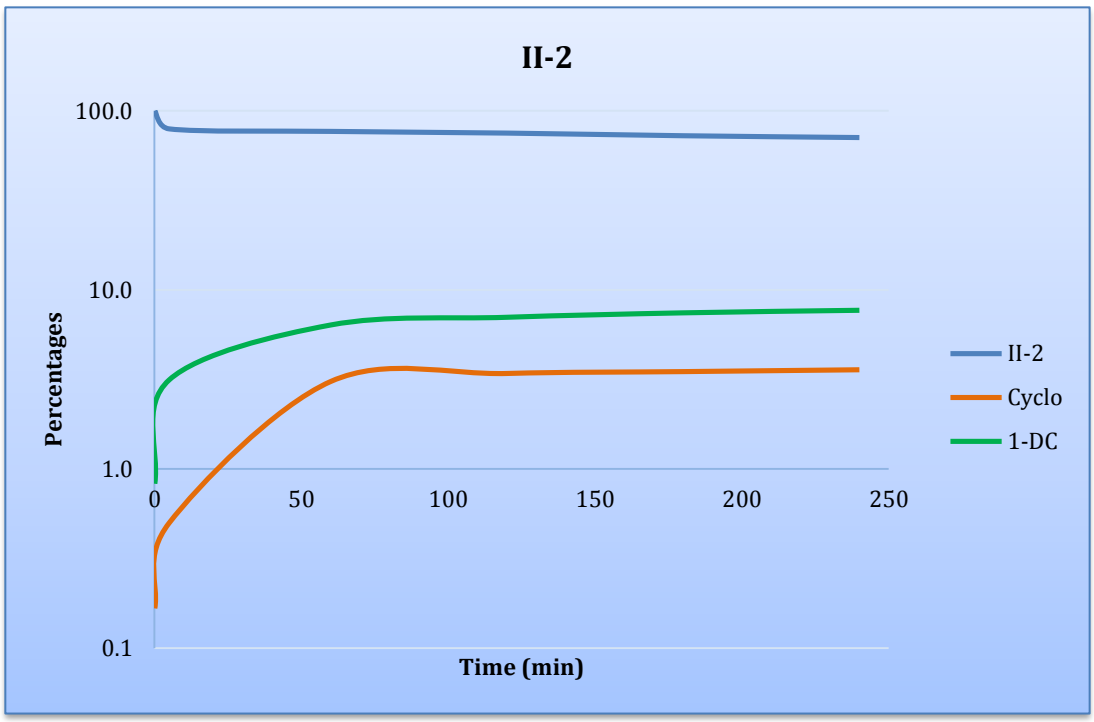


Figure 2.16 Variations of the percentages of the derivatives of II-2 as a function of time (min) during the incubation with TG-2 and DC (37 °C).

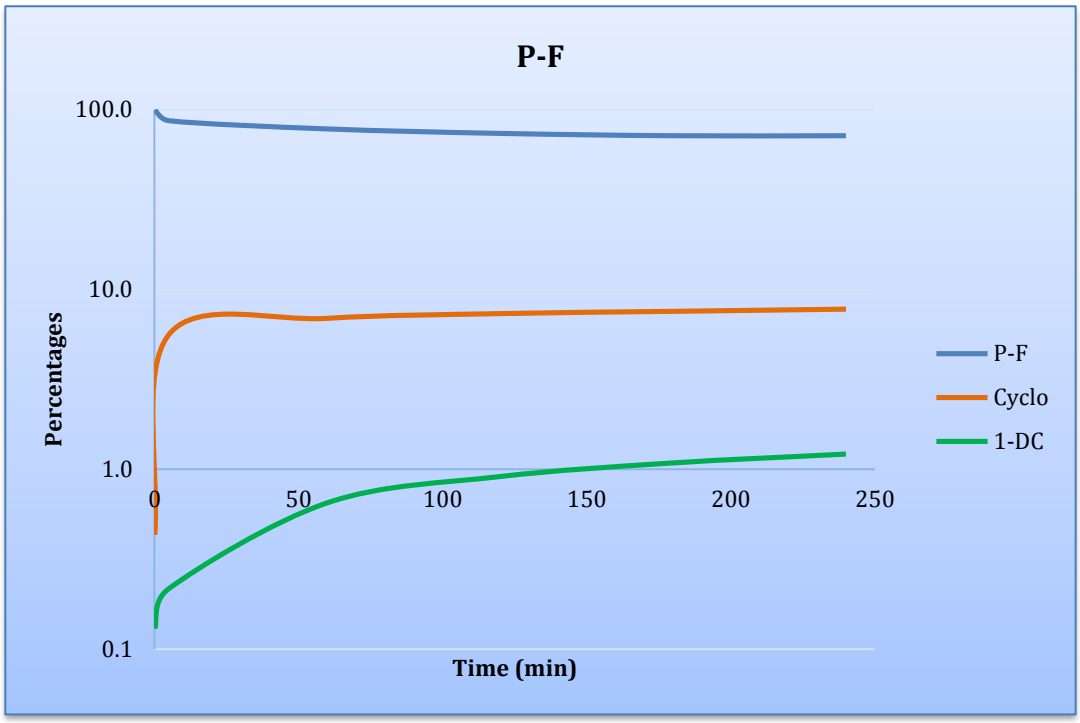


Figure 2.17 Variations of the percentages of the derivatives of P-F as a function of time (min) during the incubation with TG-2 and DC (37 °C).

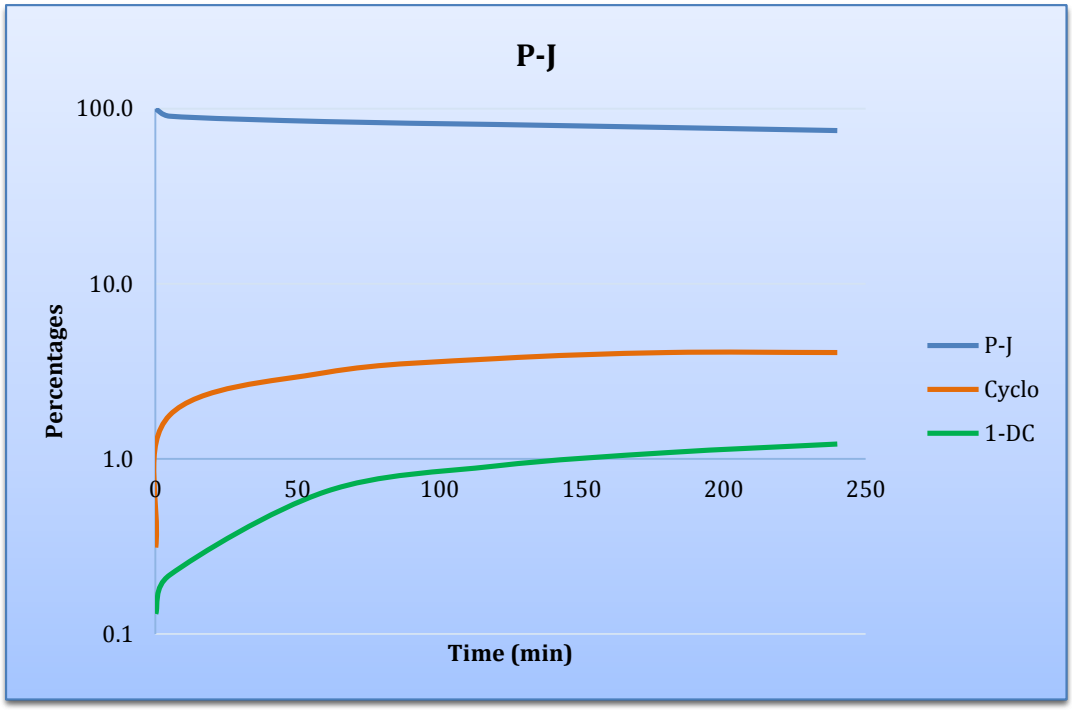


Figure 2.18 Variations of the percentages of the derivatives of P-J as a function of time (min) during the incubation with TG-2 and DC (37 °C).

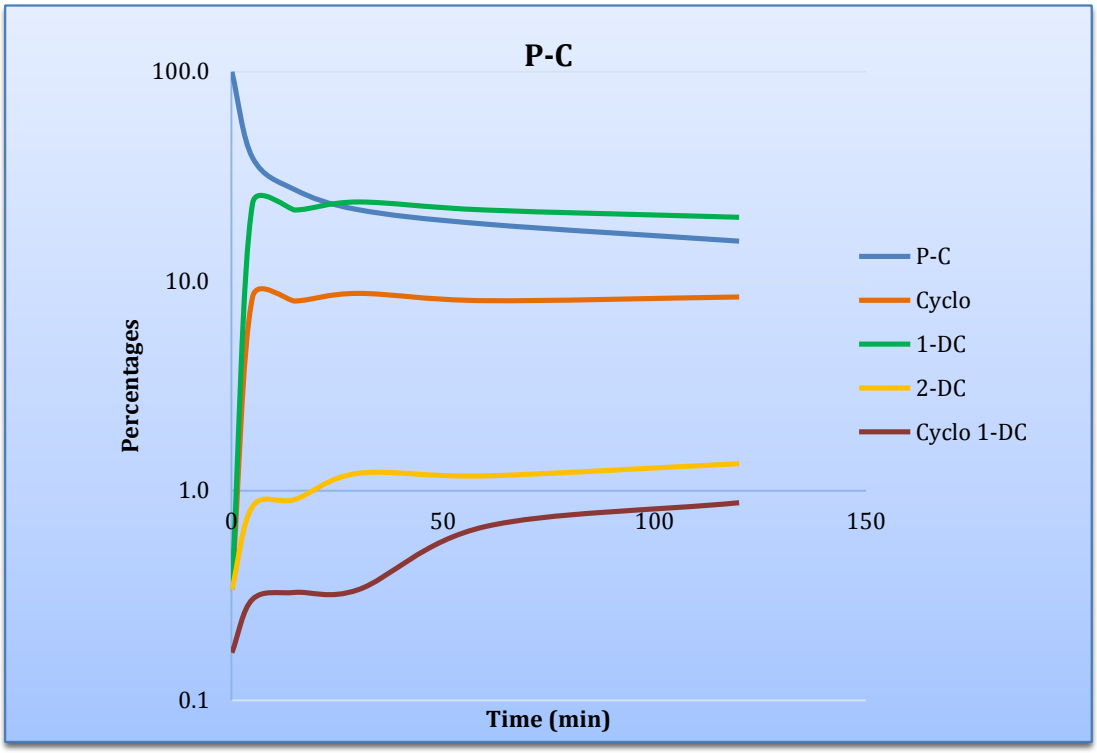


Figure 2.19 Variations of the percentages of the derivatives of P-C as a function of time (min) during the incubation with TG-2 and DC (37 °C).

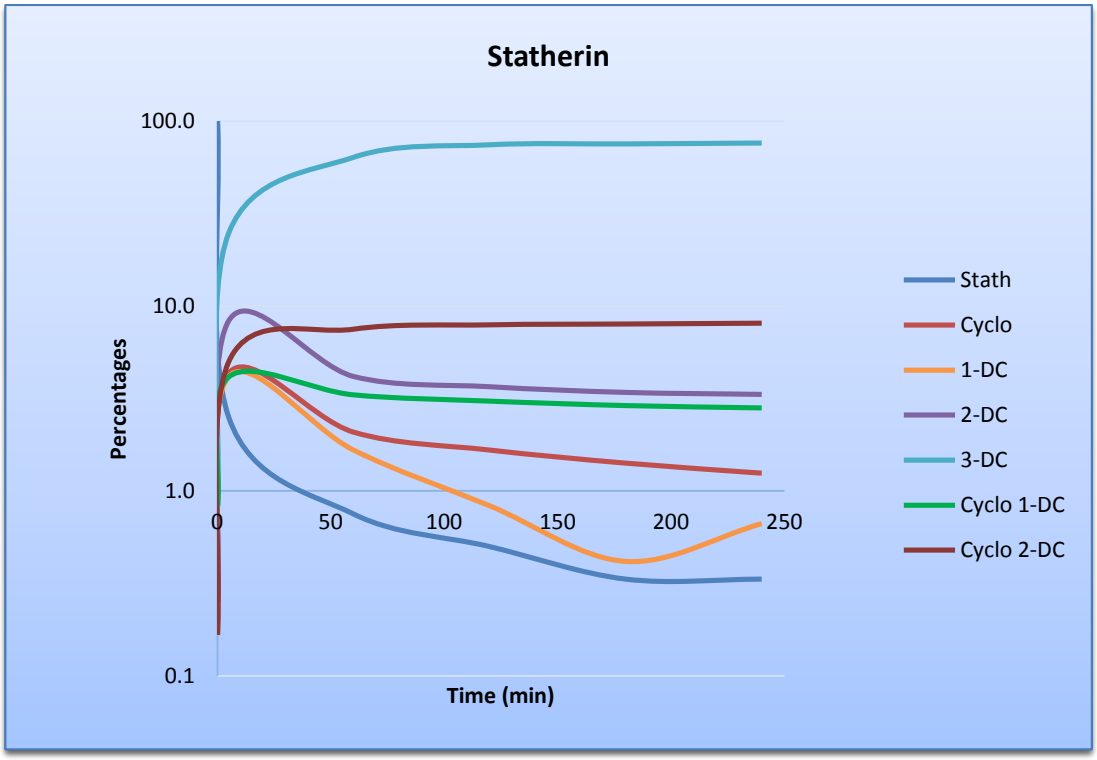


Figure 2.20 Variations of the percentages of the derivatives of statherin as a function of time (min) during the incubation with TG-2 and DC (37 °C).

2.3.6 Reaction of purified peptides with TG-2 in the presence of benzoyl-glutamine-glycine (BQG).

All the peptides under study were incubated (37 °C) with TG-2 (guinea pig) in the presence of benzoyl-glutamine-glycine (BQG) as substrate acceptor of the lone pair of the lysine residues present in the peptide sequence. Under the conditions utilized none product of reaction with BQG was observed for all the peptides and the kinetics of formation of the cyclo-derivatives was comparable to those observed with TG-2 alone.

2.3.7 Reaction of purified peptides with TG-2 alone at different temperatures.

In order to investigate the effect of temperatures, incubations with TG-2 alone were carried out at 25 °C and 45°C only for P-H, II-2 and P-C peptides. Results are summarized in Table 2.4, where the *t(1/2)* and *%max* obtained by best fitting of the percentage of cyclo-derivatives generated by the action of TG-2 by equation 1 are shown. The values of *t(1/2)* and *%max* clearly indicated that the physiological temperature (37 °C) is the best for the reaction. Similar results were obtained at different temperatures incubating the peptides at different temperature with TG-2 in the presence of DC. The analyses of the MSMS of the various DC-derivatives showed that changes of temperatures did not modify the specificity of the reaction.

Table 2.5 Comparison between values of %max and t(1/2) at different temperatures (37 °C, 25°C and 45°C) for P-H, II-2 and P-C pepyides

Peptide Name	*t(1/2) (min)*	*%max*	*t(1/2 (min)*	*%max*	*t(1/2) (min)*	*%max*
	37 °C		25 °C		45 °C	
cyclo P-H	4	15	22	10	10	10
cyclo II-2	2	5	2	4	1	3
cyclo-P-C	16	49	38	58	10	30

2.3.8 Reaction of purified peptides with human TG-2.

In order to verify if differences could be observed using the more expensive human TG-2, instead of guinea pig TG-2, several peptides (P-C, II-2 and statherin) were incubated with human TG-2 alone and in the presence of DC and BQG. Except a lower specific activity of the human enzyme than the guinea pig enzyme, no other significant differences were observed, suggesting that the guinea pig TG-2 can be used as a good substitute of human TG-2.

2.4 Discussion

As reported in the introduction, purpose of this second section of the PhD thesis was to verify if several peptides belonging to the family of bPRPs are substrate of TG-2 and therefore potential components of the protein pellicle covering the oral mucosa. This protein network has been hypothesized essential for the protection and the health of the oral mucosa (Presland RB and Dale BA. 2000). This hypothesis derived at first from the observation of the weakness of the mucosal oral epithelia in patients with primary Sjögren syndrome, which can derive from the absence of salivary proteins and peptides, characteristic of this pathology (Mathews SA, et al. 2008). Various studies (Bradway SD, et al. 1992; Yao Y, et al. 1999; Cabras T, et al. 2006; Gibbins HL, et al. 2014; Blotnick E, et al. 2017) have shown that staterin, histatins and aPRPs are substrates of TG-2 and it was demonstrated by immunohistochemistry that some of these peptides are linked to the oral mucosal epithelia (Schüpbach P, et al. 2001), confirming their contribution to the formation of the protein pellicle. The study carried out in this thesis showed that some of the bPRPs tested were substrates of TG-2 and are therefore potential components of the oral mucosal pellicle. Unfortunately, probably due to their very repetitive and similar structure, antibodies against human salivary bPRPs are not available in the market, and therefore it was not possible to evidence directly their presence as covalent components of the mucosal epithelia until now. In this context, it is relevant to remark the very different response of the oral mucosa to injuries and wound with respect the normal skin. Recently interesting wound healing properties have been demonstrated for histatin 1 (Oudhoff MJ, et al. 2009; Khurshid Z, et al. 2017; Shah D, et al. 2017). Nonetheless, probably other salivary components play relevant roles in this specific mucosal response, and bPRPs can be among the contributors.

Surprisingly, despite their very similar sequences, the bPRPs studied disclosed different response towards TG-2 reaction. Indeed, the two P-D isoforms and P-H were good substrates for TG-2, the reactivity of II-2 was good enough, while P-F and P-J displayed very low reactivity. The reactivity of P-C was remarkable too, but P-C, as a fragment released by the cleavage of aPRPs cannot be considered an authentic bPRPs. Nonetheless, its reactivity suggests its potential participation to the protein pellicle too. Interesting was the better reactivity of P-D P_{32} when compared with that of P-D

A_{32}, indicating that even small structural differences can generate a different response. In this case the more rigid P-D P_{32} peptide was better recognized as TG-2 substrate.

Mapping of glutamine residues involved in the reaction, investigated by MSMS fragmentation of several multi-charged ions of DC- and cyclo-derivatives, demonstrated the high specificity of the TG-2 recognition. The evaluation of the sequences surrounding the reactive residues of P-D, P-H and II-2 suggested that the principal consensus sequence of bPRPs for TG-2 recognition is GNPQ. This consensus sequence is not present in P-F and P-J peptides and, likely for this reason, they were less reactive substrates of TG-2. In this thesis, the bPRPs studied were selected on the basis of their dimension, choosing the smaller in the lists reported in section 1. Bigger bPRPs are difficult to purify and to investigate even with high resolution mass spectrometry apparatus, and others (i.e. P-E) were rarely detectable in human saliva. However, from the sequences reported in the Tables of Section 1 it is possible to verify that the GNPQ sequence is present in P-Ko (res. 77-80), IB-6 (res. 88-91), Ps-1 (res. 210-213), Ps-2 (res. 271-274), IB-1 (res. 18-21), while it is not present- in P-E and IB8a Con 1^- and Con 1^+. Among the gPRPs, the GNPQ sequence is present in all the Gl isoforms coded by the *PRB3* locus (i.e. Gl-1, Gl-2 and Gl-3) and it is absent in all the gPRPs coded by the *PRB4* locus (i.e. Glicosyl. Protein A, II-1 and Cd-IIg). Therefore, the bPRPs and gPRPs with the GNPQ residues in their sequences could be satisfactory substrates for TG-2 and are potential components of the mucosal protein pellicle.

Reactivity of P-C followed other recognition rules (Esposito C and Caputo I. 2005). Mapping of the glutamines of statherin recognized by TG-2 confirmed, as demonstrated by Cabras et al. (Cabras T, et al. 2006) that Q_{37} is the more reactive residue and thus it is the main glutamine responsible for the formation of cyclo-statherin. However, also Q_{39} and Q_{40} were reactive, even though less than Q_{37}. The MS/MS data on the different DC-derivatives suggested that the reaction of DC was hierarchical, i.e Q_{37} is the first glutamine recognized, followed by Q_{39} and Q_{40}. The incubation in the presence of DC inhibited partly the cycle formation. However, the results obtained in this thesis and the detection of small amount of cyclo-statherin Q_{37} in human saliva (Cabras T, et al. 2006) strongly suggest that the formation of the cyclo-statherin Q_{37} has biological significance *in vivo* and that the other residues recognizable by TG-2 could have a role in the slow maturation of the oral protein pellicle.

Strange is the absence of reactivity of benzoyl-glutamine-glycine (BQG) as acceptor of lysine lone pair, which made difficult the mapping of several lysine residues involved in TG-2 recognition. In the hypothesis that BQG was not recognized by guinea pig TG-2 the incubation was also investigated with recombinant human TG-2, but with identical results. Nonetheless, no matter for the involvement of the lysine residues of the peptides studied, because all the peptides generated cyclo-derivatives. The use of human TG-2 provided results comparable with those obtained by guinea pig TG-2, suggesting that the latter is a good substitute for the human enzyme. The experiments at different temperature showed that 37 °C is the best among those investigated. This result stimulates hypothesis on potential molecular differences in the formation of the oral protein pellicle when the oral cavity is stressed by food and environment at very low or very high temperature.

In conclusion, even though this study has not allowed to clarify definitively the involvement of bPRPs in the formation of the oral protein pellicle, it strongly suggests the probable participation of some of them in its formation and results obtained represent an interesting stimulus for future investigations on the role of bPRPs in the protection of the oral cavity.

References

Agnihotri N, Kumar S, Mehta K. Tissue transglutaminase as a central mediator in inflammation-induced progression of breast cancer. *Breast Cancer Res*. **2013**, 15(1):202

Ahvazi, B.; Boeshans, KM.; Idler, W.; Baxa, U.; Steinert, PM. Roles of Calcium Ions in the Activation and Activity of the Transglutaminase 3 Enzyme. *J. Biol. Chem*. **2003**, 27;278(26):23834-41

Akbar, A.; McNeil, NMR.; Albert, MR.; Ta, V.; Adhikary, G.; Bourgeois, K.; Eckert, RL.; Keillor, JW. Structure-Activity Relationships of Potent, Targeted Covalent Inhibitors That Abolish Both the Transamidation and GTP Binding Activities of Human Tissue Transglutaminase. *J Med Chem*. **2017**; 28;60(18):7910-7927.

Azen, E.A. Genetics of salivary protein polymorphisms. *Crit. Rev. Oral Biol. Med.* **1993,** *4*, 479–485

Azen, E.A.; Amberger, E.; Fisher, S.; Prakobphol, A.; Niece, R.L. PRB1, PRB2, and PRB4 coded polymorphisms among human salivary concanavalin-A binding, II-1, and Po proline-rich proteins. *Am. J. Hum. Genet*. **1996**, *58(1)*, 143–153

Azen, E.A.; Goodman, P.A.; Lalley, P.A. Human salivary proline-rich protein genes on chromosome 12. *Am. J. Hum. Genet*. **1985,** *37*, 418–424

Azen, E.A.; Hurley, C.K.; Denniston, C. Genetic polymorphism of the major parotid salivary glycoprotein (Gl) with linkage to the genes for Pr, Db, and Pa. *Biochem. Genet.* **1979,** *17*, 257–279

Azen, E.A.; Minaguchi, K.; Latreille, P.; Kim. H.S. Alleles at the PRB3 locus coding for a disulfide-bonded human salivary proline-rich glycoprotein (Gl 8) and a null in an Ashkenazi Jew. *Am. J. Hum. Genet.* **1990,** *47*, 686–697

Blotnick, E.; Sol, A.;, Bachrach, G.; Muhlrad, A. Interactions of histatin-3 and histatin-5 with actin. *BMC Biochem*. **2017**,18(1):3

Bobek, L.A.; Levine, M.J. Cystatins--inhibitors of cysteine proteinases. *Crit Rev Oral Biol Med*. **1992,** 3(4):307-32.

Bobek, L.A.; Liu, J.; Sait, S.N.; Shows, T.B.; Bobek, Y.A.; Levine, M.J. Structure and chromosomal localization of the human salivary mucin gene, MUC7. *Genomics*. **1996**, 31(3):277– 282

Bradway, S D.; Bergey, E J.; Jones, P C.; Levine, M J. Oral mucosal pellicle. Adsorption and transpeptidation of salivary components to buccal epithelial cells. *Biochem J*. **1989**, 261(3):887–896

Bradway, S D.; Bergey, E J.; Scannapieco, F A.; Ramasubbu, N.; Zawacki, S.; Levine, M J. Formation of salivary-mucosal pellicle: the role of transglutaminase. *Biochem J*. **1992**, 284(2):557-564

Cabras, T.; Boi, R.; Pisano, E.; Iavarone, F.; Fanali, C.; Nemolato, S.; Faa, G.; Castagnola, M.; Messana, I. HPLC–ESI–MS and MSMS structural characterization of multifucosylated N-glycoforms of the basic proline-rich protein IB-8a CON1^{+} in human saliva. *J. Sep. Sci.* **2012a,** *35,* 1079–1086

Cabras, T.; Inzitari, R.; Fanali, C.; Scarano, E.; Patamia, M.; Sanna, MT.; Pisano, E.; Giardina, B.; Castagnola, M.; Messana, I. HPLC-MS characterization of cyclo-statherin Q-37, a specific cyclization product of human salivary statherin generated by transglutaminase 2. *J Sep Sci*. **2006,** 29(17):2600-8

Cabras, T.; Melis, M.; Castagnola, M.; Padiglia, A.; Tepper, B.J.; Messana, I.; Tomassini Barbarossa, I. Responsiveness to 6-n-propylthiouracil (PROP) is associated with salivary levels of two specific basic proline-rich proteins in humans. *PLoS One.* **2012b,** *7,* e30962

Cabras, T.; Pisano, E.; Boi, R.; Olianas, A.; Manconi, B.; Inzitari, R.; Fanali, C.; Giardina, B.; Castagnola, M.; Messana, I. Age-dependent modifications of the human salivary secretory protein complex. *J. Proteome Res*. **2009,** *8,* 4126–4134

Carlson, D.M. Proline-rich proteins and glycoproteins: expression of salivary gland multigene families. *Biochimie.* **1988**, *70,* 1689–1695

Carpenter, G.H.; Proctor, G.B. O-linked glycosylation occurs on basic parotid salivary proline-rich proteins. *Oral. Microbiol. Immunol.* **1999,** *14,* 309–315

Castagnola M, Cabras T, Iavarone F, Fanali C, Nemolato S, Peluso G, Bosello SL, Faa G, Ferraccioli G, Messana I. The human salivary proteome: a critical overview of the results obtained by different proteomic platforms. *Expert Rev Proteomics*. **2012a,** 9(1):33-46

Castagnola, M.; Cabras, T.; Iavarone, F.; Vincenzoni, F.; Vitali, A.; Pisano, E.; Nemolato, S.; Scarano, E.; Fiorita, A.; Vento, G.; Tirone, C.; Romagnoli, C.; Cordaro, M.; Paludetti, G.; Faa, G.; Messana, I. Top-down platform for deciphering the human salivary proteome. *J. Matern. Fetal Neonatal. Med.* **2012b,** *25,* 27–43

Chan, M.; Bennick, A. Proteolytic processing of a human salivary proline-rich protein precursor by proprotein convertases. *Eur. J. Biochem*. **2001,** *268,* 3423–3431

Clouthier, CM.; Mironov, GG.; Okhonin, V.; Berezovski, MV.; Keillor, JW. Real-time monitoring of protein conformational dynamics in solution using kinetic capillary electrophoresis. *Angew Chem Int Ed Engl*. **2012**, 7;51(50):12464-8. Epub 2012 Nov 6

Coughtrie, MW. Function and organization of the human cytosolic sulfotransferase (SULT) family. *Chem Biol Interact*. **2016**, 25;259 (Pt A):2-7

Dickinson, D.P. Cysteine peptidases of mammals: their biological roles and potential effects in the oral cavity and other tissues in health and disease. *Crit Rev Oral Biol Med*. **2002,** 13(3):238-75.

Diraimondo, TR.; Klöck, C.; Khosla, C. Interferon-γ activates transglutaminase 2 via a phosphatidylinositol-3-kinase-dependent pathway: implications for celiac sprue therapy. *J Pharmacol Exp Ther*. **2012**, 341(1):104-14. Epub 2012 Jan 6

Eckert RL, Kaartinen MT, Nurminskaya M, Belkin AM, Colak G, Johnson GV, Mehta K. Transglutaminase Regulation of Cell Function. *Physiol Rev*. **2014**, 94(2): 383–417.

Ekström, J.; Khosravani, N.; Castagnola, M.; Messana, I. Saliva and the Control of Its Secretion. Chapter in Med Radiol Diagn Imaging. **2017**, in press DOI 10.1007/174_2017_143, © Springer International Publishing AG

Esposito, C.; Caputo, I. Mammalian transglutaminases. Identification of substrates as a key to physiological function and physiopathological relevance. *FEBS J*. **2005**, 272(3):615-31

Gatta NG, Cammarota G, Gentile V .Possible roles of transglutaminases in molecular mechanisms responsible for human neurodegenerative diseases. *AIMS Biophysics*, **2016**, 3(4): 529-545.

Gatta NG, Romano R, Fioretti E, Gentile V. Transglutaminase inhibition: possible therapeutic mechanisms to protect cells from death in neurological disorders. *AIMS Molecular Science*, **2017**, 4(4): 399-414.

Gendler, S.J.; Spicer, A.P. Epithelial mucin genes. *Ann Rev Physiol*. **1995**, 57:607–634

Gibbins, H.L.; Yakubov, G.L.; Proctor, G.B.; Wilson, S.; Carpenter, G.H. What interactions drive the salivary mucosal pellicle formation? *Colloids Surf B Biointerfaces*. **2014**, 120(100): 184–192.

Gillece-Castro, B.L.; Prakobphol, A.; Burlingame, A.L.; Leffler, H.; Fisher, S.J. Structure and bacterial receptor activity of a human salivary proline-rich glycoprotein, *J. Biol. Chem.* **1991,** *266*, 17358–17368

Gusman, H.; Leone, C.; Helmerhorst, EJ.; Nunn, M.; Flora, B.; Troxler, RF.; Oppenheim, FG. Human salivary gland-specific daily variations in histatin concentrations determined by a novel quantitation technique. *Arch Oral Biol*. **2004**, 49(1):11-22.

Halgand, F.; Zabrouskov, V.; Bassilian, S.; Souda, P.; Loo, J.A.; Faull, K.F.; Wong, D.T.; Whitelegge, J.P. Defining intact protein primary structures from saliva: a step toward the human proteome project. *Anal. Chem.* **2012**, *84*, 4383–4395

Hannig, C.; Hannig, M.; Attin, T. Enzymes in the acquired enamel pellicle. *Eur J Oral Sci*. **2005**, 113(1):2-13

Hardt, M.; Thomas, L.R.; Dixon, S.E.; Newport, G.; Agabian, N.; Prakobphol, A.; Hall, S.C.; Witkowska, H.E.; Fisher, S.J. 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*, 2885–2899

Helmerhorst, E.J.; Sun, X.; Salih, E.; Oppenheim, F.G. Identification of Lys-Pro-Gln as a novel cleavage site specificity of saliva-associated proteases. *J. Biol. Chem.* **2008**, *283*, 19957–19966

Huelsz-Prince G, Belkin AM, VanBavel E, Bakker EN. Activation of extracellular transglutaminase 2 by mechanical force in the arterial wall. *J Vasc Res*. **2013**;50(5):383-95.

Huq, N.L.; Cross, K.J.; Ung, M.; Myroforidis, H.; Veith, P. D.; Chen, D.; Stanton, D.; He, H.; Ward, B.R.; Reynolds, E.C. A Review of the Salivary Proteome and Peptidome and Saliva-derived Peptide Therapeutics. *Int. J. Pept. Res. Ther*. **2007,** *13,* 547–564

Inzitari, R.; Vento, G.; Capoluongo, E.; Boccacci, S.; Fanali, C.; Cabras, T.; Romagnoli, C.; Giardina, B.; Messana, 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. Epub 2007 Mar 7

Jin, X.; Stamnaes, J.; Klöck, C.; DiRaimondo, TR.; Sollid, LM.; Khosla, C. Activation of extracellular transglutaminase 2 by thioredoxin. *J Biol Chem*. **2011**, 286(43):37866-73. Epub 2011 Sep 9

Kauffman, D.L.; Keller, P.J.; Bennick, A.; Blum, M. Alignment of amino acid and DNA sequences of human proline-rich proteins. *Crit. Rev. Oral Biol. Med.* **1993,** *4(3/4),* 287–292

Keillor, JW.; Apperley, KY.; Akbar, A. Inhibitors of tissue transglutaminase. *Trends Pharmacol Sci*. **2015**; 36(1):32-40

Klöck, C.; Diraimondo, TR.; Khosla, C. Role of transglutaminase 2 in celiac disease pathogenesis. *Semin Immunopathol*. **2012**, 34(4):513-22. Epub 2012 Mar 22

Khurshid, Z.; Najeeb, S.; Mali, M.; Moin, S.F., Raza, S.Q.; Zohaib, S.; Sefat, F.; Zafar, M.S. Histatin peptides: Pharmacological functions and their applications in dentistry. *Saudi Pharm J*. **2017**, 25(1):25-31

Lorand, L.; Graham, RM. Transglutaminases: crosslinking enzymes with pleiotropic functions. *Nat Rev Mol Cell Biol*. **2003**, 4(2):140-56

Lyons, K.M.; Azen, E.A.; Goodman, P.A.; Smithies, O. Many protein products from a few loci: Assignment of human salivary proline-rich proteins to specific loci. *Genetics* **1988a,** *120,* 255–265

Lyons, K.M.; Stein, J.H.; Smithies, O. Length polymorphisms in human proline-rich protein genes generated by intragenic unequal crossing over. *Genetics* **1988b**, *120,* 267–278

Macias, M.J.; Wiesner, S.; Sudol, M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett.* **2002,** *513,* 30–37

Maeda, N.; Kim, H.S.; Azen, E.A.; Smithies, O. Differential RNA splicing and post-translational cleavages in the human salivary proline-rich protein gene system. *J. Biol. Chem.* **1985,** *260,* 11123–11130

Mamula, P.W.; Heerema, N.A.; Palmer, C.G.; Lyons, K.M.; Karn, R.C. Localization of the human salivary protein complex (SPC) to chromosome band 12p13.2. *Cytogenet. Cell. Genet.* **1985**, *39,* 279–284

Manconi, B.; Cabras, T.; Sanna, M.; Piras, V.; Liori. B.; Pisano, E.; Iavarone, F.; Vincenzoni, F.; Cordaro, M.; Faa, G.; Castagnola, M.; Messana, I. N- and O-linked glycosylation site profiling of the human basic salivary proline-rich protein 3M. *J. Sep. Sci.* **2016a,** *39*, 987–997

Manconi, B.; Castagnola, M.; Cabras, T.; Olianas, A.; Vitali, A.; Desiderio, C.; Sanna, M.T.; Messana, I. The intriguing heterogeneity of human salivary proline-rich proteins: Short title: Salivary proline-rich protein species. *J. Proteomics*. **2016b**, *134*, 47–56

Mathews, SA.; Kurien, BT.; Scofield, RH. Oral manifestations of Sjögren's syndrome. J Dent Res. 2008; 87(4):308-18.

Messana, I.; Cabras, T.; Iavarone, F.; Vincenzoni, F.; Urbani, A.; Castagnola, M. Unraveling the different proteomic platforms. *J Sep Sci.* **2013**, 36(1):128-39

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

Messana, I.; Cabras, T.; Pisano, E.; Sanna, M.T.; Olianas, A.; Manconi, B.; Pellegrini. M.; Paludetti, G.; Scarano, E.; Fiorita, A.; Agostino, S.; Contucci, A.M.; Calò, L.; Picciotti, P.M.; 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. 2008a, 7, 911–926

Messana, 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.* **2008b**, *31*, 1948–1963

Minaguchi, K.; Takaesu, Y.; Tsutsumi, T.; Suzuki, K. Studies of genetic markers in human saliva. (VII). Frequencies of the major parotid salivary glycoprotein (GI) system in a Japanese population. *Bull. Tokyo Dent. Coll.* **1981**, *22*, 1–6

Nemes, Z.; Petrovski, G.; Fésüs, L. Tools for the detection and quantitation of protein transglutamination. *Anal Biochem*. **2005**, 342(1):1-10

Nikolov M, Schmidt C, Urlaub H. Quantitative mass spectrometry-based proteomics: an overview. *Methods Mol Biol.* **2012**;893:85-100

Offner, G.D.; Troxler, R.F. Heterogeneity of high-molecularweight human salivary mucins. *Adv Dent Res*. **2000**, 14:69–75

Oppenheim, F.G.; Salih, E.; Siqueira, W.L.; Zhang, W.; Helmerhorst, E.J. Salivary proteome and its genetic polymorphisms. *Ann. NY Acad. Sci.* **2007,** *1098*, 22–50

Oudhoff, M.J.; Kroeze, K.L.; Nazmi, K.; van den Keijbus, P.A.; van 't Hof, W.; Fernandez-Borja, M.; Hordijk, P.L.; Gibbs, S.; Bolscher, J.G., Veerman, E.C. Structure-activity

analysis of histatin, a potent wound healing peptide from human saliva: cyclization of histatin potentiates molar activity 1,000-fold. *FASEB J*. **2009**, 23(11):3928-35.

Padiglia, A.; Orrù, R.; Boroumand, M.; Olianas, A.; Manconi, B.; Sanna, M.T.; Desiderio, C.; Iavarone, F.; Liori, B.; Messana, I.; Castagnola, M.; Cabras, T. Extensive Characterization of the Human Salivary Basic Proline-Rich Protein Family by Top-Down Mass Spectrometry. *J Proteome Res*. **2018**, 17(9):3292-3307.

Palmerini, C.A.; Mazzoni, M.; Radicioni, G.; Marzano, V.; Granieri, L.; Iavarone, F.; Longhi, R.; Messana, I.; Cabras, T.; Sanna, M.T.; Castagnola, M.; Vitali, A. Antagonistic Effect of a Salivary Proline-Rich Peptide on the Cytosolic Ca^{2+} Mobilization Induced by Progesterone in Oral Squamous Cancer Cells. *PLoS One*. **2016,** *11*, e0147925

Parameswaran KN, Velasco PT, Wilson J, Lorand L. Labeling of epsilon-lysine crosslinking sites in proteins with peptide substrates of factor XIIIa and transglutaminase. *Proc Natl Acad Sci USA*. **1990**, 87(21):8472-8475.

Pastor, MT.; Diez, A.; Pérez-Payá, E.; Abad, C. Addressing substrate glutamine requirements for tissue transglutaminase using substance P analogues. *FEBS Lett*. **1999**, 451(3):231-4

Pinkas, DM.; Strop, P.; Brunger, AT.; Khosla, C. Transglutaminase 2 undergoes a large conformational change upon activation. *PLoS Biol*. **2007**, 5(12):e327

Piper, JL, Gray, GM, Khosla, C. High selectivity of human tissue transglutaminase for immunoactive gliadin peptides: implications for celiac sprue. Biochemistry **2002**, 41(1), 386-393.

Presland, R.B; Dale, B.A. Epithelial structural proteins of the skin and oral cavity: function in health and disease. *Crit Rev Oral Biol Med*. **2000**;11(4):383-408.

Robinson, R.; Kauffman, D.L.; Waye, M.M.; Blum, M.; Bennick, A.; Keller, P.J. Primary structure and possible origin of the non-glycosylated basic proline-rich protein of human submandibular/sublingual saliva. *Biochem. J.* **1989,** *263*, 497–503

Ruhl, S.; Sandberg, A.L.; Cisar, J.O. Salivary receptors for the proline-rich protein-binding and lectin-like adhesins of oral actinomyces and streptococci. *J. Dent. Res*. **2004,** *83*, 505–510

Ryan, C.M.; Souda, P.; Halgand, F.; Wong, D.T.; Loo, J.A.; Faull, K.F.; Whitelegge, J.P. Confident assignment of intact mass tags to human salivary cystatins using top-down Fouriertransform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom*. **2010**, 21(6):908–917

Sabatini, L.M.; Azen, E.A. Histatins, a family of salivary histidine-rich proteins, are encoded by at least two loci (HIS1 and HIS2). *Biochem Biophys Res Commun*. **1989**, 160(2):495-502.

Scannapieco, F.A.; Torres, G.; Levine, M.J. Salivary alphaamylase: role in dental plaque and caries formation. *Crit Rev Oral Biol Med*. **1993**, 4:301–307

Scherer, S.E.; Muzny, D.M.; Buhay, C.J.; Chen, R.; Cree, A.; et al. The finished DNA sequence of human Chromosome 12. *Nature* **2006,** *440*, 346–351

Schüpbach, P.; Oppenheim, F.G.; Lendenmann, U.; Lamkin, M.S.; Yao, Y.; Guggenheim, B.Electron-microscopic demonstration of proline-rich proteins, statherin, and histatins in acquired enamel pellicles in vitro. *Eur J Oral Sci*. **2001**, 109(1):60-8.

Schwartz, S.S; Hay, D.I.; Schluckebier, S.K. Inhibition of calcium phosphate precipitation by human salivary statherin: structure-activity relationships. *Calcif Tissue Int*. **1992,** 50(6):511-7.

Shah, D.; Ali, M.; Shukla, D.; Jain, S.; Aakalu, V.K. Effects of histatin-1 peptide on human corneal epithelial cells. *PLoS One*. **2017**, 12(5):e0178030

Siqueira, W.L.; Oppenheim, F.G. Small molecular weight proteins/peptides present in the in vivo formed human acquired enamel pellicle. *Arch. Oral. Biol.* **2009,** *54*, 437–444

Strous, G.J.; Dekker, J. Mucin-like glycoproteins. *Crit Rev Biochem Mol Biol*. **1992**, 27(1-2):57–92

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*, 753–770

Sun, X.; Salih, E.; Oppenheim, FG.; Helmerhorst, EJ. Kinetics of histatin proteolysis in whole saliva and the effect on bioactive domains with metal-binding, antifungal, and wound-healing properties. *FASEB J*. **2009**, 23(8):2691-701.

Tagliabracci, V.S.; Engel, J.L.; Wen, J.; Wiley, S.E.; Worby, C.A.; Kinch, L.N.; Xiao, J.; Grishin. N.V.; Dixon, J.E. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science.* **2012,** *336*, 1150–1153

Thomsson, K.A.; Prakobphol, A.; Leffler, H.; Reddy, M.S.; Levine, M.J.; Fisher, S.J.; Hansson, G.C. The salivary mucin MG1 (MUC5B) carries a repertoire of unique oligosaccharides that is large and diverse. *Glycobiology*. **2002**, 12(1):1–14

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*, 911–926

Vitorino, R.; Alves, R.; Barros, A.; Caseiro, A.; Ferreira, R.; Lobo, M.C.; Bastos, A.; Duarte, J.; Carvalho, D.; Santos, L.L.; Amado, F.L. Finding new posttranslational modifications in salivary proline-rich proteins. *Proteomics.* **2010**, *10*, 3732–3742

Vitorino, R.; Barros, A.; Caseiro, A.; Domingues, P.J.; Amado, D.F. Towards defining the whole salivary peptidome. *PROTEOMICS Clinical Applications* **2009,** *3*, 528–540

Vitorino, R.; Calheiros-Lobo, M.J.; Williams, J.; Ferrer-Correia, A.J.; Tomer, K.B.; Duarte, J.A.; Domingues, P.M.; Amado, F.M. Peptidomic analysis of human acquired enamel pellicle. *Biomed. Chromatogr.* **2007,** *21*, 1107–1117

Vizcaíno, J.A.; Csordas, A.; del-Toro, N.; Dianes, J.A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q.W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.* **2016,** *44*, D447–D456

Wang, Z.; Griffin, M. TG2, a novel extracellular protein with multiple functions. *Amino Acids*. **2012**, 42(2-3):939-49

Yao, Y.; Lamkin, MS.; Oppenheim, FG. Pellicle precursor proteins: acidic proline-rich proteins, statherin, and histatins, and their crosslinking reaction by oral transglutaminase. *J Dent Res*. **1999**, 78(11):1696-703

Yao, Y.; Lamkin, MS.; Oppenheim, FG. Pellicle precursor protein crosslinking characterization of an adduct between acidic proline-rich protein (PRP-1) and statherin generated by transglutaminase. *J Dent Res*. **2000**, 79(4):930-8

Zamakhchari, M.; Wei, G.; Dewhirst, F.; Lee, J.; Schuppan, D.; Oppenheim, F.G.; Helmerhorst, E.J. Identification of Rothia bacteria as gluten-degrading natural colonizers of the upper gastro-intestinal tract. *PLoS One*. **2011,** *6*, e24455

Zhang, Z.; Marshall, A.G. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass. Spectrom.* **1998,** *9*, 225–233

Supplemental Figures.

P-D P_{32}

SPPG**K**P**Q**GPP**QQ**EGN**K**P**Q**GPPPPG**K**$_{25}$P**Q**GPPP**P**$_{32}$GGNP**Q**$_{37}$**Q**P**Q**APPAG**K**P**Q**G PPPPP**Q**GGRPPRPA**Q**G**QQ**PP**Q**

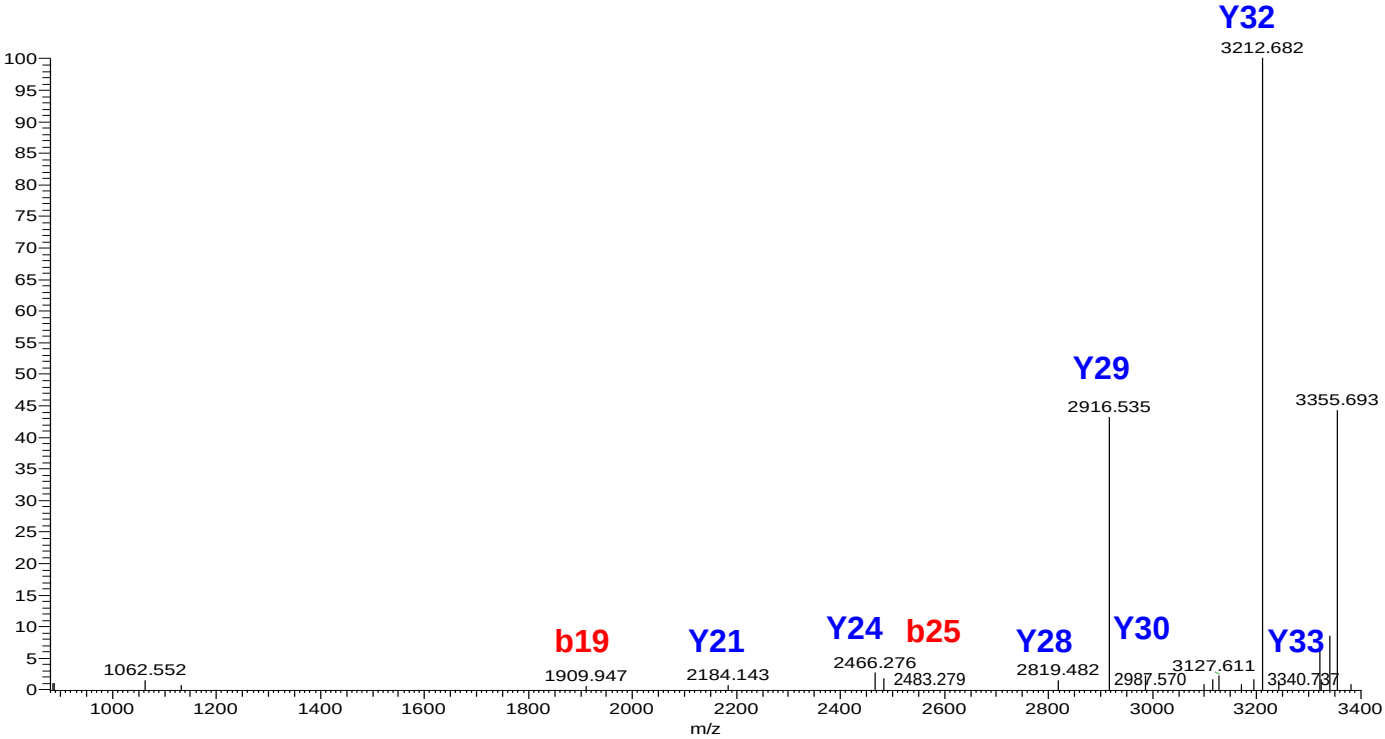


Figure S1 MSMS spectrum for cyclo P-D on $[M+7H]^{7+}$ 990.23 *m/z* (CID)

Q_{37} and K_{25} are the involved residue in the formation of the cycle.

(PD – NH_3 => 6946.55 - 17.03 = 6929.53).

Table S1 Theoretical fragments of cyclo P-D P_{32} (less NH3=17.03 after b_{37})

fragment number	*b*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.076	21	2104.06	2184.14	41	4014	4068.07	61	5980.07	6084.09
2	185.092	244.129	22	2201.11	2241.16	42	4111.05	4165.12	62	6077.12	6181.14
3	282.145	341.182	23	2298.16	2369.22	43	4208.11	4222.14	63	6148.16	6238.16
4	339.166	469.241	24	2355.18	2466.27	44	4279.14	4350.2	64	6276.21	6366.22
5	467.261	597.299	25	2483.28	2594.37	45	4336.16	4447.26	65	6333.24	6463.27
6	564.314	654.321	26	2580.33	2651.39	46	4464.26	4575.35	66	6461.29	6591.37
7	692.373	782.379	27	2708.39	2722.43	47	4561.31	4632.37	67	6589.35	6648.39
8	749.394	853.416	28	2765.41	2819.48	48	4689.37	4729.42	68	6686.41	6745.44
9	846.447	950.469	29	2862.46	2916.53	49	4746.39	4826.48	69	6783.46	6842.5
10	943.5	1106.57	30	2959.52	2987.57	50	4843.45	4923.53	70	---	---
11	1071.56	1203.62	31	3056.57	3115.63	51	4940.5	5020.58			
12	1199.62	1300.68	32	3153.62	3212.68	52	5037.55	5077.6			
13	1328.66	1456.78	33	3210.64	3340.74	53	5134.6	5205.66			
14	1385.68	1513.8	34	3267.67	3451.77	54	5231.66	5302.72			
15	1499.72	1570.82	35	3381.71	3548.83	55	5359.71	5430.81			
16	1627.82	1698.88	36	3478.76	3662.87	56	5416.74	5544.85			
17	1724.87	1795.93	37	3589.79	3719.89	57	5473.76	5601.88			
18	1852.93	1892.98	38	3717.85	3776.91	58	5629.86	5730.92			
19	1909.95	1990.04	39	3814.9	3873.96	59	5726.91	5858.98			
20	2007	2087.09	40	3942.96	3971.02	60	5823.96	5987.03			

P-D A_{32}

SPPG**K**P**Q**GPP**QQ**EGN**K**P**Q**GPPPPG**K**P**Q**GPPP**A_{32}**GGNP**Q_{37}**QP**Q**APPAG**K**P**Q**GP PPPP**Q**GGRPPRPA**Q**G**QQ**PP**Q**

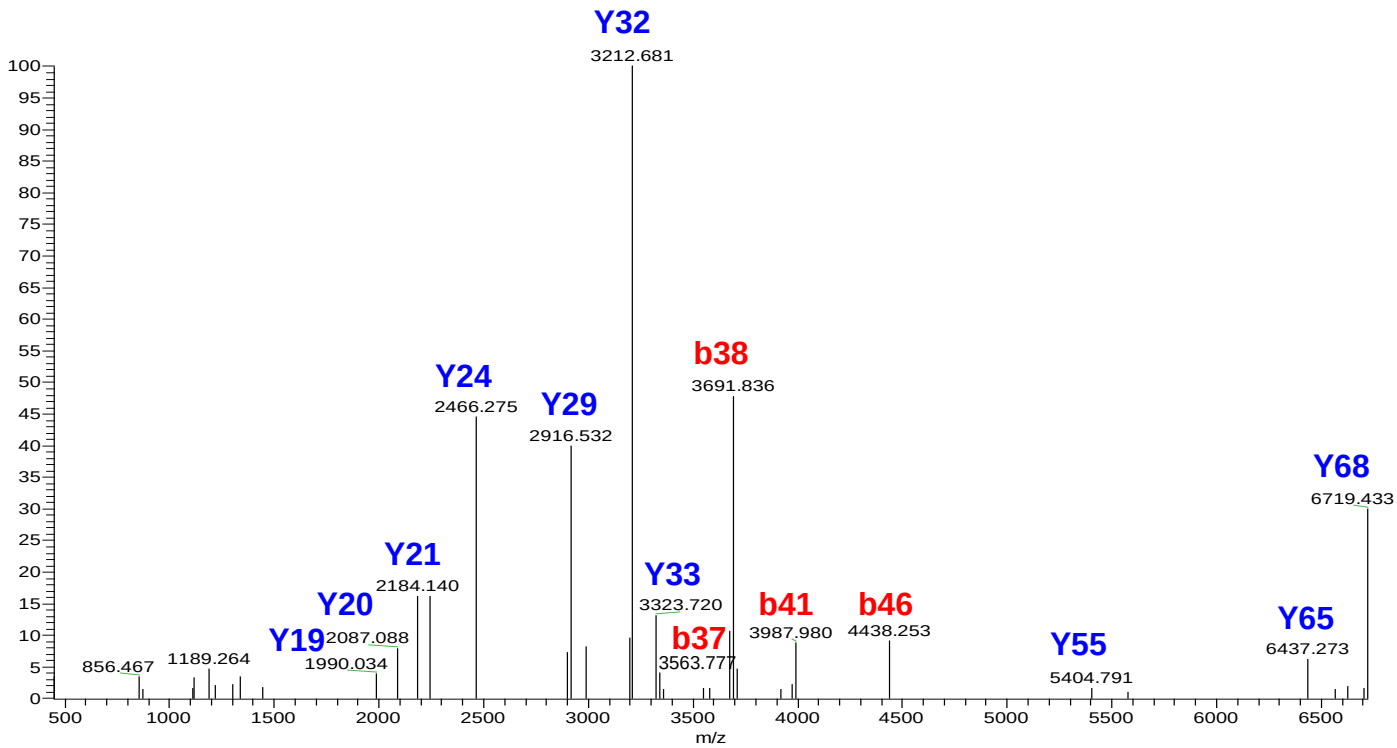


Figure S2 MSMS spectrum for cyclo P-D P_{32}→A on $[M+6H]^{6+}$ 1152.09 *m/z* (CID)

Q_{37} is the involved residue in the formation of the cycle.

(PD P_{32}→A – NH_3 => 6920.54 - 17.03 = 6903.54).

Table S2 Theoretical fragments of cyclo P-D A_{32} (less NH3=17.03 after b_{37})

fragment number	*b*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.08	21	2104.1	2184.1	41	3988	4042.1	61	5954	6058.1
2	185.09	244.13	22	2201.1	2241.2	42	4085	4139.1	62	6051.1	6155.1
3	282.14	341.18	23	2298.2	2369.2	43	4182.1	4196.1	63	6122.1	6212.1
4	339.17	469.24	24	2355.2	2466.3	44	4253.1	4324.2	64	6250.2	6340.2
5	467.26	597.3	25	2483.3	2594.4	45	4310.1	4421.2	65	6307.2	6437.3
6	564.31	654.32	26	2580.3	2651.4	46	4438.2	4549.3	66	6435.3	6565.4
7	692.37	782.38	27	2708.4	2722.4	47	4535.3	4606.4	67	6563.3	6622.4
8	749.39	853.42	28	2765.4	2819.5	48	4663.4	4703.4	68	6660.4	6719.4
9	846.45	950.47	29	2862.5	2916.5	49	4720.4	4800.5	69	6757.4	6816.5
10	943.5	1106.6	30	2959.5	2987.6	50	4817.4	4897.5	70	---	---
11	1071.6	1203.6	31	3056.6	3115.6	51	4914.5	4994.6			
12	1199.6	1300.7	32	3127.6	3212.7	52	5011.5	5051.6			
13	1328.7	1456.8	33	3184.6	3340.7	53	5108.6	5179.6			
14	1385.7	1513.8	34	3241.7	3451.8	54	5205.6	5276.7			
15	1499.7	1570.8	35	3355.7	3548.8	55	5333.7	5404.8			
16	1627.8	1698.9	36	3452.7	3662.9	56	5390.7	5518.8			
17	1724.9	1795.9	37	3563.8	3719.9	57	5447.7	5575.9			
18	1852.9	1893	38	3691.8	3776.9	58	5603.8	5704.9			
19	1910	1990	39	3788.9	3847.9	59	5700.9	5833			
20	2007	2087.1	40	3916.9	3945	60	5797.9	5961			

P-H peptide

SPPGK_5PQGPPQQEGNNPQGPPPPAGGNPQ_{29}QPQAPPAGQPQGPPRPPQGGR PSRPPQ

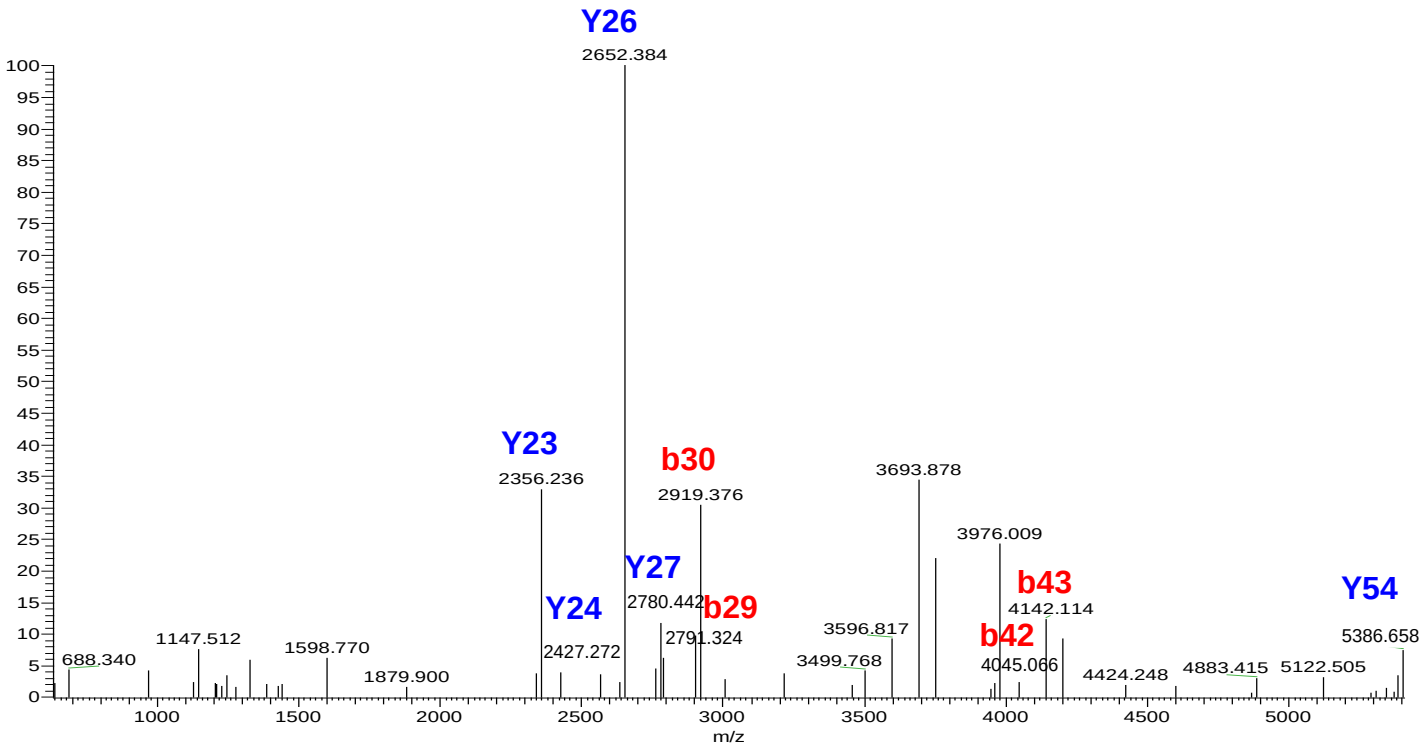


Figure S3 MSMS spectrum for cyclo P-H on $[M+5H]^{5+}$ 1115.76 *m/z* (CID)

The Q_{29} was the engaged residue in the formation of cross-link.

(PH – NH_3 => 5587.78 - 17.03 = 5570.75).

Table S3 Theoretical fragments of cyclo P-H (less NH3=17.03 after b_{29})

fragment number	*b*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	21	2090.005	2162.1326	41	3947.8807	4072.0397
2	185.0921	244.1292	22	2187.0578	2259.1854	42	4044.9335	4186.0826
3	282.1448	341.1819	23	2284.1105	2356.2381	43	4141.9863	4243.104
4	339.1663	497.2831	24	2355.1476	2427.2752	44	4298.0874	4372.1466
5	467.2613	584.3151	25	2412.1691	2555.3338	45	4395.1402	4500.2052
6	564.314	681.3678	26	2469.1906	2652.3866	46	4492.1929	4628.2638
7	692.3726	837.469	27	2583.2335	2780.4452	47	4620.2515	4725.3166
8	749.3941	894.4904	28	2680.2863	2891.4771	48	4677.273	4822.3693
9	846.4468	951.5119	29	2791.3182	2988.5299	49	4734.2944	4879.3908
10	943.4996	1079.5705	30	2919.3768	3102.5728	50	4890.3955	5007.4494
11	1071.5582	1176.6232	31	3016.4296	3159.5943	51	4987.4483	5104.5021
12	1199.6167	1273.676	32	3144.4881	3216.6157	52	5074.4803	5232.5971
13	1328.6593	1429.7771	33	3215.5253	3287.6529	53	5230.5814	5289.6185
14	1385.6808	1526.8299	34	3312.578	3384.7056	54	5327.6342	5386.6713
15	1499.7237	1623.8826	35	3409.6308	3481.7584	55	5424.687	5483.7241
16	1613.7667	1680.9041	36	3480.6679	3578.8112	56	---	---
17	1710.8194	1808.9627	37	3537.6894	3675.8639			
18	1838.878	1906.0154	38	3665.7479	3732.8854			
19	1895.8995	2034.074	39	3762.8007	3860.944			
20	1992.9522	2091.0955	40	3890.8593	3957.9967			

II-2 peptide

<QNLNEDVSQEESPSLIAGNPQ$_{21}$GPSPQGGNKPQGPPPPPGKPQGPPPQGGNK PQGPPPPGK PQGPPPQGDK SRSPR

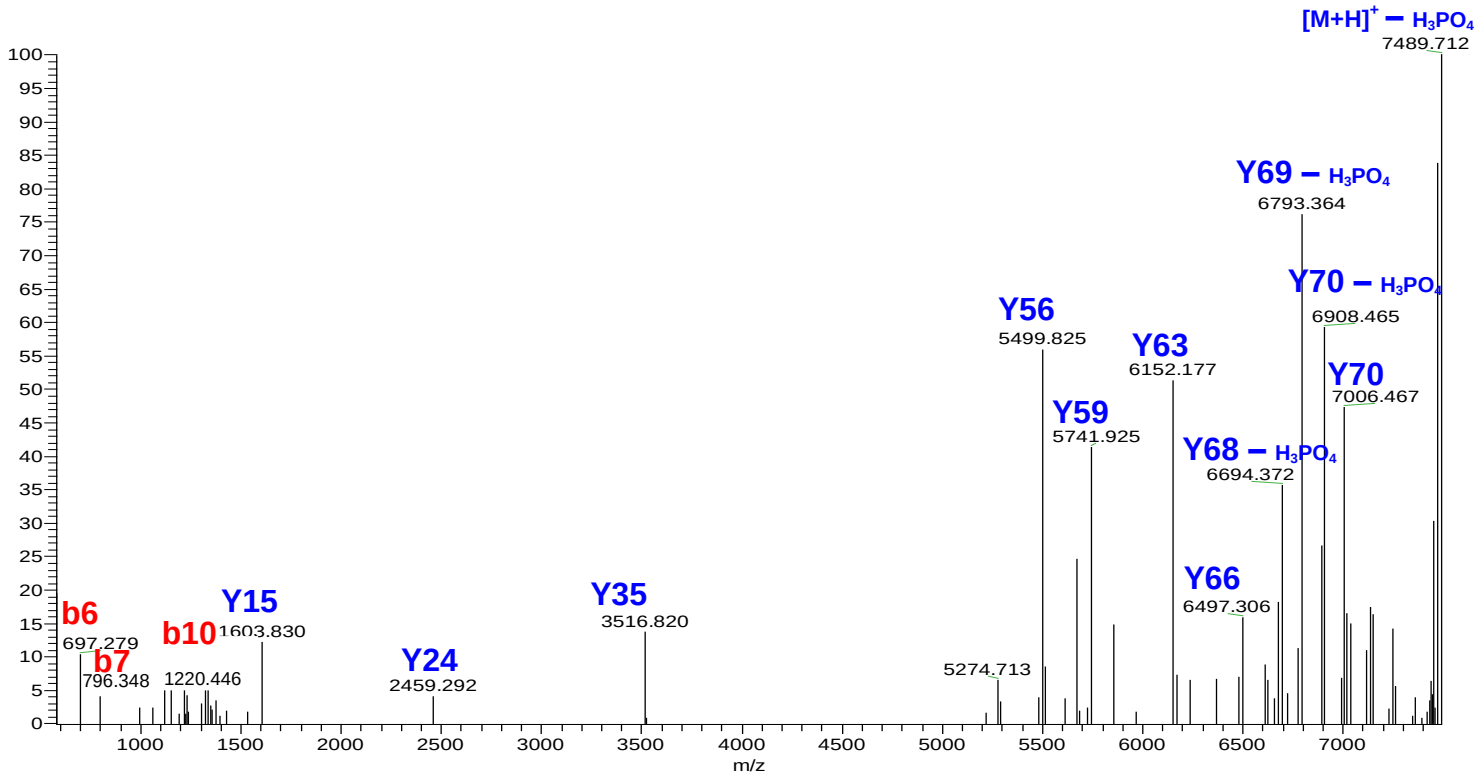


Figure S4 MSMS spectrum for cyclo II-2 on $[M+6H]^{6+}$ 1266.12 *m/z* (CID)

The Q_{21} was the engaged residue in the formation of cross-link.

(II-2 – NH_3 => 7604.69 - 17.03 = 7587.66).

Table S4 Theoretical fragments of cyclo II-2 (less NH3=17.03 after b_{21})

fragment number	*b-H_3PO_4*	*b*	*y*	*fragment number*	*b-H_3PO_4*	*b*	*y*	*y-H_3PO_4*
1	---	---	175.119	41	4070.9478	4168.9247	4090.1482	---
2	---	226.0822	272.1717	42	4199.0064	4296.9832	4187.201	---
3	---	339.1662	359.2037	43	4256.0278	4354.0047	4244.2224	---
4	---	453.2092	515.3049	44	4353.0806	4451.0575	4372.281	---
5	---	582.2518	602.3369	45	4450.1333	4548.1102	4469.3338	---
6	---	697.2787	730.4318	46	4547.1861	4645.163	4597.4287	---
7	---	796.3471	845.4588	47	4675.2447	4773.2216	4711.4716	---
8	865.3686	963.3455	902.4803	48	4732.2661	4830.243	4768.4931	---
9	993.4272	1091.4041	1030.5388	49	4789.2876	4887.2645	4825.5146	---
10	1122.4698	1220.4467	1127.5916	50	4903.3305	5001.3074	4953.5732	---
11	1251.5124	1349.4892	1224.6444	51	5031.4255	5129.4024	5050.6259	---
12	1338.5444	1436.5213	1321.6971	52	5128.4783	5226.4552	5137.6579	---
13	1435.5971	1533.574	1378.7186	53	5256.5368	5354.5137	5234.7107	---
14	1522.6292	1620.6061	1506.7772	54	5313.5583	5411.5352	5291.7322	---
15	1635.7132	1733.6901	1603.8299	55	5410.6111	5508.588	5402.7642	---
16	1748.7973	1846.7742	1731.9249	56	5507.6638	5605.6407	5499.8169	---
17	1819.8344	1917.8113	1788.9464	57	5604.7166	5702.6935	5613.8598	---
18	1876.8559	1974.8328	1885.9991	58	5701.7694	5799.7463	5670.8813	---
19	1990.8988	2088.8757	1983.0519	59	5758.7908	5856.7677	5741.9184	---
20	2087.9516	2185.9285	2080.1046	60	5886.8858	5984.8627	5855.0025	---
21	2198.9835	2296.9604	2177.1574	61	5983.9386	6081.9154	5968.0865	---
22	2256.005	2353.9819	2234.1789	62	6111.9971	6209.974	6055.1186	---
23	2353.0578	2451.0347	2362.2375	63	6169.0186	6266.9955	6152.1713	---
24	2440.0898	2538.0667	2459.2902	64	6266.0714	6364.0483	6239.2034	---
25	2537.1426	2635.1195	2587.3852	65	6363.1241	6461.101	6368.246	---
26	2665.2011	2763.178	2701.4281	66	6460.1769	6558.1538	6497.2886	---
27	2722.2226	2820.1995	2758.4496	67	6588.2355	6686.2124	6625.3471	---
28	2779.2441	2877.221	2815.471	68	6645.2569	6743.2338	6792.3455	6694.3686
29	2893.287	2991.2639	2943.5296	69	6760.2839	6858.2608	6891.4139	6793.437
30	3021.382	3119.3589	3040.5824	70	6888.3788	6986.3557	7006.4408	6908.464
31	3118.4347	3216.4116	3137.6351	71	6975.4109	7073.3878	7135.4834	7037.5065
32	3246.4933	3344.4702	3234.6879	72	7131.512	7229.4889	7249.5264	7151.5495
33	3303.5148	3401.4917	3291.7094	73	7218.544	7316.5209	7362.6104	7264.6335
34	3400.5675	3498.5444	3419.7679	74	7315.5968	7413.5737	7476.6534	7378.6765
35	3497.6203	3595.5972	3516.8207	75	---	---	---	---
36	3594.6731	3692.65	3644.9157					
37	3691.7258	3789.7027	3701.9371					
38	3788.7786	3886.7555	3798.9899					
39	3845.8	3943.7769	3896.0427					
40	3973.895	4071.8719	3993.0954					

P-C peptide

GRPQGPPQQGGHQQGPPPPPPGK$_{23}$PQGPPPQGGRPQGPPQGQ$_{41}$SPQ

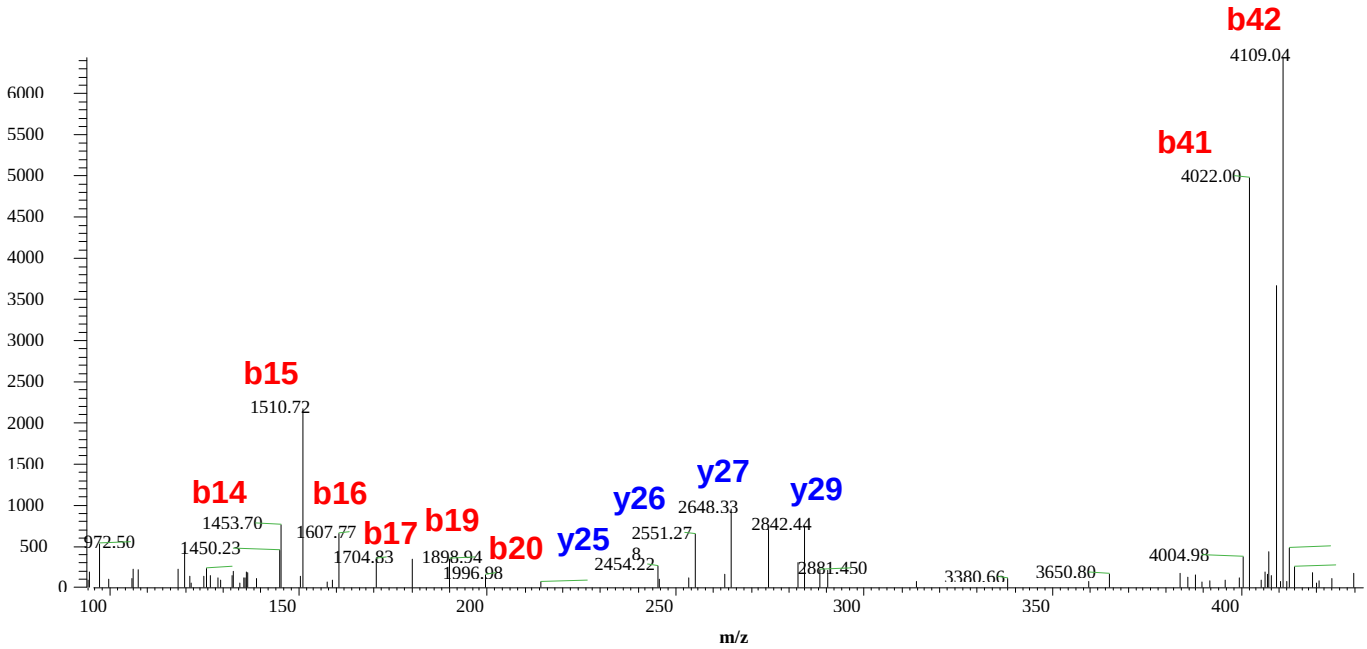


Figure S5 MSMS spectrum for cyclo P-C on $[M+4H]^{4+}$ 1089.29 *m/z* (CID).

Q41 is the involved residue in the formation of the cross-link.

(PC – NH_3 => 4369.18 - 17.03 = 4352.15).

Table S5 Theoretical fragments of cyclo P-C (less NH3=17.03 after b_{41})

fragment number	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	23	2278.1588	2260.1217
2	214.1299	244.1292	24	2375.2116	2357.1745
3	311.1826	331.1612	25	2503.2701	2454.2272
4	439.2412	442.1932	26	2560.2916	2551.28
5	496.2627	499.2147	27	2657.3444	2648.3328
6	593.3154	627.2732	28	2754.3971	2745.3855
7	690.3682	724.326	29	2851.4499	2842.4383
8	818.4268	821.3788	30	2979.5085	2899.4597
9	946.4853	878.4002	31	3036.5299	3027.5183
10	1003.5068	1006.4588	32	3093.5514	3155.5769
11	1060.5283	1103.5116	33	3249.6525	3292.6358
12	1197.5872	1259.6127	34	3346.7053	3349.6573
13	1325.6458	1316.6341	35	3474.7639	3406.6787
14	1453.7043	1373.6556	36	3531.7853	3534.7373
15	1510.7258	1501.7142	37	3628.8381	3662.7959
16	1607.7786	1598.7669	38	3725.8909	3759.8487
17	1704.8313	1695.8197	39	3853.9494	3856.9014
18	1801.8841	1792.8725	40	3910.9709	3913.9229
19	1898.9369	1849.8939	41	4022.0029	4041.9815
20	1995.9896	1977.9525	42	4109.0349	4139.0342
21	2093.0424	2075.0053	43	4206.0877	4295.1353
22	2150.0638	2203.1002	44	---	---

PD P_{32} + 1DC

SPPGKPQGPPQQEGNKPQGPPPPGK$_{25}$PQGPPPP_{32}GGNPQ$_{37}$QPQAPPAGKPQG PPPPPQGGRPPRPAQGQQPPQ

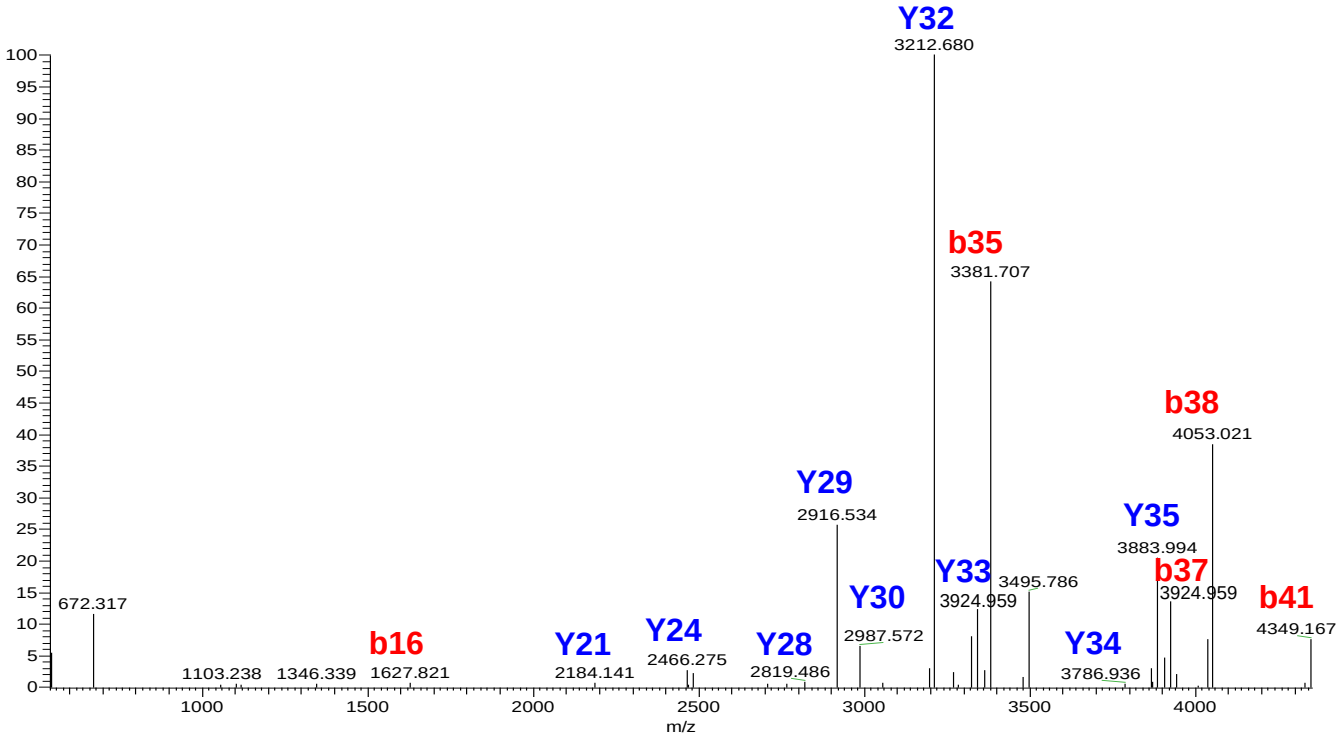


Figure S6 MSMS spectrum for P-D + 1DC on $[M+7H]^{+7}$ on 1039.53 *m/z* (CID)

Q_{37} is the principal acceptor for DC and Q_{40} can be for the second DC molecule.
(PD-peptide + DC – NH_3 => 6946.55 + 335.17 - 17.03 = 7264.69).

Table S6 Theoretical fragments of P-D P_{32} + 1DC (DC–NH3=318.14 after b_{37})

fragment number	*b*	*y*	*fragment number*	*B*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.076	19	1909.95	1990.04	37	3924.9602	4055.0569	55	5694.8815	5765.9775
2	185.092	244.129	20	2007	2087.09	38	4053.0188	4112.0784	56	5751.903	5880.0204
3	282.145	341.182	21	2104.06	2184.14	39	4150.0716	4209.1312	57	5808.9245	5937.0419
4	339.166	469.241	22	2201.11	2241.16	40	4278.1302	4306.1839	58	5965.0256	6066.0845
5	467.261	597.299	23	2298.16	2369.22	41	4349.1673	4403.2367	59	6062.0783	6194.1431
6	564.314	654.321	24	2355.18	2466.27	42	4446.22	4500.2895	60	6159.1311	6322.2016
7	692.373	782.379	25	2483.28	2594.37	43	4543.2728	4557.3109	61	6315.2322	6419.2544
8	749.394	853.416	26	2580.33	2651.39	44	4614.3099	4685.3695	62	6412.285	6516.3072
9	846.447	950.469	27	2708.39	2722.43	45	4671.3314	4782.4223	63	6483.3221	6573.3286
10	943.5	1106.57	28	2765.41	2819.48	46	4799.4263	4910.5172	64	6611.3807	6701.3872
11	1071.56	1203.62	29	2862.46	2916.53	47	4896.4791	4967.5387	65	6668.4021	6798.44
12	1199.62	1300.68	30	2959.52	2987.57	48	5024.5377	5064.5914	66	6796.4607	6926.5349
13	1328.66	1456.78	31	3056.57	3115.63	49	5081.5591	5161.6442	67	6924.5193	6983.5564
14	1385.68	1513.8	32	3153.62	3212.68	50	5178.6119	5258.697	68	7021.5721	7080.6092
15	1499.72	1570.82	33	3210.64	3340.74	51	5275.6647	5355.7497	69	7118.6248	7177.6619
16	1627.82	1698.88	34	3267.67	3786.9398	52	5372.7174	5412.7712	70	---	---
17	1724.87	1795.93	35	3381.71	3883.9925	53	5469.7702	5540.8298			
18	1852.93	1892.98	36	3478.76	3998.0355	54	5566.823	5637.8825			

PD A_{32} + 1DC

SPPGKPQGPPQQEGNKPQGPPPPGK_{25}PQGPPPA_{32}GGNPQ_{37}QPQAPPAGKPQ GPPPPPQGGRPPRPAQGQQPPQ

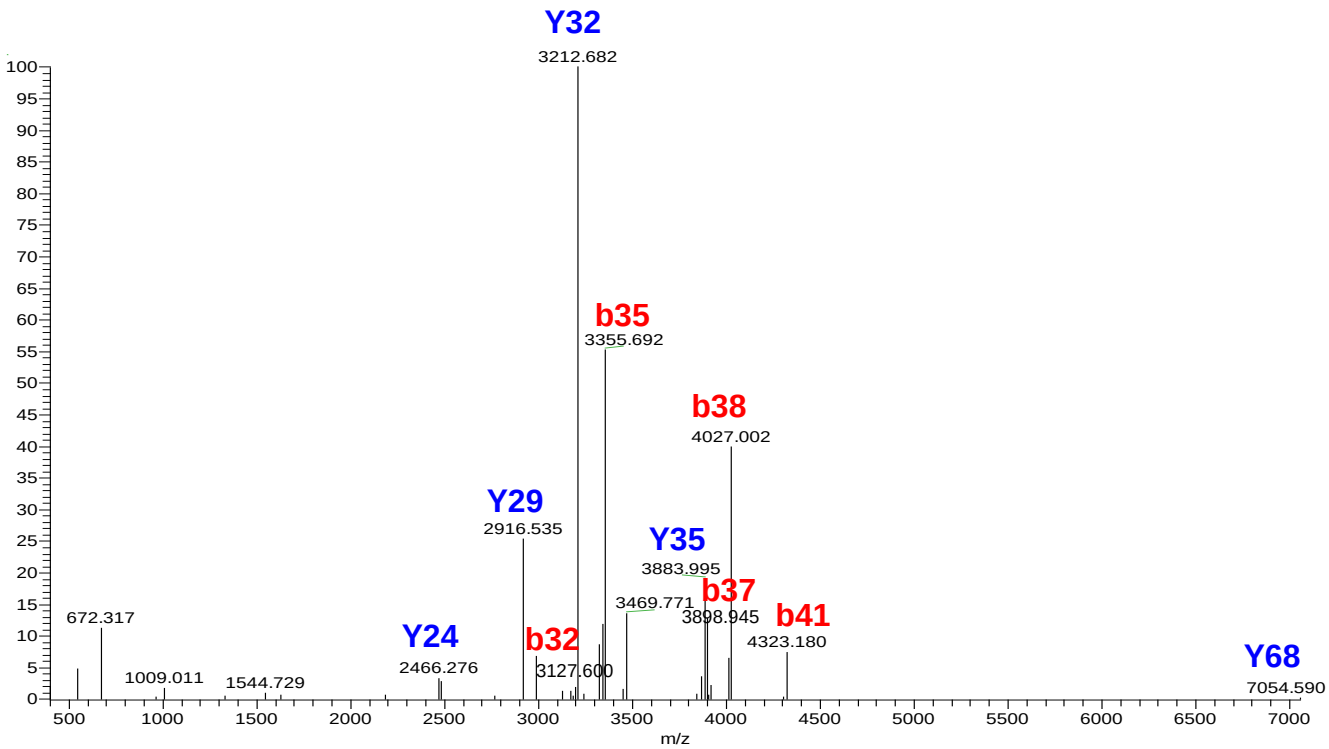


Figure S7 MSMS spectrum for P-D P_{32}→A+1DC on $[M+7H]^{7+}$ on 1035.82 *m/z* (CID). Q_{37} is the principal acceptor for DC molecule.
(PD P_{32}→A + DC – NH_3 => 6920.54 + 335.17 - 17.03 = 7238.68).

Table S7 Theoretical fragments of P-D A_{32} + 1DC (DC–NH3=318.14 after b_{37})

fragment number	*b*	*y*	*fragment number*	*B*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.076	19	1909.95	1990.04	37	3898.9446	4055.0569	55	5668.8659	5739.9618
2	185.092	244.129	20	2007	2087.09	38	4027.0032	4112.0784	56	5725.8874	5854.0047
3	282.145	341.182	21	2104.06	2184.14	39	4124.0559	4183.1155	57	5782.9088	5911.0262
4	339.166	469.241	22	2201.11	2241.16	40	4252.1145	4280.1682	58	5939.0099	6040.0688
5	467.261	597.299	23	2298.16	2369.22	41	4323.1516	4377.221	59	6036.0627	6168.1274
6	564.314	654.321	24	2355.18	2466.27	42	4420.2044	4474.2738	60	6133.1155	6296.1859
7	692.373	782.379	25	2483.28	2594.37	43	4517.2571	4531.2952	61	6289.2166	6393.2387
8	749.394	853.416	26	2580.33	2651.39	44	4588.2943	4659.3538	62	6386.2693	6490.2915
9	846.447	950.469	27	2708.39	2722.43	45	4645.3157	4756.4066	63	6457.3064	6547.3129
10	943.5	1106.57	28	2765.41	2819.48	46	4773.4107	4884.5015	64	6585.365	6675.3715
11	1071.56	1203.62	29	2862.46	2916.53	47	4870.4635	4941.523	65	6642.3865	6772.4243
12	1199.62	1300.68	30	2959.52	2987.57	48	4998.522	5038.5757	66	6770.4451	6900.5192
13	1328.66	1456.78	31	3056.57	3115.63	49	5055.5435	5135.6285	67	6898.5036	6957.5407
14	1385.68	1513.8	32	3127.6072	3212.68	50	5152.5963	5232.6813	68	6995.5564	7054.5935
15	1499.72	1570.82	33	3184.6287	3340.74	51	5249.649	5329.734	69	7092.6092	7151.6462
16	1627.82	1698.88	34	3241.6501	3786.9398	52	5346.7018	5386.7555	70	---	---
17	1724.87	1795.93	35	3355.693	3883.9925	53	5443.7545	5514.8141			
18	1852.93	1892.98	36	3452.7458	3998.0355	54	5540.8073	5611.8668			

PD P_{32} + 2DC

SPPGKPQGPPQQEGNKPQGPPPPGK$_{25}$PQGPPPP_{32}GGNPQ$_{37}$QPQ$_{40}$APPAGKPQ GPPPPPQGGRPPRPAQGQQPPQ

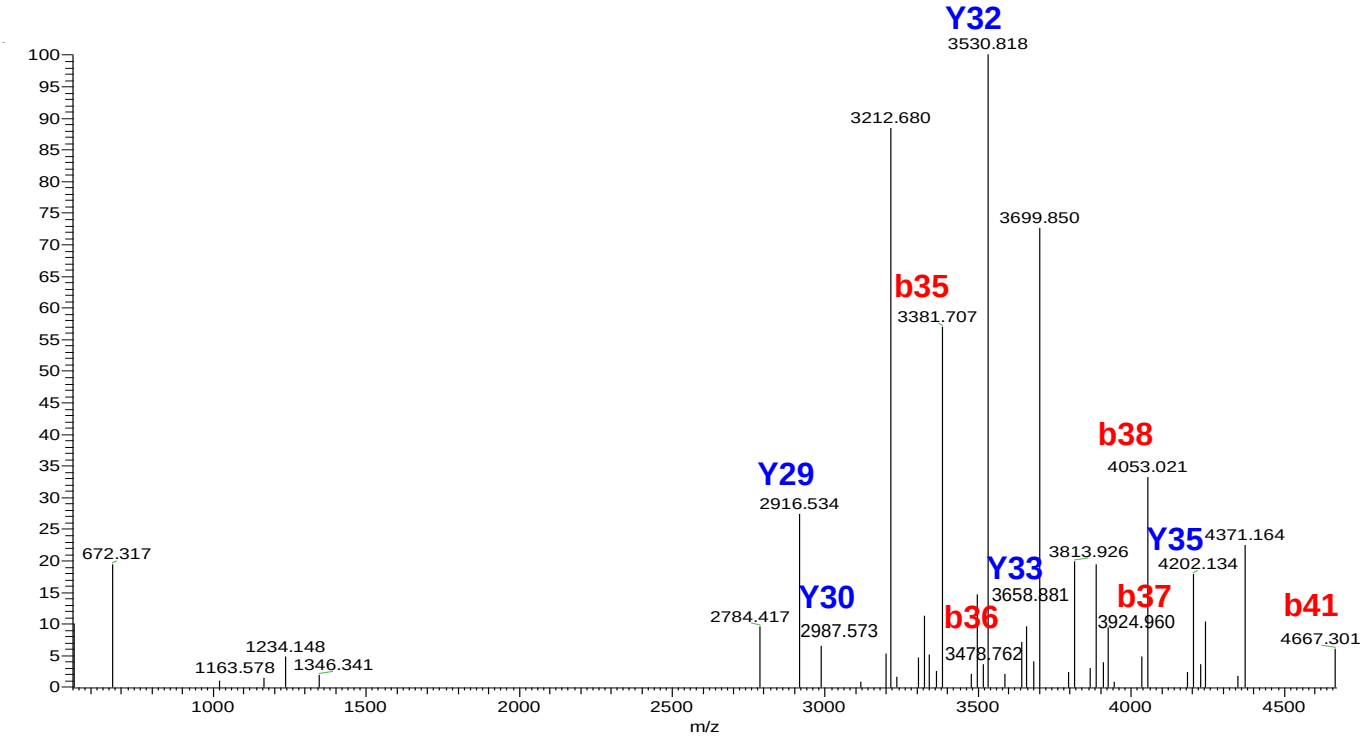


Figure S8 MSMS spectrum for P-D + 2DC on $[M+7H]^{7+}$ on 1084.7 *m/z* (CID)
Q_{37} is the principal acceptor for DC and Q_{40} can be for the second DC molecule.
(PD-peptide +2 DC – 2 NH_3 => 6946.55 + 670.34 - 34.06 = 7582.84)

Table S8 Theoretical fragments of P-D P32 + 2DC (DC–NH3=318.14 after b37 and DC–NH3=318.14 after b40)

fragment number	*b*	*y*	*fragment number*	*B*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.076	19	1909.95	1990.04	37	3924.9602	4373.1971	55	6013.0217	6084.1177
2	185.092	244.129	20	2007	2087.09	38	4053.0188	4430.2186	56	6070.0432	6198.1606
3	282.145	341.182	21	2104.06	2184.14	39	4150.0716	4527.2714	57	6127.0647	6255.1821
4	339.166	469.241	22	2201.11	2241.16	40	4596.2704	4624.3241	58	6283.1658	6384.2247
5	467.261	597.299	23	2298.16	2369.22	41	4667.3075	4721.3769	59	6380.2185	6512.2833
6	564.314	654.321	24	2355.18	2466.27	42	4764.3602	4818.4297	60	6477.2713	6640.3418
7	692.373	782.379	25	2483.28	2594.37	43	4861.413	4875.4511	61	6633.3724	6737.3946
8	749.394	853.416	26	2580.33	2651.39	44	4932.4501	5003.5097	62	6730.4252	6834.4474
9	846.447	950.469	27	2708.39	2722.43	45	4989.4716	5100.5625	63	6801.4623	6891.4688
10	943.5	1106.57	28	2765.41	2819.48	46	5117.5665	5228.6574	64	6929.5209	7019.5274
11	1071.56	1203.62	29	2862.46	2916.53	47	5214.6193	5285.6789	65	6986.5423	7116.5802
12	1199.62	1300.68	30	2959.52	2987.57	48	5342.6779	5382.7316	66	7114.6009	7244.6751
13	1328.66	1456.78	31	3056.57	3433.7702	49	5399.6993	5479.7844	67	7242.6595	7301.6966
14	1385.68	1513.8	32	3153.62	3530.8202	50	5496.7521	5576.8372	68	7339.7123	7398.7494
15	1499.72	1570.82	33	3210.64	3658.8802	51	5593.8049	5673.8899	69	7436.765	7495.8021
16	1627.82	1698.88	34	3267.67	4105.08	52	5690.8576	5730.9114	70	---	---
17	1724.87	1795.93	35	3381.71	4202.1327	53	5787.9104	5858.97			
18	1852.93	1892.98	36	3478.76	4316.1757	54	5884.9632	5956.0227			

P-H + 1DC

SPPGK_5PQGPPQ_{11}QEGNNPQGPPPPAGGNPQ_{29}QPQAPPAGQPQGPPRPPQGG RPSRPPQ

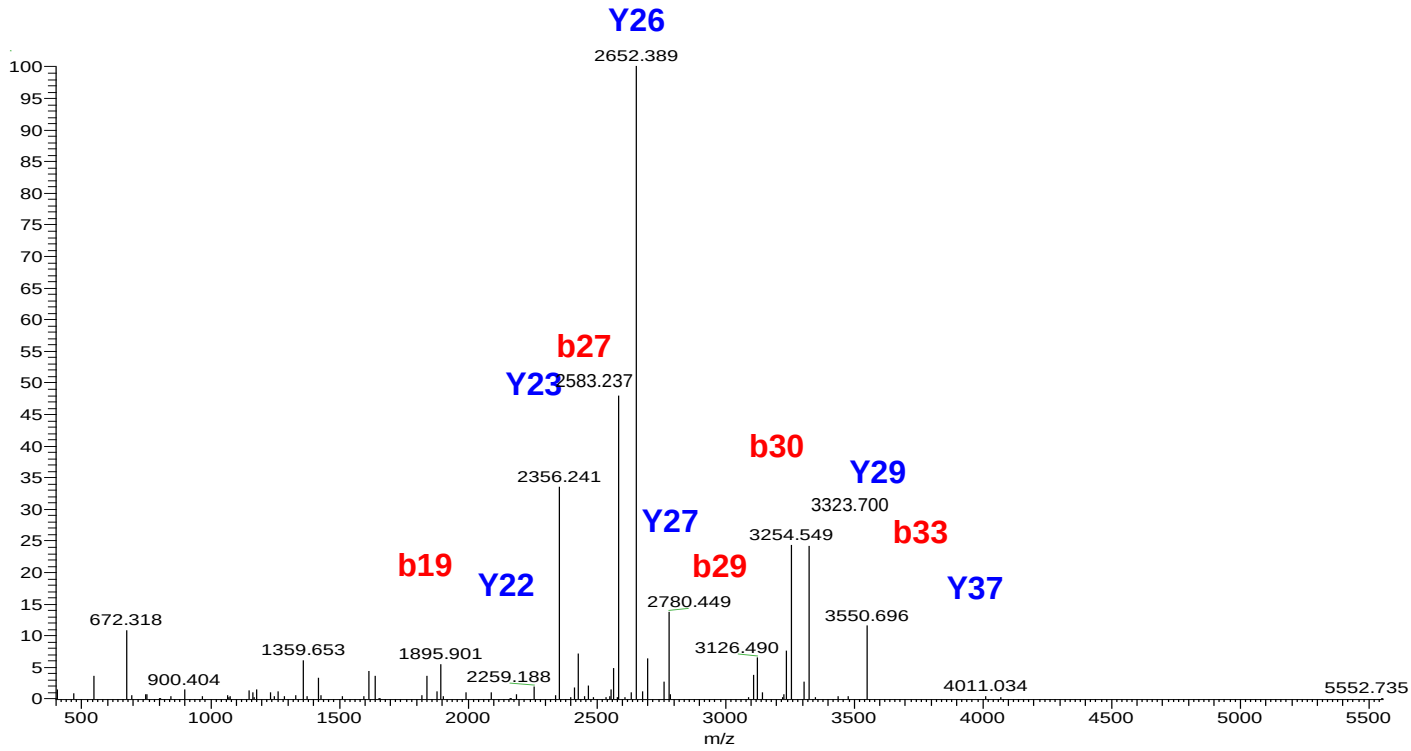


Figure S9 MSMS spectrum for P-H + 1DC on $[M+6H]^{6+}$ on 985.66 *m/z* (CID)

Q_{29} was established as a principal acceptor for DC.

(P-H peptide + DC – NH_3 => 5587.78 + 335.17 - 17.03 = 5905.92).

Table S9 Theoretical fragments of P-H + 1DC (DC–NH3=318.14 after b_{29})

fragment number	*b*	*y*	*fragment number*	*B*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	16	1613.7667	1680.9041	31	3351.5964	3494.7611	46	4827.3597	4963.4306
2	185.0921	244.1292	17	1710.8194	1808.9627	32	3479.6549	3551.7825	47	4955.4183	5060.4834
3	282.1448	341.1819	18	1838.878	1906.0154	33	3550.6921	3622.8197	48	5012.4398	5157.5361
4	339.1663	497.2831	19	1895.8995	2034.074	34	3647.7448	3719.8724	49	5069.4612	5214.5576
5	467.2613	584.3151	20	1992.9522	2091.0955	35	3744.7976	3816.9252	50	5225.5623	5342.6162
6	564.314	681.3678	21	2090.005	2162.1326	36	3815.8347	3913.978	51	5322.6151	5439.6689
7	692.3726	837.469	22	2187.0578	2259.1854	37	3872.8562	4011.0307	52	5409.6471	5567.7639
8	749.3941	894.4904	23	2284.1105	2356.2381	38	4000.9147	4068.0522	53	5565.7482	5624.7853
9	846.4468	951.5119	24	2355.1476	2427.2752	39	4097.9675	4196.1108	54	5662.801	5721.8381
10	943.4996	1079.5705	25	2412.1691	2555.3338	40	4226.0261	4293.1635	55	5759.8538	5818.8909
11	1071.5582	1176.6232	26	2469.1906	2652.3866	41	4283.0475	4407.2065	56	---	---
12	1199.6167	1273.676	27	2583.2335	2780.4452	42	4380.1003	4521.2494			
13	1328.6593	1429.7771	28	2680.2863	3226.6439	43	4477.1531	4578.2708			
14	1385.6808	1526.8299	29	3126.485	3323.6967	44	4633.2542	4707.3134			
15	1499.7237	1623.8826	30	3254.5436	3437.7396	45	4730.307	4835.372			

P-H + 2DC

SPPG**K**$_5$P**Q**GPP**Q**$_{11}$**Q**EGNNP**Q**GPPPPAGGNP**Q**$_{29}$**Q**P**Q**APPAG**Q**P**Q**GPPRPP**Q**GG RPSRPP**Q**

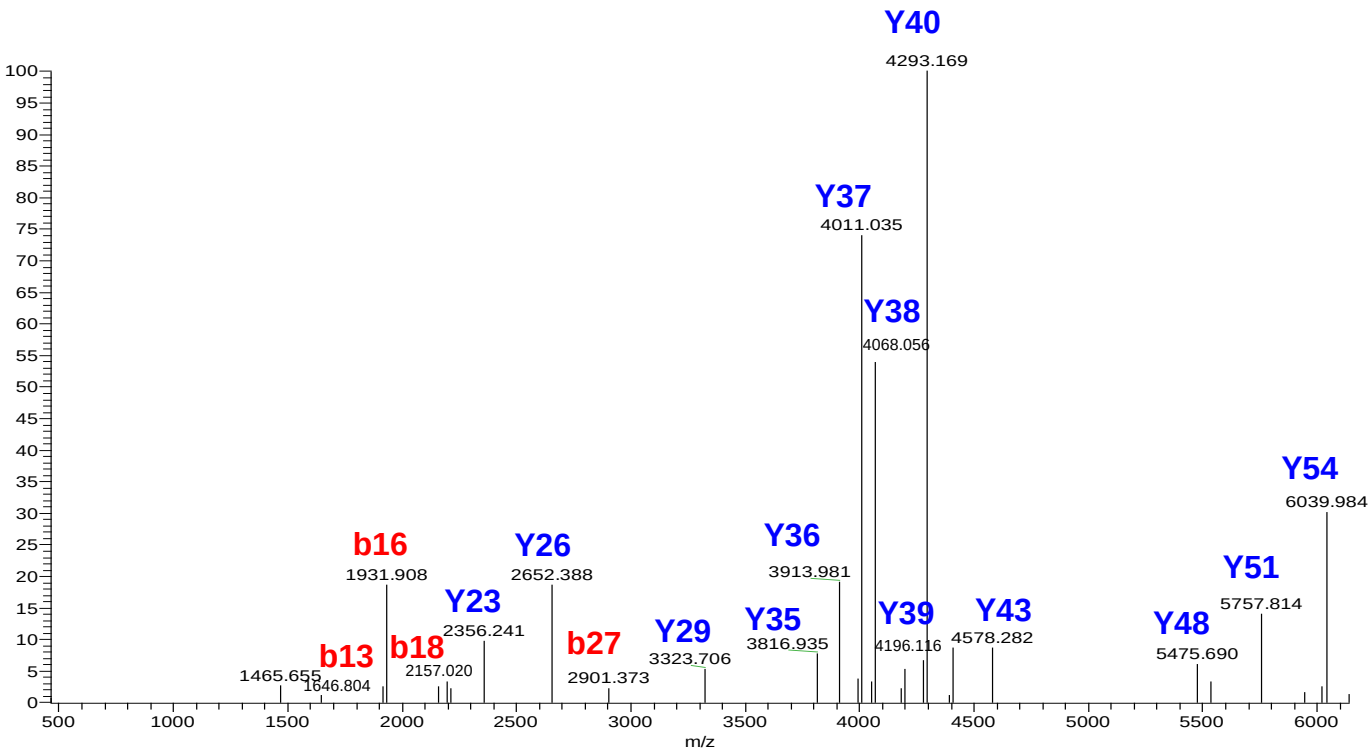


Figure S10 MSMS spectrum for P-H + 2DC on $[M+5H]^{5+}$ on 1246.42 *m/z* (CID) Q_{11} was the plausible acceptor for second DC.

(P-H peptide + 2 DC – 2 NH_3 => 5587.78 + 670.34 – 34.06 = 6224.06).

Table S10 Theoretical fragments of P-H + 2DC (DC–NH3=318.14 after b_{29} and DC–NH3=318.14 after b_{11})

fragment number	*b*	*y*	*fragment number*	*B*	*y*	*fragment number*	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	16	1931.9069	1680.9041	31	3669.7366	3494.7611	46	5145.4999	5281.5708
2	185.0921	244.1292	17	2028.9596	1808.9627	32	3797.7951	3551.7825	47	5273.5585	5378.6236
3	282.1448	341.1819	18	2157.0182	1906.0154	33	3868.8323	3622.8197	48	5330.58	5475.6763
4	339.1663	497.2831	19	2214.0397	2034.074	34	3965.885	3719.8724	49	5387.6014	5532.6978
5	467.2613	584.3151	20	2311.0924	2091.0955	35	4062.9378	3816.9252	50	5543.7025	5660.7564
6	564.314	681.3678	21	2408.1452	2162.1326	36	4133.9749	3913.978	51	5640.7553	5757.8091
7	692.3726	837.469	22	2505.198	2259.1854	37	4190.9964	4011.0307	52	5727.7873	5885.9041
8	749.3941	894.4904	23	2602.2507	2356.2381	38	4319.0549	4068.0522	53	5883.8884	5942.9255
9	846.4468	951.5119	24	2673.2878	2427.2752	39	4416.1077	4196.1108	54	5980.9412	6039.9783
10	943.4996	1079.5705	25	2730.3093	2555.3338	40	4544.1663	4293.1635	55	6077.994	6137.0311
11	1389.6984	1176.6232	26	2787.3308	2652.3866	41	4601.1877	4407.2065	56	---	---
12	1517.7569	1273.676	27	2901.3737	2780.4452	42	4698.2405	4521.2494			
13	1646.7995	1429.7771	28	2998.4265	3226.6439	43	4795.2933	4578.2708			
14	1703.821	1526.8299	29	3444.6252	3323.6967	44	4951.3944	4707.3134			
15	1817.8639	1623.8826	30	3572.6838	3437.7396	45	5048.4472	4835.372			

II-2 + 1DC

<**Q**NLNEDVS**Q**EESPSLIAGNP**Q_{21}**GPSP**Q**GGNKP**Q**GPPPPPGKP**Q**GPPP**Q**GGNK P**Q**GPPPPGK P**Q**GPPP**Q**GDK SRSPR

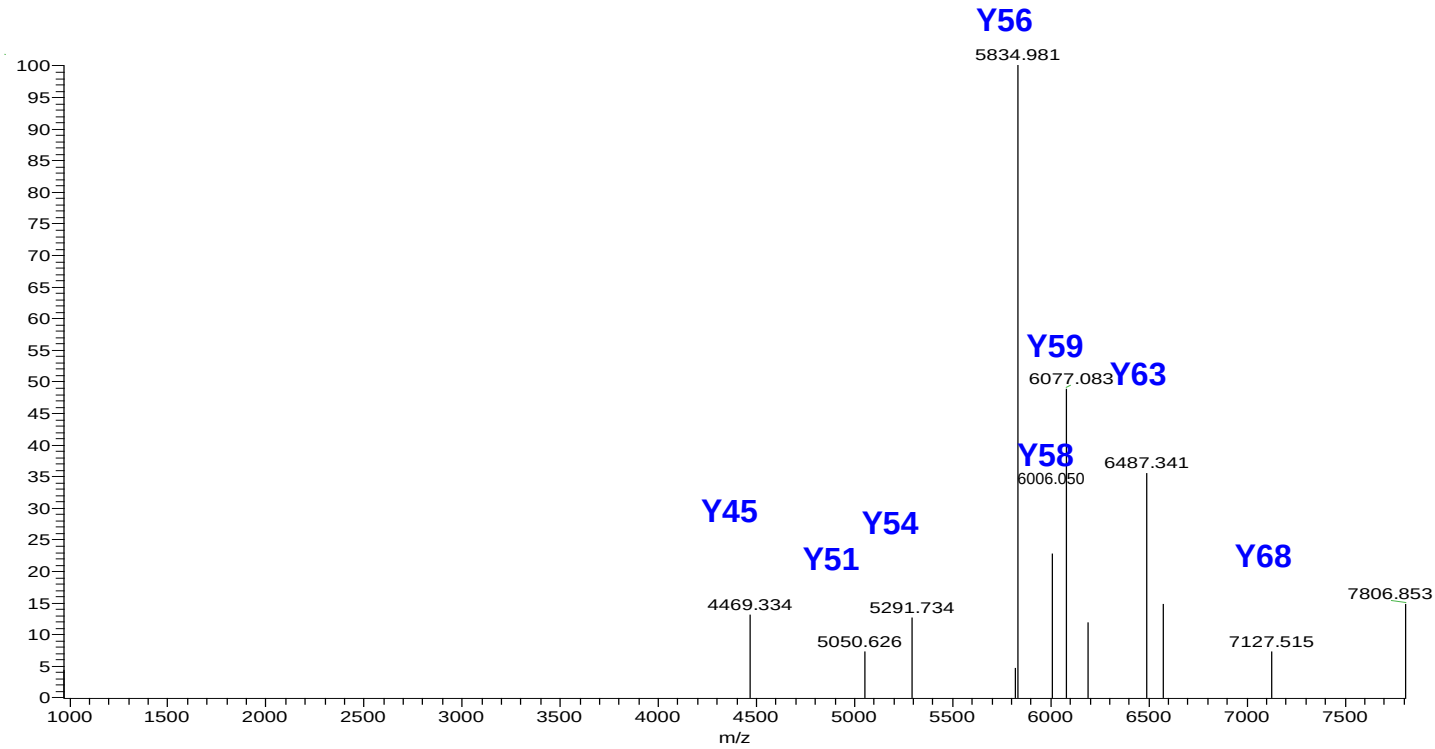


Figure S11 MSMS spectrum for II-2 + 1DC on $[M+7H]^{7+}$ on 1133.42 *m/z* (CID)

Q_{21} was the principal acceptor for DC.

(II-2 + DC – NH_3 => 7604.69 + 335.17 - 17.03 = 7922.83).

Table S11 Theoretical fragments of II-2 + 1DC (DC–NH3=318.14 after b_{21})

fragment number	*b-H_3PO_4*	*b*	*y*	*fragment number*	*b-H_3PO_4*	*b*	*y*	*y-H_3PO_4*
1	---	---	175.119	41	4406.1146	4504.0915	4090.1482	---
2	---	226.0822	272.1717	42	4534.1732	4632.15	4187.201	---
3	---	339.1662	359.2037	43	4591.1946	4689.1715	4244.2224	---
4	---	453.2092	515.3049	44	4688.2474	4786.2243	4372.281	---
5	---	582.2518	602.3369	45	4785.3001	4883.277	4469.3338	---
6	---	697.2787	730.4318	46	4882.3529	4980.3298	4597.4287	---
7	---	796.3471	845.4588	47	5010.4115	5108.3884	4711.4716	---
8	865.3686	963.3455	902.4803	48	5067.4329	5165.4098	4768.4931	---
9	993.4272	1091.4041	1030.5388	49	5124.4544	5222.4313	4825.5146	---
10	1122.4698	1220.4467	1127.5916	50	5238.4973	5336.4742	4953.5732	---
11	1251.5124	1349.4892	1224.6444	51	5366.5923	5464.5692	5050.6259	---
12	1338.5444	1436.5213	1321.6971	52	5463.6451	5561.622	5137.6579	---
13	1435.5971	1533.574	1378.7186	53	5591.7036	5689.6805	5234.7107	---
14	1522.6292	1620.6061	1506.7772	54	5648.7251	5746.702	5291.7322	---
15	1635.7132	1733.6901	1603.8299	55	5745.7779	5843.7548	5737.931	---
16	1748.7973	1846.7742	1731.9249	56	5842.8306	5940.8075	5834.9837	---
17	1819.8344	1917.8113	1788.9464	57	5939.8834	6037.8603	5949.0266	---
18	1876.8559	1974.8328	1885.9991	58	6036.9362	6134.9131	6006.0481	---
19	1990.8988	2088.8757	1983.0519	59	6093.9576	6191.9345	6077.0852	---
20	2087.9516	2185.9285	2080.1046	60	6222.0526	6320.0295	6190.1693	---
21	2534.1503	2632.1272	2177.1574	61	6319.1054	6417.0822	6303.2533	---
22	2591.1718	2689.1487	2234.1789	62	6447.1639	6545.1408	6390.2854	---
23	2688.2246	2786.2015	2362.2375	63	6504.1854	6602.1623	6487.3381	---
24	2775.2566	2873.2335	2459.2902	64	6601.2382	6699.2151	6574.3702	---
25	2872.3094	2970.2863	2587.3852	65	6698.2909	6796.2678	6703.4128	---
26	3000.3679	3098.3448	2701.4281	66	6795.3437	6893.3206	6832.4554	---
27	3057.3894	3155.3663	2758.4496	67	6923.4023	7021.3792	6960.5139	---
28	3114.4109	3212.3878	2815.471	68	6980.4237	7078.4006	7127.5123	7029.5354
29	3228.4538	3326.4307	2943.5296	69	7095.4507	7193.4276	7226.5807	7128.6038
30	3356.5488	3454.5257	3040.5824	70	7223.5456	7321.5225	7341.6076	7243.6308
31	3453.6015	3551.5784	3137.6351	71	7310.5777	7408.5546	7470.6502	7372.6733
32	3581.6601	3679.637	3234.6879	72	7466.6788	7564.6557	7584.6932	7486.7163
33	3638.6816	3736.6585	3291.7094	73	7553.7108	7651.6877	7697.7772	7599.8003
34	3735.7343	3833.7112	3419.7679	74	7650.7636	7748.7405	7811.8202	7713.8433
35	3832.7871	3930.764	3516.8207	75	---	---	---	---
36	3929.8399	4027.8168	3644.9157					
37	4026.8926	4124.8695	3701.9371					
38	4123.9454	4221.9223	3798.9899					
39	4180.9668	4278.9437	3896.0427					
40	4309.0618	4407.0387	3993.0954					

P-C + 1 DC

GRP**Q**GPP**QQ**GGH**QQ**GPPPPPPG**K**$_{23}$P**Q**GPPP**Q**GGRP**Q**GPP**Q**G**Q**$_{41}$SP**Q**

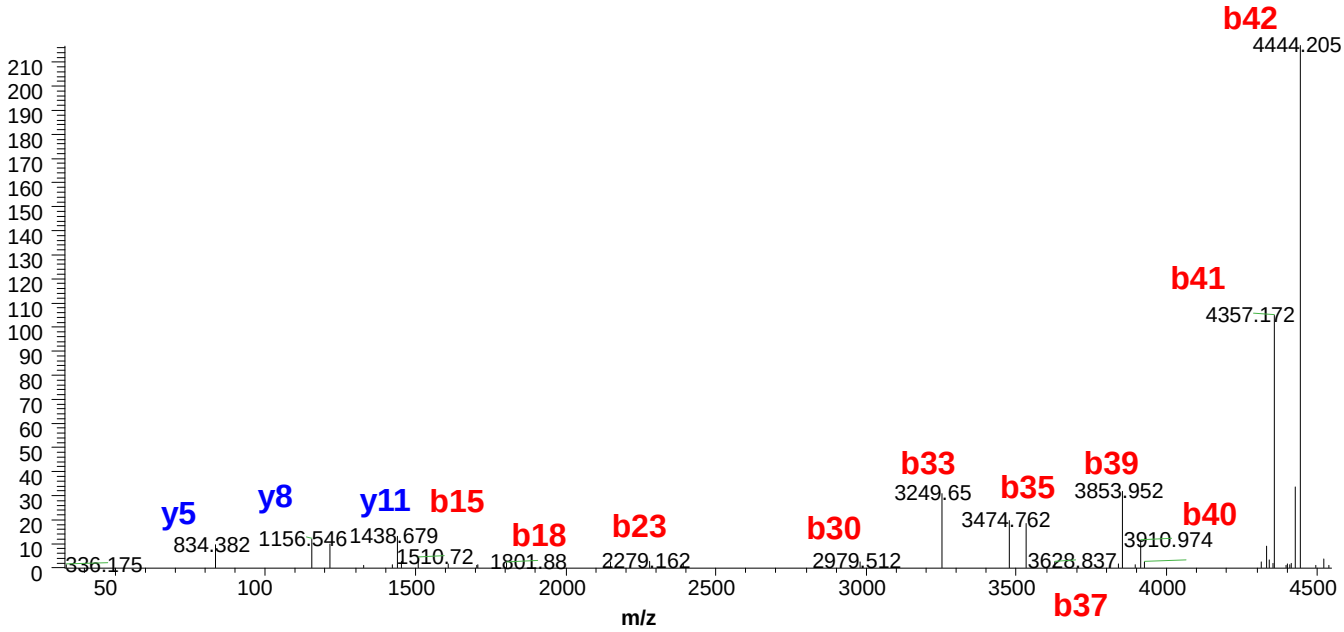


Figure S12 MSMS spectrum for PC + 1DC on $[M+5H]^{5+}$ on 938.8 *m/z* (CID)

Q_{41} is the principal acceptor for DC.

(PC-peptide + DC – NH_3 => 4369.18 + 335.17 - 17.03 = 4687.36).

Table S12 Theoretical fragments of P-C + DC (DC–NH3=318.14 after b_{41})

fragment number	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	23	2278.1588	2595.2885
2	214.1299	244.1292	24	2375.2116	2692.3413
3	311.1826	331.1612	25	2503.2701	2789.394
4	439.2412	777.36	26	2560.2916	2886.4468
5	496.2627	834.3815	27	2657.3444	2983.4996
6	593.3154	962.44	28	2754.3971	3080.5523
7	690.3682	1059.4928	29	2851.4499	3177.6051
8	818.4268	1156.5456	30	2979.5085	3234.6265
9	946.4853	1213.567	31	3036.5299	3362.6851
10	1003.5068	1341.6256	32	3093.5514	3490.7437
11	1060.5283	1438.6784	33	3249.6525	3627.8026
12	1197.5872	1594.7795	34	3346.7053	3684.8241
13	1325.6458	1651.8009	35	3474.7639	3741.8455
14	1453.7043	1708.8224	36	3531.7853	3869.9041
15	1510.7258	1836.881	37	3628.8381	3997.9627
16	1607.7786	1933.9337	38	3725.8909	4095.0155
17	1704.8313	2030.9865	39	3853.9494	4192.0682
18	1801.8841	2128.0393	40	3910.9709	4249.0897
19	1898.9369	2185.0607	41	4357.1697	4377.1483
20	1995.9896	2313.1193	42	4444.2017	4474.201
21	2093.0424	2410.1721	43	4541.2545	4630.3021
22	2150.0638	2538.267	44	---	---

P-C + 2 DC

GRP**Q**GPP**QQ**GGH**QQ**GPPPPPPG**K**$_{23}$P**Q**GPPP**Q**GGRP**Q**GPP**Q**$_{39}$G**Q**$_{41}$SP**Q**

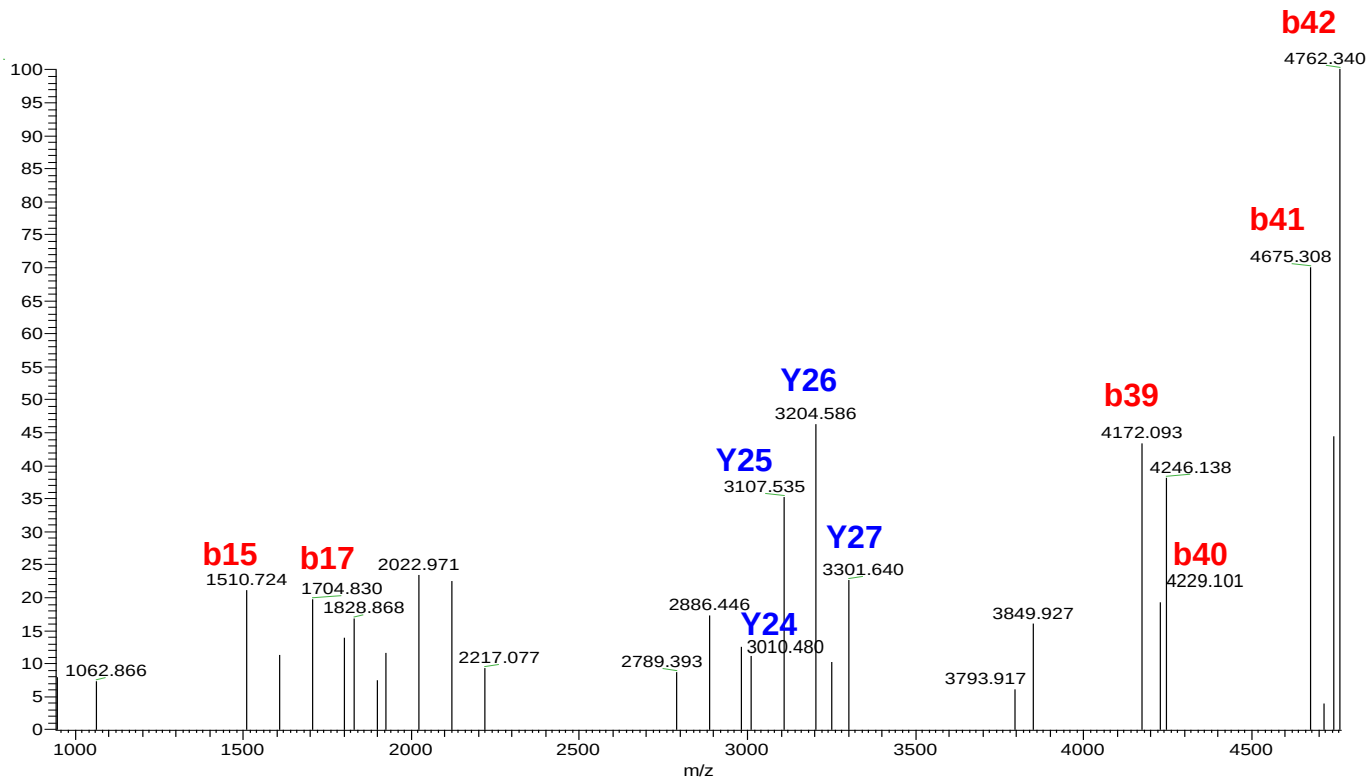


Figure S13 MSMS spectrum for PC + 2DC on $[M+6H]^{6+}$ on 835.58 *m/z* (CID)
Q_{41} and Q_{39} were the acceptors for DC.
(PC-peptide + 2DC – $2NH_3$ => 4369.18 + 670.34 - 34.06 = 5005.46).

Table S13 Theoretical fragments of P-C + 2DC (DC–NH3=318.14 after b_{41} and DC–NH3=318.14 after b_{39})

fragment number	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	23	2278.1588	2913.4287
2	214.1299	244.1292	24	2375.2116	3010.4815
3	311.1826	331.1612	25	2503.2701	3107.5342
4	439.2412	777.36	26	2560.2916	3204.587
5	496.2627	834.3815	27	2657.3444	3301.6398
6	593.3154	1280.5802	28	2754.3971	3398.6925
7	690.3682	1377.633	29	2851.4499	3495.7453
8	818.4268	1474.6858	30	2979.5085	3552.7667
9	946.4853	1531.7072	31	3036.5299	3680.8253
10	1003.5068	1659.7658	32	3093.5514	3808.8839
11	1060.5283	1756.8186	33	3249.6525	3945.9428
12	1197.5872	1912.9197	34	3346.7053	4002.9643
13	1325.6458	1969.9411	35	3474.7639	4059.9857
14	1453.7043	2026.9626	36	3531.7853	4188.0443
15	1510.7258	2155.0212	37	3628.8381	4316.1029
16	1607.7786	2252.0739	38	3725.8909	4413.1557
17	1704.8313	2349.1267	39	4172.0896	4510.2084
18	1801.8841	2446.1795	40	4229.1111	4567.2299
19	1898.9369	2503.2009	41	4675.3099	4695.2885
20	1995.9896	2631.2595	42	4762.3419	4792.3412
21	2093.0424	2728.3123	43	4859.3947	4948.4423
22	2150.0638	2856.4072	44	---	---

cyclo P-C + 1 DC

GRP**Q**GPP**QQ**GGH**QQ**GPPPPPPG**K**$_{23}$P**Q**GPPP**Q**GGRP**Q**GPP**Q**$_{39}$G**Q**$_{41}$SP**Q**

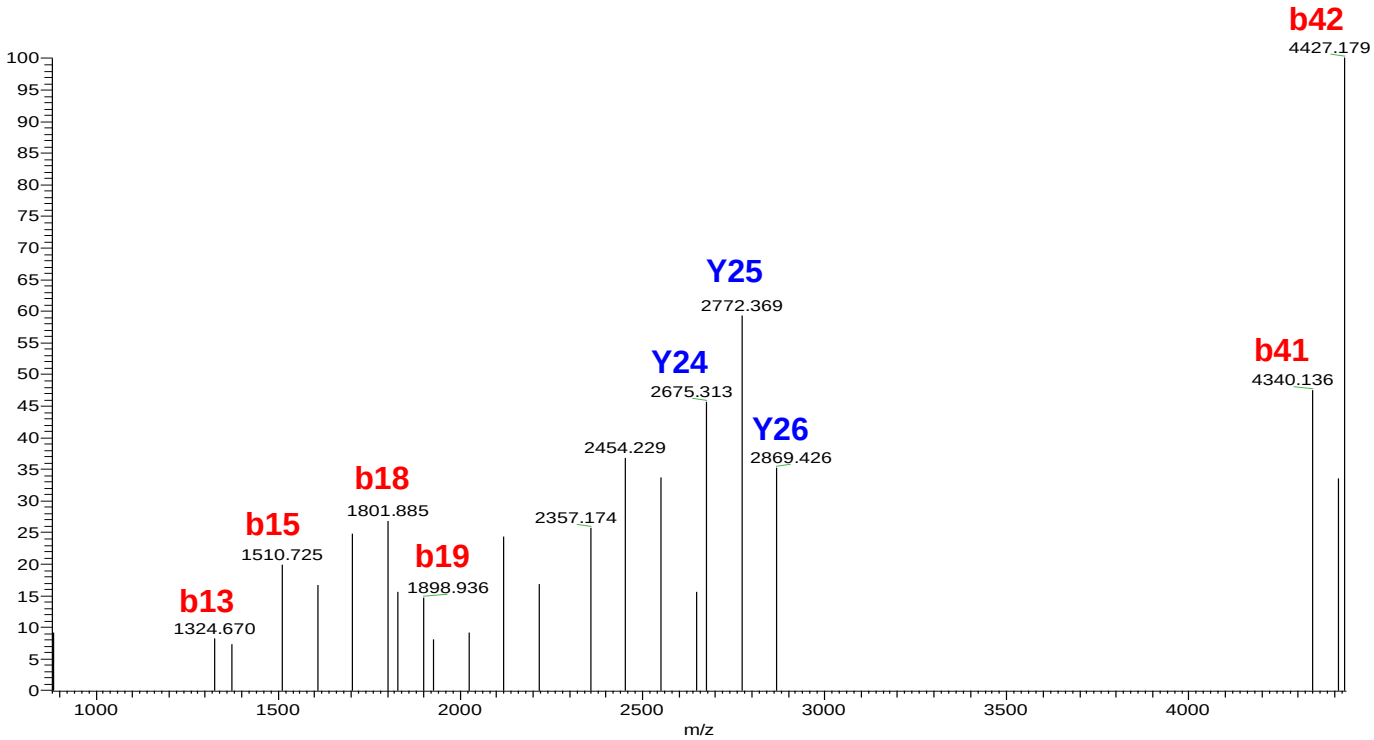


Figure S14 MSMS spectrum for cyclo PC + 1DC on $[M+5H]^{5+}$ on 935.478 *m/z* (CID) (cyclo-PC + DC – NH_3 => 4352.15 + 335.17 - 17.03 = 4670.3)

Table S14 Theoretical fragments of cyclo P-C + 1DC (less NH3=17.03 after b_{41} and DC–NH3=318.14 after b_{39})

fragment number	*b*	*y*	*fragment number*	*b*	*y*
1	---	147.0764	23	2278.1588	2578.2619
2	214.1299	244.1292	24	2375.2116	2675.3147
3	311.1826	331.1612	25	2503.2701	2772.3674
4	439.2412	442.1932	26	2560.2916	2869.4202
5	496.2627	499.2147	27	2657.3444	2966.473
6	593.3154	945.4134	28	2754.3971	3063.5257
7	690.3682	1042.4662	29	2851.4499	3160.5785
8	818.4268	1139.519	30	2979.5085	3217.5999
9	946.4853	1196.5404	31	3036.5299	3345.6585
10	1003.5068	1324.599	32	3093.5514	3473.7171
11	1060.5283	1421.6518	33	3249.6525	3610.776
12	1197.5872	1577.7529	34	3346.7053	3667.7975
13	1325.6458	1634.7743	35	3474.7639	3724.8189
14	1453.7043	1691.7958	36	3531.7853	3852.8775
15	1510.7258	1819.8544	37	3628.8381	3980.9361
16	1607.7786	1916.9071	38	3725.8909	4077.9889
17	1704.8313	2013.9599	39	4172.0896	4175.0416
18	1801.8841	2111.0127	40	4229.1111	4232.0631
19	1898.9369	2168.0341	41	4340.1431	4360.1217
20	1995.9896	2296.0927	42	4427.1751	4457.1744
21	2093.0424	2393.1455	43	4524.2279	4613.2755
22	2150.0638	2521.2404	44	---	---

Statherin

DssEE**K**$_6$FLRRIGRFGYGYGPY**Q**PVPE**Q**PLYP**Q**PY**Q**P**Q**$_{37}$Y**QQ**YTF

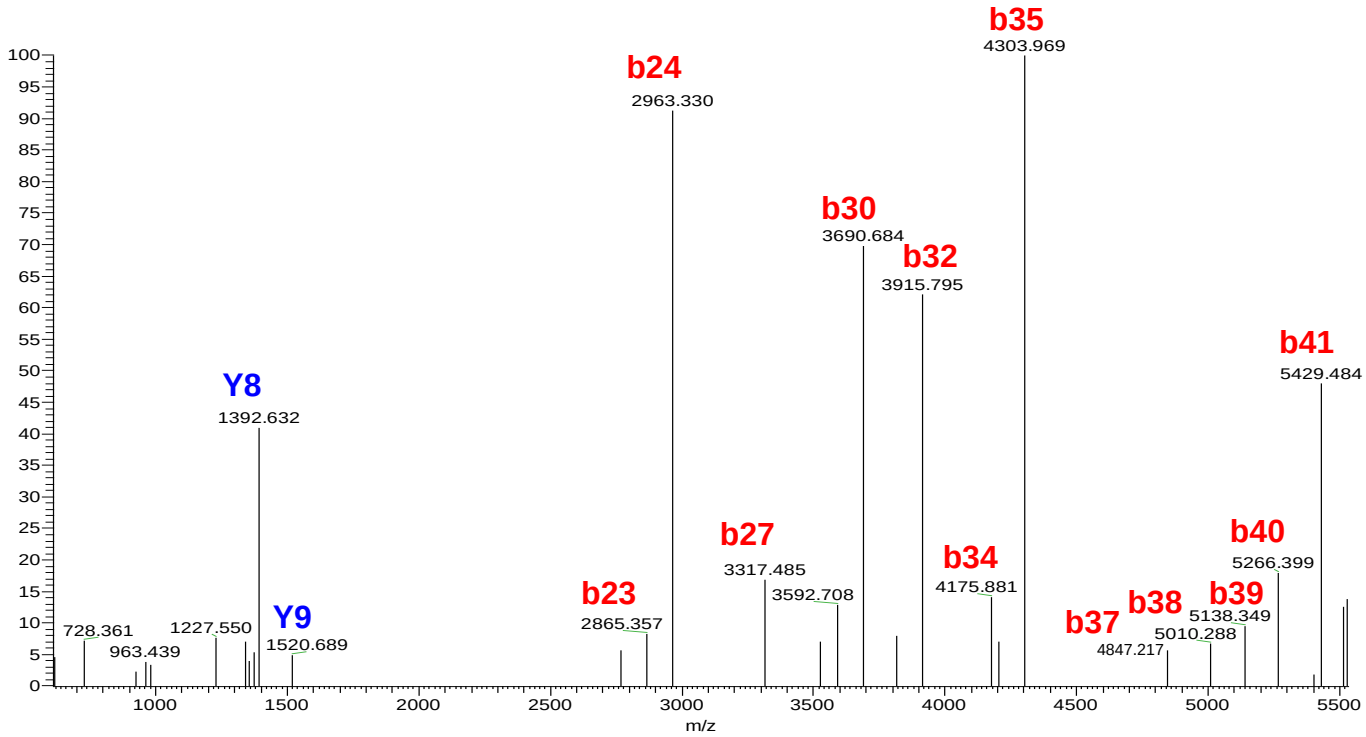


Figure S15 MSMS spectrum for statherin + 1DC on $[M+5H]^{5+}$ on 1140.33 *m/z* (CID). Q_{37} is the principal acceptor for DC-involved.
(Statherin + DC – NH_3 => 5378.44 + 335.17 - 17.03 = 5696.58).

Table S15 Theoretical fragments of statherin + 1DC (DC–NH3=318.14 after b_{37})

fragment number	*b-H_3PO_4*	*b*	*y*	*fragment number*	*b-H_3PO_4*	*b*	*y*	*y-H_3PO_4*
1	---	---	166.0863	23	2766.2824	2864.2593	3220.5125	---
2	185.0557	283.0326	267.1339	24	2865.3508	2963.3277	3317.5652	---
3	352.054	450.0309	430.1973	25	2962.4036	3060.3805	3374.5867	---
4	481.0966	579.0735	558.2558	26	3091.4462	3189.4231	3537.65	---
5	610.1392	708.1161	686.3144	27	3219.5048	3317.4817	3594.6715	---
6	738.2342	836.2111	849.3777	28	3316.5575	3414.5344	3757.7348	---
7	885.3026	983.2795	1295.5765	29	3429.6416	3527.6185	3814.7563	---
8	998.3867	1096.3636	1392.6293	30	3592.7049	3690.6818	3961.8247	---
9	1154.4878	1252.4647	1520.6879	31	3689.7577	3787.7346	4117.9258	---
10	1310.5889	1408.5658	1683.7512	32	3817.8163	3915.7932	4174.9473	---
11	1423.673	1521.6499	1780.804	33	3914.869	4012.8459	4288.0313	---
12	1480.6944	1578.6713	1908.8625	34	4077.9324	4175.9093	4444.1324	---
13	1636.7955	1734.7724	2005.9153	35	4205.9909	4303.9678	4600.2336	---
14	1783.8639	1881.8408	2168.9786	36	4303.0437	4401.0206	4713.3176	---
15	1840.8854	1938.8623	2282.0627	37	4749.2425	4847.2194	4860.386	---
16	2003.9487	2101.9256	2379.1155	38	4912.3058	5010.2827	4988.481	---
17	2060.9702	2158.9471	2507.174	39	5040.3644	5138.3413	5117.5236	---
18	2224.0335	2322.0104	2636.2166	40	5168.423	5266.3999	5246.5662	---
19	2281.055	2379.0319	2733.2694	41	5331.4863	5429.4632	5413.5645	5315.5877
20	2378.1078	2476.0847	2832.3378	42	5432.534	5530.5109	5580.5629	5482.586
21	2541.1711	2639.148	2929.3906	43	---	---	---	---
22	2669.2297	2767.2066	3057.4491					

Statherin + 2DC

DssEEK_6FLRRIGRFGYGYGPY**Q**PVPE**Q**PLYP**Q**PY**Q**PQ_{37}YQ_{39}**Q**YTF

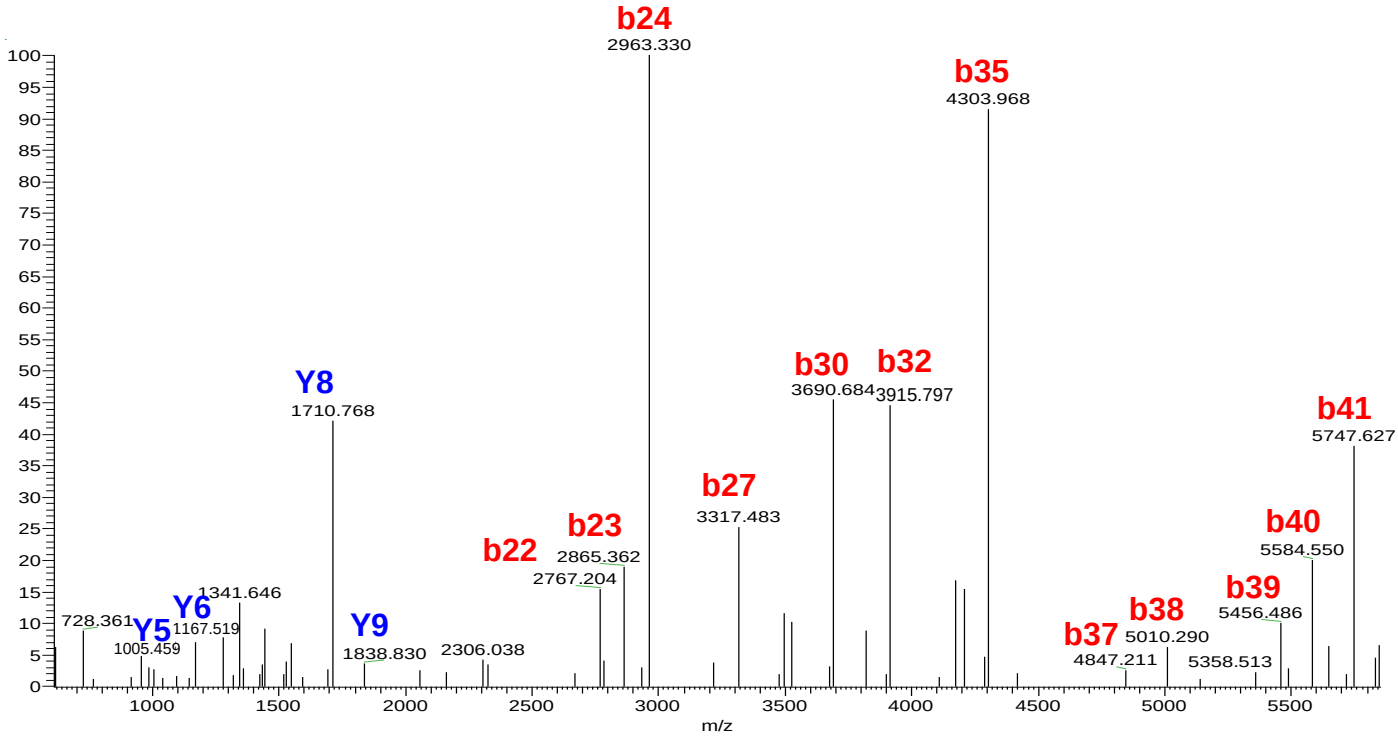


Figure S16 MSMS spectrum for statherin + 2DC on $[M+5H]^{5+}$ on 1204.55 *m/z* (CID). It was possible to establish Q_{39} as second acceptor for DC. (Statherin + 2DC – $2NH_3$ => 5378.44 + 670.34 – 34.06 = 6014.72).

Table S16 Theoretical fragments of statherin + 2DC (DC–NH3=318.14 after b_{37} and DC–NH3=318.14 after b_{39})

fragment number	*b- H_3PO_4*	*b*	*y*	*fragment number*	*b- H_3PO_4*	*b*	*y*	*y-H_3PO_4*
1	---	---	166.0863	23	2766.2824	2864.2593	3538.6527	---
2	185.0557	283.0326	267.1339	24	2865.3508	2963.3277	3635.7054	---
3	352.054	450.0309	430.1973	25	2962.4036	3060.3805	3692.7269	---
4	481.0966	579.0735	558.2558	26	3091.4462	3189.4231	3855.7902	---
5	610.1392	708.1161	1004.4546	27	3219.5048	3317.4817	3912.8117	---
6	738.2342	836.2111	1167.5179	28	3316.5575	3414.5344	4075.875	---
7	885.3026	983.2795	1613.7167	29	3429.6416	3527.6185	4132.8965	---
8	998.3867	1096.3636	1710.7695	30	3592.7049	3690.6818	4279.9649	---
9	1154.4878	1252.4647	1838.8281	31	3689.7577	3787.7346	4436.066	---
10	1310.5889	1408.5658	2001.8914	32	3817.8163	3915.7932	4493.0875	---
11	1423.673	1521.6499	2098.9442	33	3914.869	4012.8459	4606.1715	---
12	1480.6944	1578.6713	2227.0027	34	4077.9324	4175.9093	4762.2726	---
13	1636.7955	1734.7724	2324.0555	35	4205.9909	4303.9678	4918.3738	---
14	1783.8639	1881.8408	2487.1188	36	4303.0437	4401.0206	5031.4578	---
15	1840.8854	1938.8623	2600.2029	37	4749.2425	4847.2194	5178.5262	---
16	2003.9487	2101.9256	2697.2557	38	4912.3058	5010.2827	5306.6212	---
17	2060.9702	2158.9471	2825.3142	39	5358.5046	5456.4815	5435.6638	---
18	2224.0335	2322.0104	2954.3568	40	5486.5632	5584.5401	5564.7064	---
19	2281.055	2379.0319	3051.4096	41	5649.6265	5747.6034	5731.7047	5633.7279
20	2378.1078	2476.0847	3150.478	42	5750.6742	5848.6511	5898.7031	5800.7262
21	2541.1711	2639.148	3247.5308	43	---	---	---	---
22	2669.2297	2767.2066	3375.5893					

Statherin + 3DC

DssEEK_6FLRRIGRFGYGYGPY**Q**PVPE**Q**PLYP**Q**PY**Q**PQ_{37}Y$Q_{39}Q_{40}$YTF

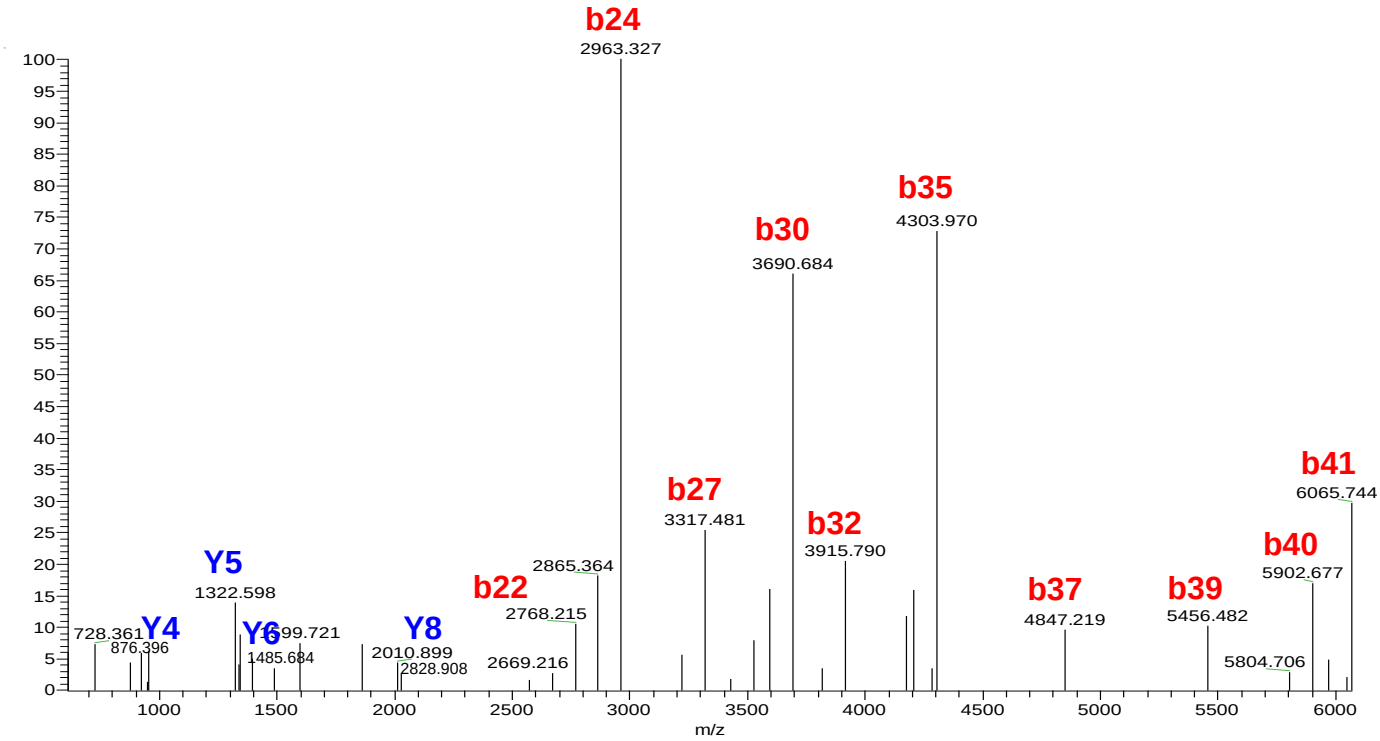


Figure S17 MSMS spectrum for statherin + 3DC on $[M+5H]^{5+}$ on 1267.98 *m/z* (CID). It was possible to establish Q_{40} as third acceptor for DC. (Statherin + 3DC – $3NH_3$ => 5378.44 + 1005.51 – 51.09 = 6332.86).

Table S16 Theoretical fragments of statherin + 3DC (DC–NH3=318.14 after b_{37} and DC–NH3=318.14 after b_{39} and DC–NH3=318.14 after b_{40})

fragment number	*b-H_3PO_4*	*b*	*y*	*fragment number*	*b-H_3PO_4*	*b*	*y*	*y-H_3PO_4*
1	---	---	166.0863	23	2766.2824	2864.2593	3856.7929	---
2	185.0557	283.0326	267.1339	24	2865.3508	2963.3277	3953.8456	---
3	352.054	450.0309	430.1973	25	2962.4036	3060.3805	4010.8671	---
4	481.0966	579.0735	876.396	26	3091.4462	3189.4231	4173.9304	---
5	610.1392	708.1161	1322.5948	27	3219.5048	3317.4817	4230.9519	---
6	738.2342	836.2111	1485.6581	28	3316.5575	3414.5344	4394.0152	---
7	885.3026	983.2795	1931.8569	29	3429.6416	3527.6185	4451.0367	---
8	998.3867	1096.3636	2028.9097	30	3592.7049	3690.6818	4598.1051	---
9	1154.4878	1252.4647	2156.9683	31	3689.7577	3787.7346	4754.2062	---
10	1310.5889	1408.5658	2320.0316	32	3817.8163	3915.7932	4811.2277	---
11	1423.673	1521.6499	2417.0844	33	3914.869	4012.8459	4924.3117	---
12	1480.6944	1578.6713	2545.1429	34	4077.9324	4175.9093	5080.4128	---
13	1636.7955	1734.7724	2642.1957	35	4205.9909	4303.9678	5236.514	---
14	1783.8639	1881.8408	2805.259	36	4303.0437	4401.0206	5349.598	---
15	1840.8854	1938.8623	2918.3431	37	4749.2425	4847.2194	5496.6664	---
16	2003.9487	2101.9256	3015.3959	38	4912.3058	5010.2827	5624.7614	---
17	2060.9702	2158.9471	3143.4544	39	5358.5046	5456.4815	5753.804	---
18	2224.0335	2322.0104	3272.497	40	5804.7034	5902.6803	5882.8466	---
19	2281.055	2379.0319	3369.5498	41	5967.7667	6065.7436	6049.8449	5951.8681
20	2378.1078	2476.0847	3468.6182	42	6068.8144	6166.7913	6216.8433	6118.8664
21	2541.1711	2639.148	3565.671	43	---	---	---	---
22	2669.2297	2767.2066	3693.7295					

Journal of proteome research

Cite This: J. Proteome Res. 2018, 17, 3292−3307

Article

pubs.acs.org/jpr

Extensive Characterization of the Human Salivary Basic Proline-Rich Protein Family by Top-Down Mass Spectrometry

Alessandra Padiglia,[†] Roberto Orrù,[†] Mozhgan Boroumand,[†] Alessandra Olianas,[*,†] Barbara Manconi,[†] Maria Teresa Sanna,[†] Claudia Desiderio,[‡] Federica Iavarone,[§,∥] Barbara Liori,[†] Irene Messana,[‡] Massimo Castagnola,[‡,§,∥] and Tiziana Cabras[†]

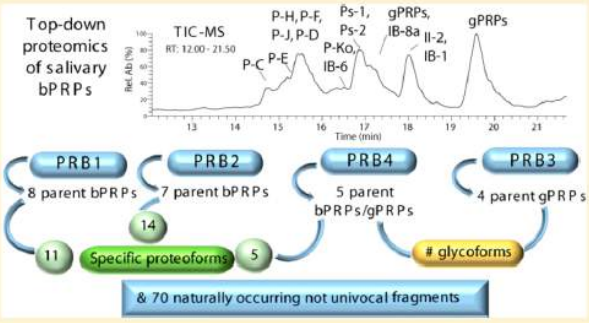
[†]Department of Life and Environmental Sciences, University of Cagliari, Cittadella Univ. Monserrato, Monserrato 09042, Cagliari, Italy

[‡]Institute of Chemistry of Molecular Recognition, CNR, Rome 00168, Italy

[§]Institute of Biochemistry and Clinical Biochemistry, Università Cattolica del Sacro Cuore, Rome 00168, Italy

[∥]Department of Laboratory Diagnostic and Infectious Diseases, Fondazione Policlinico Universitario Agostino Gemelli-IRCCS, Rome 00168, Italy

ABSTRACT: Human basic proline-rich proteins and basic glycosylated proline-rich proteins, encoded by the polymorphic *PRB1−4* genes and expressed only in parotid glands, are the most complex family of adult salivary proteins. The family includes 11 parent peptides/proteins and more than 6 parent glycosylated proteins, but a high number of proteoforms with rather similar structures derive from polymorphisms and post-translational modifications. 55 new components of the family were characterized by top-down liquid chromatography−mass spectrometry and tandem-mass platforms, bringing the total number of proteoforms to 109. The new components comprise the three variants P-H $S_1 \rightarrow A$, P-Ko $P_{36} \rightarrow S$, and P-Ko $A_{41} \rightarrow S$ and several of their naturally occurring proteolytic fragments. The paper represents an updated reference for the peptides included in the heterogeneous family of proteins encoded by *PRB1/PRB4*. MS data are available via ProteomeXchange with the identifier PXD009813.



KEYWORDS: *basic proline-rich proteins, top-down proteomics, mass spectrometry, human saliva*

INTRODUCTION

Proline-rich proteins (PRPs) are a family of human salivary proteins classified as acid (aPRPs), basic (bPRPs), and glycosylated basic (gPRPs). bPRPs and gPRPs are secreted only by the human parotid glands, where they represent >50% w/w of all of the parotid proteins.[1] They are the most composite family of salivary proteins[2,3] coded by a cluster of six genes, strictly associated in a segment of ∼4.0 Kb in length on chromosome 12 at band 13.2.[4−6] In particular, the cluster of genes encoding for bPRPs and gPRPs includes four loci named *PRB1−PRB4*, each one existing in several allelic forms. Each PRB gene covers four exons, the third of which is fully composed of 63-bp tandem repeats coding the proline-rich portion of the protein products. Variation in the numbers of these repeats is responsible for length differences in different alleles of the PRB genes.[7,8] At least four alleles (S, small; M, medium; L, large; and VL, very large) are present in the Western population at *PRB1* and *PRB3* loci and three (S, M, L) at *PRB2* and *PRB4* loci.[9] In addition to tandem repeat length variations, these alleles show SNPs in the coding region, polymorphic cleavage sites, and polymorphic stop codons. Moreover, alternative splicing generates multiple transcript variants encoding distinct proteins, and some alleles are still pending for their characterization.[10−12] Genetic variability, post-translational modifications (PTMs) implicated in the presecretory maturation processes, and further transformations occurring in the oral environment give a contribution to the heterogeneity of bPRPs and gPRPs. The proteolytic cleavage is the main occurring post-translational event. Indeed, except for the protein encoded by the *PRB3* locus that originates several gPRPs, the other pro-proteins before granule maturation are completely cleaved by proprotein convertases, generating smaller peptides.[3,13] Moreover, after secretion, these peptides are further cleaved by endogenous and exogenous (microbioma) proteinases generating numerous fragments.[14,15]

More than 15 years ago, our group undertook the characterization of the principal bPRPs detectable in human whole saliva by an integrated top-down/bottom-up RP−HPLC−ESI−MS platform.[16] At the time, the used ion-trap MS with a resolution of ∼1/5000 did not allow us to characterize all of the masses potentially attributable to bPRPs. In the last years the availability of a high-resolution MS apparatus (Orbitrap MS) increased our analytical skills, and by using a

Received: June 12, 2018
Published: July 31, 2018

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

top-down MS/MS platform, we were able to establish the structure of many other members of bPRPs. Although the current knowledge on the bPRP family cannot be considered conclusive, we describe here the state of the art of the naturally occurring proteins encoded by *PRB1/PRB4*. This comprehensive overview may be a useful reference for the scientists involved in the investigation of human saliva.

■ EXPERIMENTAL SECTION

Reagents

Chemicals and reagents, all of LC−MS grade, were purchased from Merck/Sigma-Aldrich (Darmstadt, Germany), Waters (Milford, MA), and Thermo Fisher Scientific (Rockford, IL).

Ethics Statements and Subjects under Study

The study protocol and written consent form were approved by the Ethical Committee of the Catholic University of Rome and have been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. All rules were respected, and written consent forms were obtained by the donors. Unstimulated whole saliva (WS) was collected from 86 adult healthy donors (40 ± 10 years old, males $n = 42$, females $n = 44$).

Sample Collection

Unstimulated WS was collected according to a standardized protocol optimized to preserve saliva proteins from proteolytic degradation. Donors did not eat or drink at least 2 h before the collection, which was performed in the morning between 10:00 A.M. and 12:00 A.M. with a soft plastic aspirator. Saliva was transferred into a plastic tube in ice bath, and 0.2% 2,2,2-trifluoroacetic acid (TFA) was immediately added in 1:1 v/v ratio. The solution was then centrifuged at 10 000*g* for 10 min at 4 °C. The acidic supernatant was separated from the precipitate and either immediately analyzed by HPLC−ESI−MS or stored at −80 °C until the analysis.

HPLC−Low-Resolution ESI−IT−MS Experiments

The acid-soluble fractions (33 μL, corresponding to 16.5 μL of whole saliva) of salivary proteins/peptides have been analyzed by reversed-phase (RP)−HPLC−low-resolution ESI−IT−MS apparatus, constituted by a Surveyor HPLC system connected to an LCQ Advantage mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped with an ESI source. The chromatographic column was a Vydac (Hesperia, CA) C8 column with 5 μm particle diameter (150 × 2.1 mm). The eluents were the following: (eluent A) 0.056% (v/v) aqueous TFA and (eluent B) 0.05% (v/v) TFA in acetonitrile/water 80/20. The gradient applied was linear from 0 to 55% of B in 40 min and from 55 to 100% of B in 10 min, at a flow rate of 0.10 mL/min toward the ESI source. During the first 5 min of separation, eluate was diverted to waste to avoid source contamination because of the high salt concentration. Mass spectra were collected every 3 ms in the m/z range 300−2000 in positive ion mode. The MS spray voltage was 5.0 kV, and the capillary temperature was 260 °C. MS resolution was 6000. Deconvolution of averaged ESI−MS spectra was performed by MagTran 1.0 software.[17] Average experimental mass values (Mav) were compared with the relative theoretical ones using the PeptideMass program available on the Swiss-Prot data bank (https://www.expasy.org/proteomics).

nanoHPLC−High-Resolution ESI−MS/MS Experiments

For the structural characterization, 67 samples were analyzed by nanoHPLC−high-resolution MS/MS with an Ultimate 3000 RSLC Nano System HPLC apparatus (Thermo Fisher Scientific, Sunnyvale, CA) coupled to an LTQ-Orbitrap Elite apparatus (Thermo Fisher Scientific). The used chromatographic column was a Zorbax 300SB-C8 (3.5 μm particle diameter; 1.0 × 150 mm). Eluents were: (eluent A) 0.1% (v/v) aqueous formic acid (FA) and (eluent B) 0.1% (v/v) FA in acetonitrile/water 80/20. The gradient was: 0−2 min 5% B, 2−40 min from 5 to 55% B (linear), and 40−45 min from 70 to 99% B at a flow rate of 50 μL/min. MS and MS/MS spectra of intact proteins and peptides were collected in positive mode with resolution of 60 000. The acquisition range was from 350 to 2000 m/z, and the tuning parameters were: capillary temperature 300 °C, source voltage 4.0 kV, and S-Lens RF level 60%. In data-dependent acquisition mode the five most intense ions were selected and fragmented by using collision-induced dissociation (CID) or higher energy collision dissociation (HCD), with 35% normalized collision energy for 1 ms, isolation width of 5 m/z, and activation q of 0.25. The injected volume was 20 μL. HPLC−ESI−MS and MS/MS data were generated by Xcalibur 2.2 SP1.48 (Thermo Fisher Scientific) using default parameters of the Xtract program for the deconvolution. MS/MS data were analyzed by both manual inspection of the MS/MS spectra recorded along the chromatogram and the Proteome Discoverer 1.4 software elaboration based on the SEQUEST HT cluster as a search engine (University of Washington, licensed to Thermo Electron Corporation, San Jose, CA) against the UniProtKB human data-bank (163 117 entries, release 2018_02). For peptide matching, high-confidence filter settings were applied: The peptide score threshold was 2.3, and the limits were Xcorr scores >1.2 for singly charged ions and 1.9 and 2.3 for doubly and triply charged ions, respectively. The false discovery rate (FDR) was set to 0.01 (strict) and 0.05 (relaxed), and the precursor and fragment mass tolerance was 10 ppm and 0.5 Da, respectively. Pyroglutamination from E or Q residues and serine phosphorylation were selected as dynamic modifications. Because of the difficulties of the automated software to detect with high confidence every bPRP and its fragments, the structural information derives in part from manual inspections of the MS/MS spectra, obtained by either CID or HCD fragmentation, against the theoretical ones generated by MS-Product software available at the Protein Prospector Web site (http://prospector.ucsf.edu/prospector/mshome.htm). All of the MS/MS spectra (HCD or CID) have been manually verified by utilizing every fragmentation spectra with a significant number of fragment ions.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://www.ebi.ac.uk/pride) via the PRIDE[18] partner repository with the data set identifier PXD009813.

■ RESULTS

bPRPs and gPRPs are usually detected in the RP−HPLC−ESI−MS TIC (total ion current) profile as a characteristic cluster in the elution window comprised between 14.5 and 20.5 min under our experimental conditions. Figure 1A shows the typical TIC profile of the soluble acid fraction of adult human whole saliva obtained by RP−HPLC−low-resolution ESI−MS analysis. The main bPRPs detectable between 14.5

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

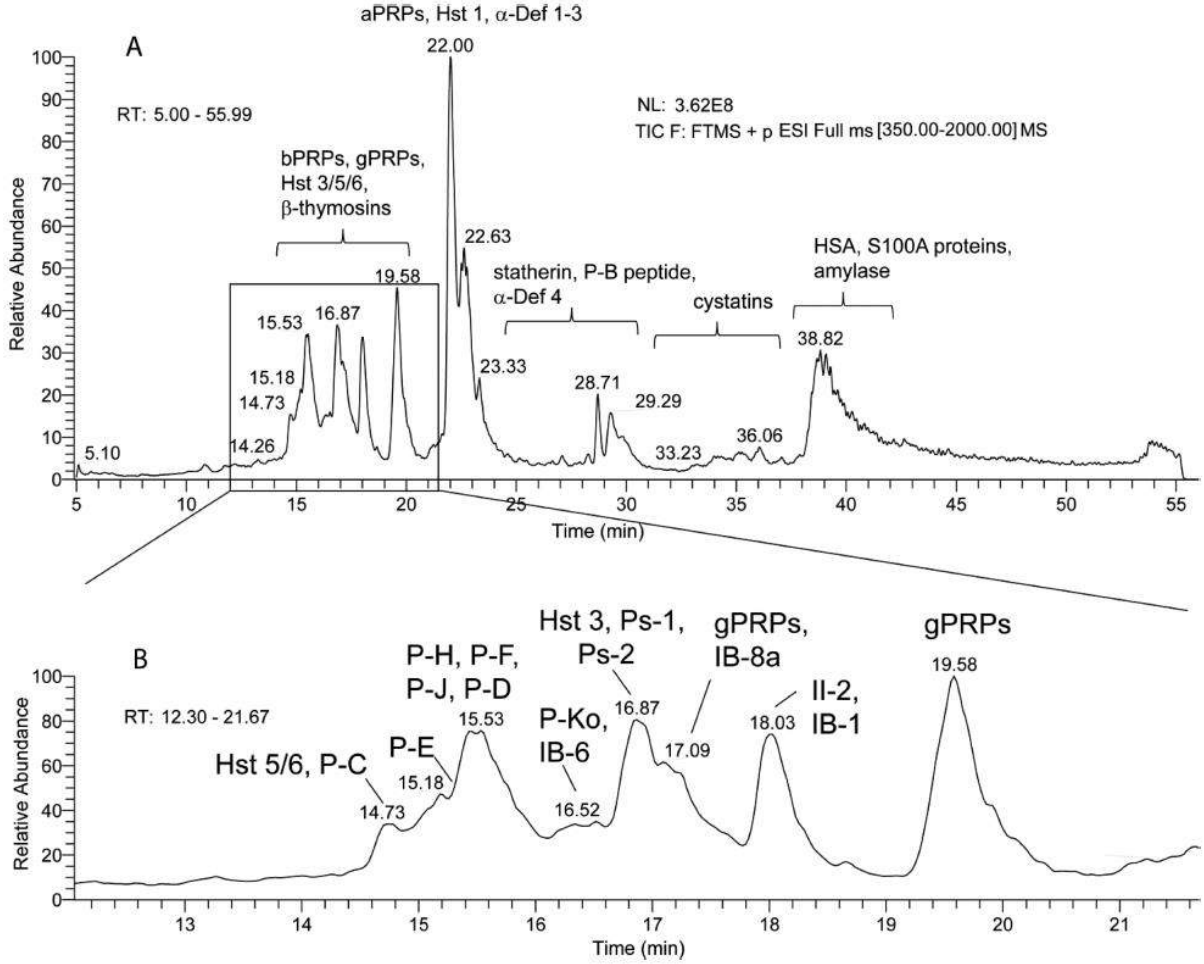


Figure 1. (A) Typical HPLC−ESI−MS TIC profile of the acid-soluble fraction of adult human whole saliva showing the elution times of the principal families of salivary proteins. Basic (bPRPs) and glycosylated basic proline-rich proteins (gPRPs) for their structural similarity elute in a cluster comprised between 14.5 and 20.5 min under the experimental conditions used. (B) Enlargement of the elution cluster of bPRPs and gPRPs, with the indication of the approximate chromatographic position. NL: normalization level; α-Def: α-defensins 1−4; Hst: histatin.

and 20.5 min and shown in the enlargement of Figure 1B elute closely, often in a single unresolved chromatographic peak, and some of them, such as P-J and P-F, are difficult to separate due to their structural similarity. bPRPs and gPRPs encoded by *PRB1*, *PRB2*, and *PRB4* loci are completely cleaved by pro-protein convertases before secretion, and thus in saliva only fragments of the pro-proteins can be detected. The proteins and peptides described in the following sections were identified by analyzing the MS/MS fragmentation spectra with both the manual inspection and the Proteome Discoverer tool, and all of the identifications provided by the software were verified by our manual examination of the spectra. The results of the top-down identification of the several proteoforms are reported in Tables 1−9 and available via ProteomeXchange with the identifier PXD009813. Proteoforms identified by only manual inspection of the MS/MS spectra are indicated in each Table with an asterisk.

Products of *PRB1* Locus

Figure 2 shows the parent peptides deriving from maturation of the proteins encoded by the different alleles of *PRB1* locus detected in the Western population. They are P-E (also named IB-9), II-2, P-Ko, IB-6, Ps-1, and Ps-2.

Table 1 reports their mass values, elution times, and sequences, together with frequencies within the cohort of 86 analyzed samples. We confirmed by manual inspection of the

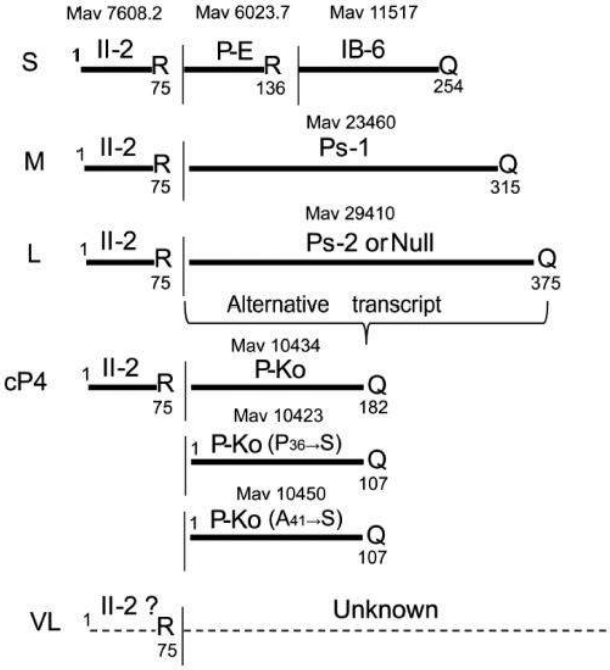


Figure 2. Schematic representation of the human salivary *PRB1* locus and their alleles, showing the coding regions for parent bPRPs.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 1. Average (Mav) and $[M + H]^{1+}$ Monoisotopic Masses, Elution Times, Frequency, and Sequence of the Products of *PRB-1* Locus (UniProtKB code P04280)[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ± 0.4)	freq (n = 86)	sequence[b]
P-E (or IB-9)*	6023.7 ± 0.7 (6023.69)	6021.09 ± 0.03 (6021.088)	14.9	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNRPQG PPPPGKPQGP PPQGDKSRSP R
II-2*	7608.2 ± 0.8 (7608.19)	7604.69 ± 0.04 (7604.712)	19.2	86	<QNLNEDVSQE ESPSLIAGNP QGPSPQGGNK PQGPPPPPGK PQGPPPQGGN KPQGPPPPGK PQGPPPQGDK SRSPR
P-Ko*	10434 ± 1.1 (10433.57)	10428.29 ± 0.05 (10428.285)	16.0	65	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKPQGPP AQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
P-Ko P_{36}→S*	10423 ± 1.0 (10423.46)	10418.28 ± 0.05 (10418.264)	15.8	1	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKSQGPP AQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
P-Ko A_{41}→S*[c]	10450 ± 1.1 (10449.57)	10444.30 ± 0.05 (10444.279)	16.0	11	SPPGKPQGPP PQGGKPQGPP PQGGNKPQGP PPPGKPQGPP SQGGSKSQSA RAPPGKPQGP PQQEGNNPQG PPPPAGGNPQ QPQAPPAGQP QGPPRPPQGG RPSRPPQ
IB-6*	11517 ± 1.2 (11516.67)	11510.80 ± 0.06 (11510.799)	16.7	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PAQGGSKSQS ARSPPGKPQG PPQQEGNNPQ GPPPPAGGNP QQPQAPPAGQ PQGPPRPPQG GRPSRPPQ
Ps-1*[d]	23460 ± 3 (23459.07)	23445.9 ± 0.11 (23445.859)	17.6	52	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDKSQSP RSPPGKPQGP PPQGGNQPQG PPPPPGKPQG PPQQGGNRPQ GPPPPGKPQG PPPQGDKSRS PQSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGGNKP QGPPPPGKPQ GPPAQGGSKS QSARAPPGKP QGPPQQEGNN PQGPPPPAGG NPQQPQAPPA GQPQGPPRPP QGGRPSRPPQ
Ps-2	29410 ± 4 (29408.72)	29391.9 ± 0.14 (29391.881)	17.6	5	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDKSQSP RSPPGKPQGP PPQGGNQPQG PPPPPGKPQG PPPQGGNKPQ GPPPPGKPQG PPPQGDKSQS PRSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPQQGGNRP QGPPPPGKPQ GPPPQGDKSR SPQSPPGKPQ GPPPQGGNQP QGPPPPPGKP QGPPPQGGNK PQGPPPPGKP QGPPAQGGSK SQSARAPPGK PQGPPQQEGN NPQGPPPPAG GNPQQPQAPP AGQPQGPPRP PQGGRPSRPP Q

[a]Proteins characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]<Q: pyro-glutamic acid; S: phosphorylated Ser. [c]UniProtKB code G5E9X6. [d]UniProtKB code Q86YA1.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 2. Average (Mav) and $[M + H]^{1+}$ Monoisotopic Masses and Elution Times of the Main Derivatives of the Products of *PRB-1* Locus (UniProtKB code P04280)[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ± 0.4)
II-2 (Fr. 18-32)[b]	1462.7 ± 0.2 (1462.54)	1462.71 ± 0.01 (1462.703)	8.9
II-2 (Fr. 1-23) non phosph. pyro-Gln*	**2406.3 ± 0.3 (2406.45)**	**2406.11 ± 0.02 (2406.106)**	**21.8**
II-2 (Fr. 18-42)[b]	2415.2 ± 0.3 (2415.67)	2415.22 ± 0.02 (2415.216)	15.9
II-2 (Fr. 1-23) pyro-Gln, S_8(phosph)*	**2486.5 ± 0.3 (2486.43)**	**2486.07 ± 0.02 (2486.072)**	**20.5**
II-2 (Fr. 18-53)	**3474.4 ± 0.6 (3473.84)**	**3472.75 ± 0.03 (3472.747)**	**16.2**
II-2 (Fr. 20-67)	**4635.4 ± 0.8 (4635.18)**	**4633.41 ± 0.05 (4633.381)**	**17.0**
II-2 (Fr. 18-75)*	**5690.9 ± 0.6 (5690.35)**	**5687.92 ± 0.03 (5687.783)**	**16.1**
P-E Des R_{61}*[c]	5867.5 ± 0.6 (5867.50)	5864.98 ± 0.03 (5864.987)	14.9
II-2 Des $R_{72}SPR_{75}$ pyro-Gln, S_8(phosph)*	**7111.7 ± 0.8 (7111.68)**	**7108.43 ± 0.04 (7108.425)**	**19.1**
II-2 Des R_{75} pyro-Gln, S_8(phosph)*[c]	7452.0 ± 0.8 (7452.01)	7448.61 ± 0.04 (7448.612)	19.2
II-2 non phosph. pyro-Gln[c]	7528.3 ± 0.8 (7528.21)	7524.75 ± 0.04 (7524.746)	19.7

[a]Peptides characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]Identified also in ref 24. [c]Identified also in refs 3 and 16.

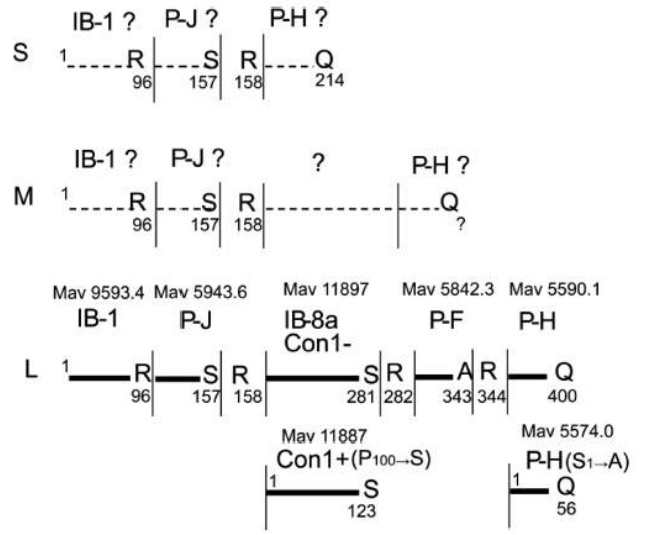


Figure 3. Schematic representation of the human salivary *PRB2* locus and their alleles, showing the coding regions for parent bPRPs. Sequences of S and M alleles are not known; however, we speculated that they should encode IB-1, P-J, and P-H peptides because these peptides were detected in all of the salivary samples investigated.

high-resolution MS/MS analysis sequences of P-E, II-2, and IB-6, bPRPs previously characterized by top-down proteomic platforms by our group[3,19] and other research groups.[20,21] MS and MS/MS data obtained on IB-6 did not match the sequence reported in UniProtKB data bank (code P04280) for the presence of a serine instead of an alanine at position 63. By our top-down proteomic approach we were able to characterize for the first time the structure of Ps-1 in its intact form (experimental monoisotopic $[M + H^+]^{1+}$ value 23 445.9 m/z), which is the protein previously identified by our group by a bottom-up approach.[22] Although sequencing of intact Ps-2 did not allow us to confirm its sequence with confidence, the experimental monoisotopic $[M + H^+]^{1+}$ value of 29 391.9 ± 0.14 m/z was in perfect agreement with the theoretical value ($[M + H^+]^{1+}$ 29 391.881 m/z) reported in databases (*PRB1-L* allele, UniProtKB code P04280). We did not detect peptides or proteins potentially deriving from *PRB1-VL* nor the peptide with an average mass of 8391.2 Da, corresponding to the splice variant classified by Maeda[8] with the acronym cP5. Moreover, we were able to detect and characterize the P-Ko protein encoded by *PRB1-L cP4* (Table 1) and two of its variants. The $P_{36} \rightarrow S$ variant, found in one sample (out of 86), was characterized from the inspection of MS/MS CID fragmentation spectra of $[M + 8H]^{8+}$ (1309.9 m/z) and $[M + 9H]^{9+}$ (1159.1 m/z) multiply charged ions. The attribution of substitution to P_{36} among the multiple proline residues present in the P-Ko sequence was based on the detection of the b_{35} (exp. 3323.74; theor. 3323.740 m/z), b_{37} (exp. 3538.84; theor. 3538.830 m/z), y_{71} (exp. 7008.51; theor. 7008.500 m/z), and y_{72} (exp. 7095.54; theor. 7095.532 m/z) ions. The $A_{41} \rightarrow S$ variant of P-Ko (Table 1) was detected in 11 samples (out of 86). Several b and y ions detected in the MS/MS CID spectra performed on the $[M + 8H]^{8+}$ (1307.2 m/z) ion were in agreement with the substitution of A_{41} or A_{50} residues, but some internal fragments were diagnostic for A_{41} substitution, in particular, the fragments $QGP_{30}PPPGKPQGPP_{40}SQ$ (exp. 1450.74; theor. 1450.744 m/z) and $PGKPQGPP_{40}SQGGSK$-$SKSQ_{50}$-NH_3 (exp. 1501.76; theor. 1501.739 m/z), in agreement with a serine residue at position 41, and the fragments $GSKSQSA_{50}RAPPGKPQGP_{60}$-$H_2O$ (exp. 1613.82; theor 1613.850 m/z) and $GSKSQSA_{50}RAPPGKPQGP_{60}$-$NH_3$ (exp. 1614.81; theor. 1614.835 m/z), in agreement with an alanine residue at position 50. Table 2 reports the most common derivatives of the main *PRB1* locus proteoforms, principally from II-2. Six out of 11 peptides of Table 2 were characterized in this study for the first time, whereas the others have also been described in previous studies.[3,23] A variant of II-2 peptide, lacking the proline residue at position 39, has been described,[21] but we were not able to detect it in any of the samples analyzed in the present study.

II-2 was detected in all of samples analyzed, as expected because it originates from all of the *PRB1* alleles (Figure 2). P-Ko was highly frequent in our cohort of healthy adult population, being detected in 68 subjects, of which 56 were homozygous for the main P-Ko variant, 2 were homozygous for the $A_{41} \rightarrow S$ variant, 1 was homozygous for the $P_{36} \rightarrow S$ variant, and 9 were heterozygous P-Ko/P-Ko $A_{41} \rightarrow S$. Also, Ps-1 protein was frequently detected (56 out of 86 subjects), while the other *PRB1* products, P-E and IB-6 from *PRB1-S* allele and Ps-2 from *PRB1-L* allele, were rarely observed.

Products of *PRB2* Locus

Figure 3 shows the bPRPs deriving from the different alleles of *PRB2* locus detected in the Western population, namely, P-H (or IB-4), P-F, P-J, IB-1, IB-8a Con 1^-, and IB-8a Con 1^+.

The sequences, mass values, elution times, and detection frequencies of bPRPs encoded by *PRB2* locus are reported in

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 3. Average (Mav) and $[M + H]^{1+}$ Monoisotopic Masses, Elution Times, Frequency, and Sequence of the Principal Products of *PRB-2* Locus (UniProtKB Code P02812)[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ±0.4)	freq (n = 86)	sequence[b]
P-H $S_1 \rightarrow A$*	5574.0 ± 0.6 (5574.14)	5571.79 ± 0.02 (5571.788)	15.2	9	APPGKPQGPP QQEGNNPQGP PPPAGGNPQQ PQAPPAGQPQ GPPRPPQGGR PSRPPQ
P-H (or IB-4)	5590.1 ± 0.6 (5590.10)	5587.77 ± 0.02 (5587.783)	15.2	85	SPPGKPQGPP QQEGNNPQGP PPPAGGNPQQ PQAPPAGQPQ GPPRPPQGGR PSRPPQ
P-F (or IB-8c)	5842.5 ± 0.7 (5842.49)	5840.00 ± 0.02 (5839.992)	14.7	83	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGGSKSRS A
P-J	5943.6 ± 0.7 (5943.56)	5941.00 ± 0.02 (5941.003)	14.5	86	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSRS S
IB-1*	9593.4 ± 1.0 (9593.38)	9588.61 ± 0.04 (9588.703)	19.4	86	<QNLNEDVSQE ESPSLIAGNP QGAPPQGGNK PQGPPSPPGK PQGPPPQGGN QPQGPPPPPG KPQGPPPQGG NKPQGPPPPG KPQGPPPQGD KSRSPR
IB-8a Con1$^-$ P_{100}*	11897 ± 2 (11896.16)	11890.05 ± 0.05 (11890.035)	16.7	42	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSQS ARSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGGNKP QGPPPPGKPQ GPPPQGGSKS RSS
IB-8a Con1$^+$ S_{100}*	11887 ± 2 (11886.12)	11880.02 ± 0.05 (11880.014)	17.6	15	SPPGKPQGPP PQGGNQPQGP PPPPGKPQGP PPQGGNKPQG PPPPGKPQGP PPQGDNKSQS ARSPPGKPQG PPPQGGNQPQ GPPPPPGKPQ GPPPQGGNKS QGPPPPGKPQ GPPPQGGSKS RSS
IB-8a Con1$^+$ S_{100} Glycoform-1	13291 ± 2 (13290.42)	ND (13283.521)	15.6	7	IB-8a Con1$^+$ S_{100} sequence + $dHex_1$+Hex_4+$HexNAc_3$
IB-8a Con1$^+$ S_{100} Glycoform-2	13656 ± 2 (13655.76)	ND (13648.653)	15.6	15	IB-8a Con1$^+$ S_{100} sequence + $dHex_1$+Hex_5+$HexNAc_4$
IB-8a Con1$^+$ S_{100} Glycoform-3	13802 ± 2 (13801.90)	ND (13794.711)	15.6	23	IB-8a Con1$^+$ S_{100} sequence + $dHex_2$+Hex_5+$HexNAc_4$
IB-8a Con1$^+$ S_{100} Glycoform-4	13948 ± 2 (13948.04)	ND (13940.769)	15.6	32	IB-8a Con1$^+$ S_{100} sequence + $dHex_3$+Hex_5+$HexNAc_4$
IB-8a Con1$^+$ S_{100} Glycoform-5	14093 ± 2 (14094.18)	ND (14086.827)	15.6	19	IB-8a Con1$^+$ S_{100} sequence + $dHex_4$+Hex_5+$HexNAc_4$
IB-8a Con1$^+$ S_{100} Glycoform-6	14239 ± 2 (14240.33)	ND (14232.885)	15.6	2	IB-8a Con1$^+$ S_{100} sequence + $dHex_5$+Hex_5+$HexNAc_4$

[a]Peptides characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]<Q: pyro-glutamic acid; S: phosphorylated Ser. dHex: deoxy-hexose, probably fucose; Hex: hexose, probably mannose or galactose; HexNAc: *N*-acetyl-hexosamine, probably *N*-acetyl-glucosamine. NKS: N-glycosylation consensus sequence; ND, not determinable.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 4. Average (Mav) and $[M + H]^{1+}$ Monoisotopic Masses and Elution Times of the Main Derivatives of the Products of *PRB-2* Locus (UniProtKB code P02812)[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ±0.4)
P-H (Fr. 8-56)	4898.5 ± 0.5 (4898.34)	4896.42 ± 0.03 (4896.417)	17.7
P-H (Fr. 1-18)[b]	1856.9 ± 0.4 (1856.97)	1856.89 ± 0.02 (1856.89)	10.5
P-F Des $R_{59}SA_{61}$	5528.4 ± 0.6 (5528.19)	5525.81 ± 0.03 (5525.821)	16.3
P-J Des $R_{59}SS_{61}$	5613.4 ± 0.6 (5613.25)	5610.84 ± 0.03 (5610.838)	16.3
IB-1 (Fr. 33-42)[b]	961.1 ± 0.2 (961.09)	961.51 ± 0.01 (961.51)	13.4
IB-1 (Fr. 18-32)	1446.7 ± 0.2 (1446.54)	1446.71 ± 0.01 (1446.708)	8.1
IB-1 Des $R_{93}SPR_{96}$ pyro-Gln, S_8(phosph)*	9097.0 ± 1.0 (9096.88)	9092.42 ± 0.05 (9092.416)	19.1
IB-1 Des R_{91} pyro-Gln (N-term) S_8(phosph)*[c]	9437.0 ± 1.0 (9437.20)	9432.61 ± 0.05 (9432.602)	19.4

[a]Peptides characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]Identified also in ref 24. [c]Identified also in refs 3 and 16.

Table 3. High-resolution MS/MS analysis confirmed the sequences previously characterized[3,19] and allowed us to identify the new P-H $S_1 \rightarrow A$ variant not reported in UniProtKB database. MS/MS experiments performed by HCD fragmentation of the multiply charged ion $[M+5H]^{5+}$ (1115.57 m/z) confirmed the $S \rightarrow A$ substitution at position 1.

The two proteoforms of IB-8a detected in human saliva derive from an SNP responsible for $S_{100} \rightarrow P$ substitution.[10] IB-8a carrying P_{100} is not glycosylated and is named Con1$^-$ because it is not able to bind concanavalin A.[10] IB-8a Con1$^+$ carries a serine at position 100, and it may be glycosylated on N_{98}. The six different glycoforms of IB-8a Con1$^+$ characterized by HPLC−ESI−MS in adult human saliva together with the nonglycosylated protein are reported in Table 3.[24] Five of the glycosylated species carry a biantennary *N*-linked glycan fucosylated in the innermost *N*-acetylglucosammine of the core and show from zero to four additional fucoses in the antennal region. The sixth glycoform carries a monoantennary monofucosylated oligosaccharide. IB-8a was detected in 64 subjects (out of 86); 25 were homozygous for IB-8a Con 1$^-$ and 22 were homozygous for IB-8a Con 1$^+$, whereas 17 subjects exhibited both the variants. Among the 39 subjects expressing IB-8a Con 1$^+$, 24 showed only the glycosylated proteoforms, whereas 15 also showed the apoprotein. In the HPLC−ESI low-resolution MS analyses it was possible to determine the Mav of the glycoforms of IB-8a-Con1$^+$, but it was not possible to determine the monoisotopic mass value by deconvolution of the high-resolution ESI spectra. IB-1, P-J, and P-H were detected in all 86 samples analyzed, whereas P-F showed a slightly lower frequency (83 out of 86 subjects) (Table 3). The $S_1 \rightarrow A$ variant of P-H peptide was detected in whole saliva of nine adult subjects, with one of them homozygous for the $S_1 \rightarrow A$ variant and the other eight heterozygous for P-H/P-H $S_1 \rightarrow A$. Several peptides characterized in this study derived specifically from the fragmentation of bPRPs encoded by *PRB2* locus, and they are reported in Table 4. Among them, five peptides were identified for the first time in this study, whereas the other three have already been characterized in previous top-down investigations.[3,23,25]

Products of the *PRB3* Locus

Figure 4 shows the asset of the *PRB3* locus. The sequence and the possible glycosylation sites of the most common glycoproteins codified by *PRB3* locus, namely, Gl-1, Gl-2, and Gl-3, are reported in Table 5.

Each Gl protein carries a different number of putative *O*- and *N*-glycosylation sites depending on the length of the polypeptide backbone.[26,27] Because of the high heterogeneity of the glyco-moiety, ESI spectra of the intact glycoproteins are crowded with m/z signals and cannot be resolved by the deconvolution software. Therefore, until now, we were unable to detect masses pertaining to these proteins by a top-down platform. Surprisingly, the Gl-2 (or PRP-3M) glycoforms were the only bPRPs detectable in significant amounts in newborn whole saliva.[28] Characterization of Gl-2 glycoforms was performed by RP−HPLC−high-resolution ESI−MS before and after *N*-deglycosylation with PNGase F of an enriched fraction isolated from newborn saliva. Furthermore, peptides obtained by Glu-C digestion were submitted to MS/MS sequencing. In this way, it was possible to characterize the peptide backbone and to identify the *N*- and *O*-glycosylation sites. The heterogeneous mixture of the glycoforms derived from the combination of 8 different neutral and sialylated glycans *O*-linked to T_{34} and 33 different glycans *N*-linked to N_{50}, N_{71}, N_{92}, N_{113}, N_{134}, N_{155}, N_{176}, and N_{197} residues. It is plausible to assume that similar glycoforms of Gl-1 and Gl-3 exist by similarity.

Products of the *PRB4* Locus

Figure 5 shows the asset of *PRB4* locus. Among the products of this locus, only the two variants of P-D peptide, carrying either P or A at position 32, were detectable under our experimental conditions.

The other products of *PRB4* locus are highly glycosylated proteins until now not completely characterized, and their sequences (reported in Table 6) derive from gene sequencing.[12,29] As for the other glycosylated bPRPs, their ESI spectra were crowded for the heterogeneous glycan moieties, and it was not possible to establish their molecular masses by our MS platform. Table 6 reports the mass values, the sequences, and the elution times of the bPRPs encoded by *PRB4* locus and the frequencies determined in healthy adults. Top-down high-resolution MS/MS experiments allowed us to confirm the sequences of the two P-D variants already characterized by us[3,19] and to identify some P-D fragments, two of which described for the first time in this study. P-D peptide was detected in 75 subjects; 57 were homozygous for the main P-D variant, 5 were homozygous for the $P_{32} \rightarrow A$ variant, and 13 were heterozygous P-D/P-D $P_{32} \rightarrow A$ (Table 6). The sequences of the glycosylated proteins encoded by the three *PRB4* alleles reported in Table 6 were obtained from the literature[10,12] and from UniProt KB database (P10163).

Other Fragments of bPRPs

The sequences, mass values, elution times, and the possible origin of 34 bPRP fragments eluting in the bPRP chromatographic cluster are reported in Table 7. Among them, 21 have

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 5. Sequence and Potential Glycosylation Sites of the Products of *PRB3* Locus (UniProtKB code Q04118)

name	sequence[a]
Gl-3 or PRP-3S (5 N-glycosyl. sites)	<QSLNEDVSQE ESPSVISGKP EGRRPQGGNQ PQRTPPPGK PEGRPPQGGN QSQGPPPRPG KPEGPPPQGG NQSQGPPPRP GKPEGQPPQG GNQSQGPPPR PGKPEGPPPQ GGNQSQGPPP RPGKPEGPPP QGGNQSQGPP PRPGKPEGSP SQGGNKPQGP PPHPGKPQGP PPQEGNKPQR PPPPGRPQGP PPPGGNPQQP LPPPAGKPQG PPPPPQGGRP HRPPQGQPPQ
Gl-2 or PRP-3M (8 N-glycosyl. sites)	<QSLNEDVSQE ESPSVISGKP EGRPPQGGNQ PQRTPPPGK PEGPPPQGGN QSQGPPPRPG KPEGQPPQGG NQSQGPPPRP GKPEGPPPQG GNQSQGPPPR PGEPEGPPPQ GGNQSQGPPP HPGKPEGPPP QGGNQSQGPP PRPGKPEGPP PQGGNQSQGP PPRPGKPEGP PPQGGNQSQG PPPRPGKPEG PPPQGGNQSQ GPPPRPGKPE GSPSQGGNKP RGPPPHPGKP QGPPPQEGNK PQRPPPPRRP QGPPPPGGNP QQPLPPPAGK PQGPPPPPQG GRPHRPPQGQ PPQ
Gl-1 or PRP-3L (9 N-glycosyl. sites)	<QSLNEDVSQE ESPSVISGKP EGRRPQGGNQ PQRTPPPLGK PEGRPPQGGN QSQGPPPRPG KPEGPPPQGG NQSQGPPPRP GKPEGQPPQG GNQSQGPPPR PGKPEGPPPQ GGNQSQGPPP RPGEPEGPPP QGGNQSQGPP PHPGKPEGPP PQGGNQSQGP PPRPGKPEGP PPQGGNQSQG PPPRPGKPEG PPPQGGNQSQ GPPPRPGKPE GPPPQGGNQS QGPPPRPGKP EGSPSQGGNK PRGPPPHPGK PQGPPPQEGN KPQRPPPPRR PQGPPPPGGN PQQPLPPPAG KPQGPPPPPQ GGRPHRPPQG QPPQ

[a] <Q: pyro-glutamic acid; S: phosphorylated Ser; NQS: N-glycosylation site; T_{34} of Gl-2 is the *O*-glycosylation site.[28] The sequence and the glycosylation sites of Gl-2 have been experimentally confirmed.[28] The PTMs of Gl-3 and Gl-1 are hypothesized by similarity.

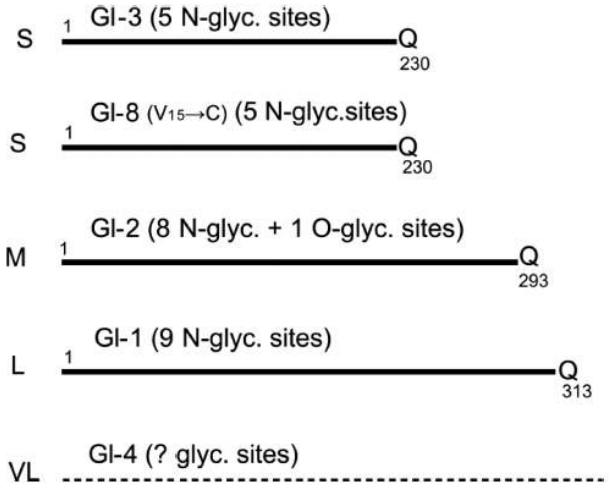


Figure 4. Schematic representation of the human salivary *PRB3* locus and their alleles, showing the coding regions for parent gPRPs.

never been detected in previous investigations, whereas the others have already been described.[3,20,23,30,31] Furthermore, a large number of very small and polar naturally occurring fragments of bPRPs eluting before the bPRP cluster were detected by HPLC–ESI–MS. Table 8 lists the 36 most abundant; 17 were never detected in previous investigations, whereas 19 were previously characterized.[3,20,23,31]

Fragments of Other Salivary Proteins That May Be Confused with Anomalous bPRPs

Several masses were often detected in the chromatographic cluster of bPRPs and characterized by our group as naturally occurring fragments deriving from other salivary proteins, mainly P-B peptide and aPRPs. These fragments usually detected in human adult saliva are listed in Table 9 and comprise 15 fragments never detected in previous investigations and 6 fragments already characterized in human saliva by other research groups.[20,23,32]

■ DISCUSSION

The top-down approach applied to the proteomic characterization of human saliva allowed us to highlight the great heterogeneity of the bPRP family, which, on the basis of the 55 new bPRPs characterized for the first time in the present study, accounts for a total number of 109 proteoforms confirmed by MS/MS sequencing. The heterogeneity of the parent bPRPs is really amazing, but the great similarity among some of them, evident by looking at the sequences reported in Tables 1, 3, 5, and 6, suggested the division of the bPRPs into two main groups and a third minor hybrid group (Figure 6).

The first group, which we named Group 1, includes P-E, P-Ko, IB-6, Ps-1, Ps-2, P-H, P-F, P-J, and P-D. The sequence of all of these bPRPs starts with the same SPPGKPQGPP motif, followed by sequences somewhat similar but showing small variations among the different components. The central part of the sequences shows similar repeats. Because P-E, IB-6, Ps-1, and Ps-2 sequences originate from DNA-length polymorphisms in exon 3 of the *PRB1* locus, they exhibit high similarity.[9,10,12] Whereas PRB1-S proprotein contains two convertase cleavage sites that generate II-2 (first cleavage), P-E, and IB-6 (second cleavage) (Figure 2), PRB1-M and L proproteins, due to the substitution $R_{131} \rightarrow Q$, which abolishes the second cleavage site, undergo only one convertase cleavage, which generates II-2 together with Ps-1 and Ps-2, respectively,

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292–3307

Table 6. Average (Mav) and $[M + H]^{1+}$ Monoisotopic Mass Values, Elution Times, Frequency, and Sequence of the Products of *PRB-4* Locus and their Derivatives (UniProtKB code P10163)[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ±0.4)	frequency (n = 86)	sequence[b]
P-D (Fr. 49–60)	1153.6 ± 0.2 (1153.32)	1153.61 ± 0.01 (1153.611)	15.7	31	GPPPPPQGGRPP
P-D (Fr. 1–18)[c]	1871.0 ± 0.3 (1871.05)	1870.94 ± 0.01 (1870.941)	13.5	4	SPPGKPQGPP QQEGNKPQ
P-D (Fr. 59–70)[d]	2242.2 ± 0.3 (2241.52)	2241.16 ± 0.02 (2241.164)	15.9	38	GPPPPPQGGRPPRPAQGQQPPQ
P-D (Fr. 1–27)[c,d]	2727.3 ± 0.4 (2727.06)	2726.42 ± 0.01 (2726.401)	15.4	23	SPPGKPQGPP QQEGNKPQGP PPPGKPQ
P-D (Fr. 1–60)*	5861.5 ± 0.7 (5861.58)	5859.00 ± 0.03 (5859.002)	15.2	3	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP PPGGNPQQPQ APPAGKPQGP PPPPQGGRPP
P-D (P_{32}→A)*	6923.6 ± 0.8 (6923.69)	6920.54 ± 0.04 (6920.538)	15.9	18	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP PAGGNPQQPQ APPAGKPQGP PPPPQGGRPP RPAQGQQPPQ
P-D (or IB-5)*	6950.0 ± 0.8 (6949.73)	6946.55 ± 0.04 (6946.554)	16.7	70	SPPGKPQGPP QQEGNKPQGP PPPGKPQGPP PPGGNPQQPQ APPAGKPQGP PPPPQGGRPP RPAQGQQPPQ
Glycos. Protein A	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGGN QSQGPPPPPG KPEGRPPQGG NQSQGPPPHP GKPERPPPQG GNQSQGTPPP PGKPERPPPQ GGNQSHRPPP PPGKPERPPP QGGNQSQGPP PHPGKPEGPP PQEGNKSRSA R
II-1	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGGN QSQGPPPPPG KPEGRPPQGG NQSQGPPPHP GKPERPPPQG GNQSQGTPPP PGKPEGRPPQ GGNQSQGPPP HPGKPERPPP QGGNQSHRPP PPPGKPERPP PQGGNQSQGP PPHPGKPEGP PPQEGNKSRS AR
Cd-IIg	ND	ND	ND	ND	<ESSSEDVSQE ESLFLISGKP EGRRPQGGNQ PQRPPPPPGK PQGPPPQGGN QSQGPPPPPG KPEGRPPQGG NQSQGPPPHP GKPERPPPQG GNQSQGPPPH PGKPESRPPQ GGHQSQGPPP TPGKPEGPPP QGGNQSQGTP PPPGKPEGRP PQGGNQSQGP PPHPGKPERP PPQGGNQSHR PPPPPGKPER PPPQGGNQSQ GPPPHPGKPE GPPPQEGNKS RSAR

[a]Peptides characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]<E: pyro-glutamic acid; S: phosphorylated Ser (hypothetical, by similarity). [c]Identified also in ref 23. [d]Identified also ref 20.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

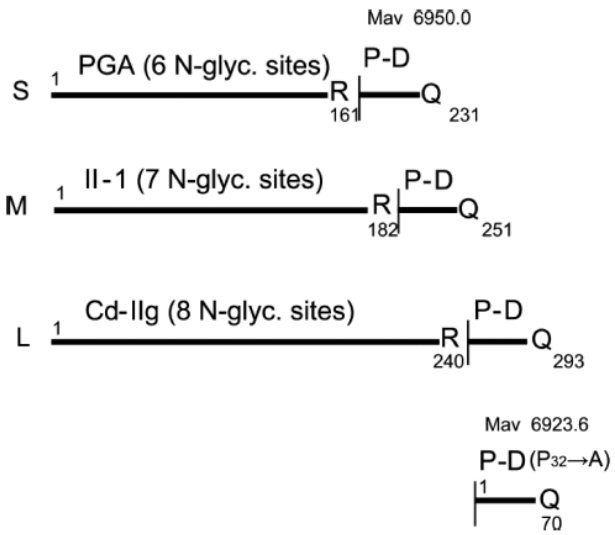


Figure 5. Schematic representation of the human salivary *PRB4* locus and their alleles, showing the coding regions for parent bPRPs and gPRPs.

as already suggested by Azen and coworkers.[9] The bPRP with a Mav of 10 433.5 Da, detected in whole saliva and in parotid secretory granules[3,16] and named P-Ko by Halgand et al.,[21] is encoded by *cP4*, a differentially spliced transcript of *PRB1-L* allele.[8] *cP4* proprotein lacks the sequence 106−299 of *PRB1-L* (P04280), and its cleavage generates II-2 peptide and P-Ko protein (Figure 2).

Group 2 includes IB-1, II-2, and the glycosylated bPRPs codified by *PRB3* and *PRB4* genes, namely, Gl-1, Gl-2, Gl-3, GPA, II-1, and Cd-IIg. Their sequences start with the similar motif (E/Q)XXXEDVSQEES, where XXX is LNE in IB-1, II-2, Gl-1, Gl-2, and Gl-3 and SSS in GPA, II-1, and Cd-IIg. The central part of the sequences comprises similar repeats with differences from the repeats of the members belonging to Group 1. The N-terminal glutamine of IB-1 and II-2 is converted to a pyro-glutamic acid moiety, and the serine at position 8 is phosphorylated for the presence of the SXE consensus sequence recognized by the Golgi casein kinase Fam20C,[33] responsible for the phosphorylation of all of the salivary peptides (aPRPs, histatin 1, statherin, and cystatin S). In a previous work, we demonstrated that in resemblance to IB-1 and II-2, the N-terminal glutamine of Gl-2 is converted to a pyro-glutamic acid moiety and serine at position 8 is phosphorylated.[28] Phosphorylation is an almost complete event because <1% of the nonphosphorylated forms can be detected in parotid granules, parotid, and whole saliva and probably occurs after the cleavage of the proprotein.[3] It can be supposed, by sequence similarity, that also Gl-1 and Gl-3 undergo the same PTMs, reported in Table 5 as hypothetical. The presence of a glutamic acid residue at the N-terminus of GPA, II-1, and Cd-IIg and the SQE consensus sequence (for Serine-8) suggests similar PTMs for these bPRPs, too, namely, the N-terminal pyro-E and phosphorylation of S_8. These PTMs are reported as hypothetical in Table 6. A second potential phosphorylation site at S_3 is present in the sequence of GPA, II-1, and Cd-IIg, but because of the absence of experimental evidence of this modification in this study and in literature, the phosphorylation of S_3 is not reported in Table 6. All of the glycosylated proteins of Group 2, after the initial sequence similar to IB-1 and II-2, contain a variable number of similar repeats characterized by the presence of the N-glycosylation consensus sequence NQS. Moreover, all of these glycosylated proteins show potential O-glycosylation sites. On the basis of structural differences, members of Group 2 can be divided in three subgroups: Group 2A, including IB-1 and II-2, without glycosylation sequons; Group 2B, including the Gl proteins codified by the alleles of *PRB3* locus; and Group 2C including the glycosylated proteins codified by the alleles of *PRB4* locus. Differently from the other *bPRP* loci, the pro-proteins expressed by the *PRB3* locus are not submitted to a proteolytic cleavage before secretion. Gl proteins are found in at least nine size variants in different populations.[11,34−36] In black and white populations, the four allelic size variants *S*, *M*, *L*, and *VL* encode for the corresponding Gl protein size variants Gl-4/*PRB3-VL* > Gl-1/*PRB3-L* > Gl-2/*PRB3-M* > Gl-3/*PRB3-S*.[11] The Gl-8 glycoprotein derives from a single nucleotide insertion in the *PRB3-S*^{Cys} allele, which converts R_{15} to C. Gl-8 protein is electrophoretically distinct from the other Gl protein variants because it forms a disulfide-bond heterodimer under the action of the salivary peroxidase.[34] In Table 5, only the three most common variants described in the Caucasian population are reported. The small Group 3 is a hybrid group, which includes the two proteoforms of IB-8a, Con1$^-$ and Con1$^+$. The initial sequence of these two proteins resembles that of Group 1, whereas the terminal sequence is similar to the repeat responsible for the glycosylation of the bPRPs of Groups 2B and 2C.

We never detected a putative PRB2-like Con2$^+$ protein in either the nonglycosylated or the glycosylated form. Indeed, it was reported that this protein, 60 residues long and encoded by a hybrid *PRB1-M CON2*$^+$ allele, had a single potential N-glycosylation site.[10]

We were able to characterize by MS/MS the structure of some variants of bPRPs. In particular, the characterization of P-H S_1 → A variant, previously detected by Kauffmann and colleagues[37] and attributed to the fragment 337−392 of the *PRB1-S* allele (corresponding to the fragment 63−118 of IB-6), resulted in not being correct from our data for two reasons: (a) IB-6 has a serine residue at position 63 instead of the alanine reported by Kauffmann; (b) in saliva of 9 subjects (out of 86) carrying this variant we never detected the complementary 1−62 fragment of IB-6. Moreover, we characterized two variants of P-Ko: the P_{36} → S variant identified for the first time in this study, detected in only one subject, and the A_{41} → S variant, detected in 11 out of 86 subjects, which corresponded to the 92−198 fragment of the sequence deposited at the UniProtKB human data bank with the code G5E9X6. This sequence, obtained from a large-scale genomic DNA investigation, is attributed to a polymorphism of *PRB1* locus that encodes for a pro-protein with a single convertase cleavage site from which II-2 and the P-Ko A_{41} → S variant are generated. The parent bPRPs reported in Tables 1, 3, 5, and 6 were submitted to naturally occurring fragmentations, and the peptide products were shown in Tables 2, 4, 7, and 8. The fragmentations observed on bPRPs can be divided into two types, those occurring before secretion and those occurring after secretion.[3] The first type commonly occurs at the C-terminal residues, and it is a widespread event observed in many secretory processes ascribed to specific carboxy-exopeptidases acting after the convertase cleavage. The postsecretory cleavage is mainly due to exogenous proteinases deriving from the oral microbiota and generates numerous small fragments recurrently found in whole saliva. Because of the great sequence similarities of bPRPs, it is

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 7. List of the Most Common Fragments from bPRPs Eluting in the bPRP Cluster (14.0−20.0 min) That Cannot Be Attributed to a Specific Parent bPRP[a]

sequence	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ± 0.4)	possible origin
QPLPPPAGKPQ[b]	1129.6 ± 0.2 (1129.34)	1129.64 ± 0.01 (1129.636)	15.7	Gl-3
GPPPPAGGNPQQPQ[b,c]	1341.7 ± 0.2 (1341.45)	1341.66 ± 0.01 (1341.655)	14.3	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPPGKPQGPPPQ	1350.7 ± 0.2 (1350.56)	1350.72 ± 0.02 (1350.716)	14.8	P-E, II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GGNQPQGPPPPPGKPQ[b]	1552.8 ± 0.2 (1552.73)	1552.79 ± 0.02 (1552.787)	15.2	IB-1, IB-6, P-F, P-J, P-E, Ps-1, Ps-2, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPRPPQGGRPSRPPQ	1680.9 ± 0.2 (1680.91)	1680.90 ± 0.02 (1680.904)	14.7	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPPGKPQGPPPQGDKS	1737.9 ± 0.3 (1737.95)	1737.89 ± 0.02 (1737.892)	14.0	II-2, P-E, Ps-1, Ps-2, IB-1
SPPGKPQGPPPQGGNQPQ[b,c,d,e,f]	1767.9 ± 0.3 (1767.92)	1767.89 ± 0.02 (1767.877)	14.3	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGKPQGPPAQGGSKSQ	1869.1 ± 0.3 (1869.08)	1868.96 ± 0.02 (1868.961)	17.2	IB-6, P-Ko, Ps-1, Ps-2
GPPPQGGNKPQGPPPPGKPQ[b,f]	1932.2 ± 0.4 (1932.17)	1932.01 ± 0.03 (1932.009)	14.7	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a $Con1^-$, IB-8a $Con1^+$
PQGGNKPQGPPPPGKPQGPP	1932.0 ± 0.4 (1932.17)	1932.01 ± 0.03 (1932.009)	14.5	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGGNPQQPLPPPAGKPQGPP	2028.3 ± 0.3 (2028.31)	2028.08 ± 0.02 (2028.066)	18.2	Gl-1, Gl-2, Gl-3
GPPPPGGNPQQPLPPPAGKPQ[b,e]	2028.3 ± 0.3 (2028.31)	2028.07 ± 0.02 (2028.066)	18.2	Gl-1, Gl-2, Gl-3
GPPPQGGNQPQGPPPPPGKPQ[b,f]	2029.2 ± 0.4 (2029.24)	2029.03 ± 0.03 (2029.025)	16.0	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PQGGNQPQGPPPPPGKPQGPP	2029.4 ± 0.3 (2029.26)	2029.03 ± 0.02 (2029.025)	16.0	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPPGKPQGPPPQGGNKPQ	2029.4 ± 0.3 (2029.30)	2029.06 ± 0.02 (2029.061)	14.8	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGKPQGPPPQGGNKPQGPPP	2029.4 ± 0.3 (2029.30)	2029.06 ± 0.02 (2029.061)	14.9	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
GPPPPGKPQGPPPQGDKSRSP	2078.3 ± 0.3 (2078.33)	2078.08 ± 0.02 (2078.078)	14.5	II-2, P-E, Ps-1, Ps-2, IB-1
GPPPQEGNKPQRPPPPGRPQ	2132.1 ± 0.3 (2131.40)	2131.12 ± 0.02 (2131.115)	14.6	GL-3
GPPPPPQGGRPHRPPQGQPPQ[e]	2180.1 ± 0.3 (2179.45)	2179.13 ± 0.02 (2179.127)	14.9	Gl-1, Gl-2, Gl-3
GPPPPAGGNPQQPQAPPAGQPQGPP[b]	2339.6 ± 0.3 (2339.57)	2339.15 ± 0.02 (2339.153)	18.1	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPPQGGNQPQGPPPPPGKPQ[d]	2721.0 ± 0.4 (2721.05)	2720.40 ± 0.03 (2720.390)	16.3	P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$
PPGKPQGPPPQGGNKPQGPPPPGKPQGPPP	2885.7 ± 0.5 (2885.31)	2884.52 ± 0.03 (2884.522)	16.1	II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a $Con1^-$, IB-8a $Con1^+$

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 7. continued

sequence	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ± 0.4)	possible origin
PPPGKPQGPPPQGGNKPQGPPPPGKPQGPP	2885.7 ± 0.5 (2885.31)	2884.54 ± 0.03 (2884.522)	16.1	II-2, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPPQGGNKPQGPPPPGKPQGPPPQGDKSRSP[d]	3136.5 ± 0.6 (3136.50)	3135.61 ± 0.03 (3135.608)	15.2	II-2, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPPPGGNPQQPLPPPAGKPQGPPPPPQGGRPH	3203.7 ± 0.6 (3203.64)	3202.67 ± 0.03 (3202.666)	18.1	Gl-1, Gl-2, Gl-3
GPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPP	3486.7 ± 0.6 (3486.75)	3485.66 ± 0.03 (3485.658)	18.4	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPPQGGNQPQGPPPPPGKPQGPPPQGGNKPQ*	3779.2 ± 0.6 (3779.22)	3777.92 ± 0.04 (3777.921)	16.4	IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPPPGGNPQQPLPPPAGKPQGPPPPPQGGRPHRPPQGQPPQ[c]	4190.2 ± 0.7 (4189.71)	4188.17 ± 0.04 (4188.175)	18.1	Gl-1, Gl-2, Gl-3
SPPGKPQGPPPQGGNQPQGPPPPPGKPQGPPPQGGNKPQGPPPPGKPQ	4635.2 ± 0.8 (4635.22)	4633.38 ± 0.05 (4633.381)	16.9	IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPPRPPQGGRPSRPPQ	4898.4 ± 0.8 (4898.35)	4896.44 ± 0.05 (4896.417)	17.9	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
QGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPPRPPQGGRPSRPP	4898.4 ± 0.8 (4898.35)	4896.44 ± 0.05 (4896.417)	17.9	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SPPGKPQGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPPRPPQGGRPSRPP	5462.7 ± 0.9 (5462.01)	5459.73 ± 0.06 (5459.724)	18.1	IB-6, P-H S_1
PPPGKPQGPPPQGGNKPQGPPPPGKPQGPPAQGGSKSQSARAPPGKPQGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQ	7611.7 ± 1.3 (7611.42)	7607.75 ± 0.07 (7607.820)	20.5	Ps-1, Ps-2
PQGGNKPQGPPPPGKPQGPPAQGGSKSQSARAPPGKPQGPPQQEGNNPQGPPPPAGGNPQQPQAPPAGQPQGPPRPPQ	7613.7 ± 1.3 (7613.39)	7609.74 ± 0.07 (7609.811)	20.4	P-Ko, Ps-1, Ps-2

[a]Peptides characterized for the first time in this study are reported in bold. The proteoform identified only by manual inspection of MS/MS spectra is labeled with an asterisk. [b]Identified also in ref 23 [c]Identified also in ref 31. [d]Identified also in ref 30. [e]Identified also in ref 20. [f]Identified also in ref 3.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 8. List of the Most Abundant Naturally Occurring Fragments of bPRPs Eluting before the bPRP Cluster[a]

sequence	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ± 0.4)	possible origin
PQGPPPQ*[b]	719.8 ± 0.2 (719.80)	720.37 ± 0.01 (720.368)	8.1	II-2, IB-1, Gl-1, II-1, CDII-g, Glycosyl. Pr. A
PPPPGKPQ[c]	816.8 ± 0.2 (816.96)	817.46 ± 0.01 (817.466)	4.9	P-E, IB-6, II-2, P-Ko, Ps-1, Ps-2, P-F, P-J, P-D P_{32}, P-D A_{32}, IB-1, IB-8a Con1$^-$, IB-8a Con1$^+$, Glycosyl. Pr. A, II-1, CD-IIg
PPPPGRPQ	844.7 ± 0.2 (844.97)	845.46 ± 0.01 (845.426)	10.9	Gl-3
GPPPPGKPQ[b,c]	874.3 ± 0.2 (874.01)	874.48 ± 0.01 (874.478)	5.5	II-2, P-E, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$, P-D P_{32}, P-D A_{32}
PPPPPGKPQ[c]	914.2 ± 0.2 (914.07)	914.51 ± 0.01 (914.509)	8.5	II-2, P-E, IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$, Glycosyl. Pr. A, II-1, CD-IIg
GPPPPGGNPQ[d]	917.0 ± 0.2 (916.99)	917.45 ± 0.01 (917.448)	7.5	P-D, Gl-3, Gl-2, Gl-1
GGRPSRPPQ	951.1 ± 0.2 (951.05)	951.51 ± 0.01 (951.512)	4.3	P-Ko, IB-6, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPPPGKPQ[b,c,e]	971.3 ± 0.2 (971.12)	971.53 ± 0.01 (971.531)	12.0	II-2, P-E, IB-6, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPPPPGKPE	972.0 ± 0.2 (972.11)	972.52 ± 0.01 (972.515)	12.8	Glycosyl. Pr. A, II-1, CD-IIg
GPPPHPGKPQ[d,e]	1011.3 ± 0.2 (1011.15)	1011.38 ± 0.01 (1011.537)	5.8	Gl-1, Gl-2, Gl-3,
GPPPHPGKPE[d]	1012.4 ± 0.2 (1012.13)	1012.52 ± 0.01 (1012.521)	7.3	Gl-1, Gl-2, Glycosyl. Pr. A, II-1, CD-IIg
SPQSPPGKPQ	1022.0 ± 0.2 (1022.13)	1022.53 ± 0.01 (1022.526)	6.5	Ps-1, Ps-2
GPPPRPGKPE	1031.3 ± 0.2 (1031.18)	1031.56 ± 0.01 (1031.563)	8.1	Gl-1, Gl-2, Gl-3
SPRSPPGKPQ	1050.1 ± 0.2 (1050.18)	1050.57 ± 0.01 (1050.569)	4.7	Ps-1, Ps-2
RPPPPPGKPQ[b]	1070.6 ± 0.2 (1070.26)	1070.61 ± 0.01 (1070.611)	9.5	Glycosyl. Pr. A, II-1, CDII-g
GPPPQGGNQPQ[b,c,d]	1076.4 ± 0.2 (1076.13)	1076.51 ± 0.01 (1076.512)	4.6	P-E, IB-6, Ps-1, Ps-2, IB-1, P-J,
GPPPQGGNKPQ[b,d,e]	1076.3 ± 0.2 (1076.18)	1076.55 ± 0.01 (1076.548)	4.7	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-J, P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
RPAQGQQPPQ	1106.5 ± 0.2 (1106.21)	1106.57 ± 0.01 (1106.570)	5.0	P-D P_{32}, P-D A_{32}
GPPQQGGNRPQ	1135.3 ± 0.2 (1135.20)	1135.56 ± 0.01 (1135.560)	4.5	Ps-1, Ps-2
GPPPQEGNKPQ	1148.0 ± 0.2 (1148.24)	1148.57 ± 0.01 (1148.569)	4.5	Gl-1, Gl-2, Gl-3
GPPQQEGNNPQ[b,d]	1165.5 ± 0.2 (1165.18)	1165.52 ± 0.01 (1165.523)	5.6	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPQQEGNKPQ	1179.5 ± 0.2 (1179.25)	1179.58 ± 0.01 (1179.575)	4.3	P-D P_{32}, P-D A_{32}
SQGTPPPPGKPE[d]	1191.1 ± 0.2 (1191.31)	1191.60 ± 0.01 (1191.600)	13.1	Glycosyl. Pr. A, II-1, CDII-g
GPPPPPQGGRPH[e]	1193.4 ± 0.2 (1193.34)	1193.62 ± 0.01 (1193.617)	9.4	Gl-1, Gl-2, Gl-3
PQGPPPPPGKPQ	1196.5 ± 0.2 (1196.37)	1196.65 ± 0.01 (1196.642)	13.9	II-1, P-E, IB-6, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPRPPQGGRPS	1202.6 ± 0.2 (1202.34)	1202.64 ± 0.01 (1202.639)	13.4	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
GPPPQGDKSRSP[b]	1222.6 ± 0.2 (1222.32)	1222.62 ± 0.01 (1222.617)	4.3	II-2, P-E, Ps-1, Ps-2, IB-1
SQGPPPHPGKPE[d]	1227.4 ± 0.2 (1227.34)	1227.61 ± 0.01 (1227.612)	11.9	Gl-1, Gl-2, Gl-3, Glycosyl. Pr. A, II-1, CDII-g
SQGPPPRPGKPE	1246.7 ± 0.2 (1246.39)	1246.65 ± 0.01 (1246.654)	12.3	Gl-1, Gl-2, Gl-3
PPQGGRPSRPPQ[c]	1273.1 ± 0.2 (1273.42)	1273.68 ± 0.01 (1273.676)	11.0	IB-6, P-Ko, Ps-1, Ps-2, P-H S_1, P-H A_1
SHRPPPPPGKPE	1295.6 ± 0.2 (1295.46)	1295.69 ± 0.01 (1295.685)	8.9	Glycosyl. Pr. A, II-1, CDII-g
GGNKPQGPPPPGKPQ	1455.8 ± 0.2 (1455.64)	1455.77 ± 0.01 (1455.770)	12.9	II-2, IB-6, P-Ko, Ps-1, Ps-2, IB-1, P-F, P-J, IB-8a Con1$^-$, IB-8a Con1$^+$
GPPPPGKPQGPPPQGGSKS	1766.9 ± 0.3 (1766.97)	1766.92 ± 0.02 (1766.918)	13.8	P-F, IB-8a Con1$^-$, IB-8a Con1$^+$
SPPGKPQGPPQQEGNKPQ[e]	1870.9 ± 0.3 (1871.04)	1870.94 ± 0.03 (1870.941)	13.8	P-D P_{32}, P-D A_{32}
GPPPQGDKSQSPRSPPGKPQ[d]	2042.1 ± 0.4 (2042.24)	2042.05 ± 0.03 (2042.041)	13.1	Ps-1, Ps-2
GPPPQGDKSRSPQSPPGKPQ	2042.1 ± 0.4 (2042.24)	2042.04 ± 0.03 (2042.041)	13.0	Ps-1, Ps-2

[a]Peptides characterized for the first time in this study are reported in bold. The proteoform identified only by manual inspection of MS/MS spectra is labeled with an asterisk. [b]Identified also in ref 3. [c]Identified also in ref 31 [d]Identified also in ref 23. [e]Identified also in ref 20.

impossible to establish the parent protein of the fragments reported in Tables 7 and 8. Many of these peptides terminate with a KPQ sequence, and this finding allowed the research group of Oppenheim to characterize a glutamine endoproteinase from *Rothia* species bacteria as responsible for this cleavage.[38]

Twenty-one peptides/proteins eluting in the chromatographic range of the bPRP cluster were identified as fragments of other salivary proteins (Table 9). Indeed, almost all of the human secreted salivary proteins are submitted to proteolysis by various proteinases acting before, during, and after glandular secretion.[3,19] The fragments shown in Table 9 derived mainly from aPRPs, P-C and P-B salivary peptides. It is important to recall that P-C and P-B peptides were sometimes ascribed to the bPRP family. However, P-C is a peptide of 44 amino acid residues resulting from the cleavage of PRP-1, PRP-2, Pif-s, and Db-s proteoforms of aPRPs; therefore, it must be considered a member of the aPRP family. P-B peptide is the product of *PROL3* gene (PBI; http://www.ensembl.org/Homo_sapiens/, ENSG00000171201) localized on chromosome 4q13.3, very close to the statherin gene. It shows high-sequence homology with statherin, and, as statherin, displays some tyrosine residues in its sequence (completely absent in bPRP family). As statherin, it is secreted from both parotid and submandibular/sublingual glands, and it does not derive from the cleavage of a bigger pro-protein. For all of these reasons, it has to be considered a member of the statherin family.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

Table 9. List of the Most Common Peptides or Fragments of Proteins, Which Elute in the bPRP Cluster and Might Be Confused with Anomalous bPRPs[a]

name	exp. Mav (theor.)	exp. $[M + H]^{1+}$ (theor.)	elution time (min ±0.4)	sequence[b]
P-B Fr. 37-45[c,d]	948.1 ± 0.1 (948.14)	948.52 ± 0.01 (948.519)	20.4	IPPPPPAPY
P-B Fr. 24-32[d,e]	960.1 ± 0.2 (960.15)	960.52 ± 0.01 (960.519)	19.7	VPPPPPPPY
P-B Fr. 23-32[e]	1107.2 ± 0.2 (1107.25)	1107.57 ± 0.01 (1107.569)	17.2	FVPPPPPPPY
P-B Fr. 33-45[c,d,e]	1315.6 ± 0.2 (1315.55)	1315.72 ± 0.01 (1315.716)	20.0	GPGRIPPPPP APY
aPRP Fr. 31-44[e]	1436.6 ± 0.2 (1436.56)	1436.72 ± 0.01 (1436.724)	14.7	RQGPPLGGQQ SQPS
P-C Fr. 15-35	2040.3 ± 0.3 (2040.33)	2040.08 ± 0.01 (2040.077)	15.5	GPPPPPPGKP QGPPPQGGRP Q
aPRP Fr. 77-105[d]	2938.3 ± 0.4 (2938.24)	2937.47 ± 0.01 (2937.473)	15.7	GPPQQGGHPP PPQGRPQGPP QQGGHPRPP
aPRP Fr. 67-105	3922.4 ± 0.5 (3922.371)	3921.00 ± 0.02 (3920.992)	16.2	GPPPPQGKPQ GPPQQGGHPP PPQGRPQGPP QQGGHPRPP
aPRP Fr. 50-106*	5852.4 ± 1.1 (5852.367)	5849.87 ± 0.03 (5849.875)	16.5	D DGPQQGPPQQ GGQQQQGPPP QGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr. 44-105	6238.7 ± 1.2 (6239.69)	6236.97 ± 0.03 (6236.967)	16.3	AGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr. 29-93*	6580.0 ± 1.2 (6580.95)	6577.12 ± 0.03 (6577.126)	17.2	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQ
aPRP Fr. 40-105	6638.1 ± 1.2 (6638.09)	6635.16 ± 0.03 (6636.158)	16.7	S QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr. 31-105*	7501.0 ± 1.4 (7501.04)	7498.60 ± 0.04 (7498.572)	17.4	QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr. 29-105*	7786.3 ± 1.6 (7786.35)	7782.73 ± 0.04 (7782.732)	18.2	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr. 18-93*	7853.1 ± 1.6 (7853.14)	7849.55 ± 0.04 (7849.545)	19.8	DGG DSEQFIDEER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQ
aPRP Fr. 29-106	7942.9 ± 1.6 (7943.53)	7938.8 ± 0.04 (7938.833)	16.9	ER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr. 26-106	8300.9 ± 1.7 (8300.88)	8296.96 ± 0.04 (8297.011)	17.5	IDEER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr. 18-105*	9060.2 ± 1.8 (9061.47)	9055.20 ± 0.05 (9056.134)	19.6	DGG DSEQFIDEER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPP
aPRP Fr. 18-106*	9216.3 ± 1.8 (9216.67)	9211.30 ± 0.05 (9211.251)	19.1	DGG DSEQFIDEER QGPPLGGQQS QPSAGDGNQN DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPR
aPRP Fr. 37-150*	11460.576 (11461.397)	11454.56 ± 0.06 (11454.563)	19.8	GQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPRGRPQ GPPQQGGHQQ GPPPPPPGKP QGPPPQGGRP QGPPQGQSP
aPRP Fr. 29-150*	12296.0 ± 2 (12296.33)	12289.03 ± 0.06 (12288.998)	19.0	ER QGPPLGGQQS QPSAGDGNQD DGPQQGPPQQ GGQQQQGPPP PQGKPQGPPQ QGGHPPPPQG RPQGPPQQGG HPRPPRGRPQ GPPQQGGHQQ GPPPPPPGKP QGPPPQGGRP QGPPQGQSPQ

[a]UniProtKB code is P02814 for P-B fragments and P02810 for P-C and aPRP fragments. The peptides/proteins characterized for the first time in this study are reported in bold. The proteoforms identified only by manual inspection of MS/MS spectra are labeled with an asterisk. [b]S: phosphorylated Ser. [c]Identified also in ref 31. [d]Identified also in ref 20. [e]Identified also in ref 33.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307

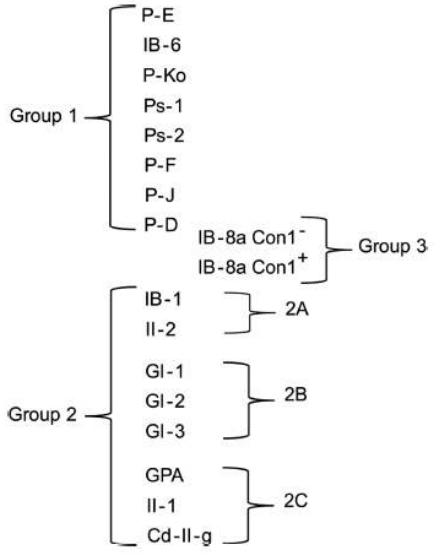


Figure 6. Schematic classification of the parent bPRPs based on sequence similarity.

As shown in this survey, polymorphisms and PTMs generate a high number of proteins/peptides with rather similar structures. Being the naturally occurring proteolytic cleavage, the most represented event, top-down proteomics, represents a key tool for the characterization of the bPRP complexity. The meaning of this amazing complexity is still largely obscure. Salivary proline-rich proteins are highly conserved in mammalian saliva, although significant structural differences are present in the class, suggesting that they play a crucial role in the oral protection. Some bPRPs exhibit the ability to bind harmful tannins,[39] others have the ability to modulate the oral flora,[40] and some others are involved in bitter taste perception.[22] Some bPRP fragments are involved in enamel pellicle formation,[14] and others act as antagonists of the progesterone-induced cytosolic Ca^{2+} mobilization.[41] The intrinsic propensity of some fragments to adopt a polyproline-II helix arrangement joined to PxxP motifs was suggestive of the interaction with the SH3 domain family.[42] Interestingly, interactions were highlighted[41] with Fyn, Hck, and c-Src SH3 domains, which are included in the Src kinases family, suggesting that some basic bPRPs can be involved in the signal transduction pathways modulated by these kinases. Only a small number of data on correlations between genes of basic PRPs and diseases linked to their allelic variants have been reported so far. In fact, for some of the alleles (PRB1VL, PRB2S, M, VL, PRB3VL) the genetic sequence is not reported. Moreover, for the small and large alleles of PRB1, the genetic sequence is incomplete[9] because the reference genome (NCBI Gene ID: 5542) encodes the medium allele. Regarding the primary structure of bPRP alleles, in the UniProtKB database (accession number P04280) is deposited the full amino acid sequence of the large variant, deduced through experimental evidence at the protein level. In human, bPRPs are secreted only by parotid glands, and this regioselectivity is puzzling. Moreover, their expression appears to be related to human growth, with different trends among the several bPRPs.[43]

■ CONCLUSIONS

Although various aspects of bPRPs still have to be defined, this survey may be considered an updated reference for the peptides included in this family. For all of the above-mentioned reasons, we hope that the information presented in this study on human salivary bPRPs might facilitate future studies devoted to establishing the specific roles of the different components of this complex family of proteins.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +39 0706754507. Fax: +39 6754523. E-mail: olianas@unica.it.

ORCID

Alessandra Olianas: 0000-0003-4238-3233
Barbara Manconi: 0000-0002-2880-9915
Tiziana Cabras: 0000-0001-7535-9825

Author Contributions

Manuscript was designed by A.P., M.C., I.M., and T.C. and supported and written with the contribution of all of the authors. Data analysis was performed by T.C., R.O., C.D., B.M., M.B., A.O., M.T.S., and F.I. Experimental procedure was performed by B.L., F.I., M.B., and B.M. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.
Mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (http://www.ebi.ac.uk/pride) with the data set identifier PXD009813.

■ ACKNOWLEDGMENTS

T.C. received funds from Cagliari University (FIR-2016), M.C. received funds from Catholic University of Rome (R4124500561), Nando Peretti Foundation (2011/28), and National Research Council of Italy (CNR) (DSB.AD004.077).

■ REFERENCES

(1) Bennick, A. Salivary proline-rich proteins. *Mol. Cell. Biochem.* **1982**, *45*, 83−99.
(2) Oppenheim, F. G.; Salih, E.; Siqueira, W. L.; Zhang, W.; Helmerhorst, E. J. Salivary proteome and its genetic polymorphisms. *Ann. N. Y. Acad. Sci.* **2007**, *1098*, 22−50.
(3) Messana, I.; Cabras, T.; Pisano, E.; Sanna, M. T.; Olianas, A.; Manconi, B.; Pellegrini, M.; Paludetti, G.; Scarano, E.; Fiorita, A.; Agostino, S.; Contucci, A. M.; Calò, L.; Picciotti, P. M.; 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*, 911−926.
(4) Azen, E. A.; Goodman, P. A.; Lalley, P. A. Human salivary proline-rich protein genes on chromosome 12. *Am. J. Hum. Genet.* **1985**, *37*, 418−424.
(5) Mamula, P. W.; Heerema, N. A.; Palmer, C. G.; Lyons, K. M.; Karn, R. C. Localization of the human salivary protein complex (SPC) to chromosome band 12p13.2. *Cytogenet. Genome Res.* **1985**, *39*, 279−284.
(6) Scherer, S. E.; Muzny, D. M.; Buhay, C. J.; Chen, R.; Cree, A.; et al. The finished DNA sequence of human Chromosome 12. *Nature* **2006**, *440*, 346−351.
(7) Lyons, K. M.; Azen, E. A.; Goodman, P. A.; Smithies, O. Many protein products from a few loci: Assignment of human salivary proline-rich proteins to specific loci. *Genetics* **1988**, *120*, 255−265.
(8) Maeda, N.; Kim, H. S.; Azen, E. A.; Smithies, O. Differential RNA splicing and post-translational cleavages in the human salivary proline-rich protein gene system. *J. Biol. Chem.* **1985**, *260*, 11123−11130.

(9) Azen, E. A. Genetics of salivary protein polymorphisms. *Crit. Rev. Oral Biol. Med.* **1993**, *4*, 479−485.
(10) Azen, E. A.; Amberger, E.; Fisher, S.; Prakobphol, A.; Niece, R. L. PBR1, PBR2, and PRB4 coded polymorphisms among human salivary concanavalin-A binding, II-1, and Po proline-rich proteins. *Am. J. Hum. Genet.* **1996**, *58*, 143−153.
(11) Lyons, K. M.; Stein, J. H.; Smithies, O. Length polymorphisms in human proline-rich protein genes generated by intragenic unequal crossing over. *Genetics* **1988**, *120*, 267−278.
(12) 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*, 753−770.
(13) Chan, M.; Bennick, A. Proteolytic processing of a human salivary proline-rich protein precursor by proprotein convertases. *Eur. J. Biochem.* **2001**, *268*, 3423−3431.
(14) Vitorino, R.; Calheiros-Lobo, M. J.; Williams, J.; Ferrer-Correia, A. J.; Tomer, K. B.; Duarte, J. A.; Domingues, P. M.; Amado, F. M. Peptidomic analysis of human acquired enamel pellicle. *Biomed. Chromatogr.* **2007**, *21*, 1107−1117.
(15) Siqueira, W. L.; Oppenheim, F. G. Small molecular weight proteins/peptides present in the in vivo formed human acquired enamel pellicle. *Arch. Oral Biol.* **2009**, *54*, 437−444.
(16) Messana, I.; Cabras, T.; Inzitari, R.; Lupi, A.; Zuppi, C.; Olmi, C.; Fadda, M. B.; 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*, 792−800.
(17) Zhang, Z.; Marshall, A. G. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 225−233.
(18) Vizcaíno, J. A.; Csordas, A.; del-Toro, N.; Dianes, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Ternent, T.; Xu, Q. W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.* **2016**, *44*, D447−D456.
(19) Castagnola, M.; Cabras, T.; Iavarone, F.; Vincenzoni, F.; Vitali, A.; Pisano, E.; Nemolato, S.; Scarano, E.; Fiorita, A.; Vento, G.; Tirone, C.; Romagnoli, C.; Cordaro, M.; Paludetti, G.; Faa, G.; Messana, I. Top-down platform for deciphering the human salivary proteome. *J. Matern.-Fetal Neonat. Med.* **2012**, *25*, 27−43.
(20) Vitorino, R.; Barros, A.; Caseiro, A.; Domingues, P. J.; Duarte, J.; Amado, F. Towards defining the whole salivary peptidome. *Proteomics: Clin. Appl.* **2009**, *3*, 528−540.
(21) Halgand, F.; Zabrouskov, V.; Bassilian, S.; Souda, P.; Loo, J. A.; Faull, K. F.; Wong, D. T.; Whitelegge, J. P. Defining intact protein primary structures from saliva: a step toward the human proteome project. *Anal. Chem.* **2012**, *84*, 4383−4395.
(22) Cabras, T.; Melis, M.; Castagnola, M.; Padiglia, A.; Tepper, B. J.; Messana, I.; Tomassini Barbarossa, I. Responsiveness to 6-n-propylthiouracil (PROP) is associated with salivary levels of two specific basic proline-rich proteins in humans. *PLoS One* **2012**, *7*, e30962.
(23) Helmerhorst, E. J.; Sun, X.; Salih, E.; Oppenheim, F. G. Identification of Lys-Pro-Gln as a novel cleavage site specificity of saliva-associated proteases. *J. Biol. Chem.* **2008**, *283*, 19957−19966.
(24) Cabras, T.; Boi, R.; Pisano, E.; Iavarone, F.; Fanali, C.; Nemolato, S.; Faa, G.; Castagnola, M.; Messana, I. HPLC−ESI−MS and MS/MS structural characterization of multifucosylated N-glycoforms of the basic proline-rich protein IB-8a $CON1^{+}$ in human saliva. *J. Sep. Sci.* **2012**, *35*, 1079−1086.
(25) Manconi, B.; Castagnola, M.; Cabras, T.; Olianas, A.; Vitali, A.; Desiderio, C.; Sanna, M. T.; Messana, I. The intriguing heterogeneity of human salivary proline-rich proteins: Short title: Salivary proline-rich protein species. *J. Proteomics* **2016**, *134*, 47−56.
(26) Carpenter, G. H.; Proctor, G. B. O-linked glycosylation occurs on basic parotid salivary proline-rich proteins. *Oral Microbiol. Immunol.* **1999**, *14*, 309−315.
(27) Gillece-Castro, B. L.; Prakobphol, A.; Burlingame, A. L.; Leffler, H.; Fisher, S. J. Structure and bacterial receptor activity of a human salivary proline-rich glycoprotein. *Arch. Biochem. Biophys.* **1991**, *266*, 17358−17368.
(28) Manconi, B.; Cabras, T.; Sanna, M.; Piras, V.; Liori, B.; Pisano, E.; Iavarone, F.; Vincenzoni, F.; Cordaro, M.; Faa, G.; Castagnola, M.; Messana, I. N- and O-linked glycosylation site profiling of the human basic salivary proline-rich protein 3M. *J. Sep. Sci.* **2016**, *39*, 1987−1997.
(29) Kauffman, D. L.; Keller, P. J.; Bennick, A.; Blum, M. Alignment of amino acid and DNA sequences of human proline-rich proteins. *Crit. Rev. Oral Biol. Med.* **1993**, *4*, 287−292.
(30) Vitorino, R.; Alves, R.; Barros, A.; Caseiro, A.; Ferreira, R.; Lobo, M. C.; Bastos, A.; Duarte, J.; Carvalho, D.; Santos, L. L.; Amado, F. L. Finding new posttranslational modifications in salivary proline-rich proteins. *Proteomics* **2010**, *10*, 3732−3742.
(31) Huq, N. L.; Cross, K. J.; Ung, M.; Myroforidis, H.; Veith, P. D.; Chen, D.; Stanton, D.; He, H.; Ward, B. R.; Reynolds, E. C. A Review of the Salivary Proteome and Peptidome and Saliva-derived Peptide Therapeutics. *Int. J. Pept. Res. Ther.* **2007**, *13*, 547−564.
(32) Hardt, M.; Thomas, L. R.; Dixon, S. E.; Newport, G.; Agabian, N.; Prakobphol, A.; Hall, S. C.; Witkowska, H. E.; Fisher, S. J. 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*, 2885−2899.
(33) Tagliabracci, V. S.; Engel, J. L.; Wen, J.; Wiley, S. E.; Worby, C. A.; Kinch, L. N.; Xiao, J.; Grishin, N. V.; Dixon, J. E. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science* **2012**, *336*, 1150−1153.
(34) Azen, E. A.; Minaguchi, K.; Latreille, P.; Kim, H. S. Alleles at the PRB3 locus coding for a disulfide-bonded human salivary proline-rich glycoprotein (Gl 8) and a null in an Ashkenazi Jew. *Am. J. Hum. Genet.* **1990**, *47*, 686−697.
(35) Azen, E. A.; Hurley, C. K.; Denniston, C. Genetic polymorphism of the major parotid salivary glycoprotein (Gl) with linkage to the genes for Pr, Db, and Pa. *Biochem. Genet.* **1979**, *17*, 257−279.
(36) Minaguchi, K.; Takaesu, Y.; Tsutsumi, T.; Suzuki, K. Studies of genetic markers in human saliva. (VII). Frequencies of the major parotid salivary glycoprotein (GI) system in a Japanese population. *Bull. Tokyo Dent. Coll.* **1981**, *22*, 1−6.
(37) Robinson, R.; Kauffman, D. L.; Waye, M. M.; Blum, M.; Bennick, A.; Keller, P. J. Primary structure and possible origin of the non-glycosylated basic proline-rich protein of human submandibular/sublingual saliva. *Biochem. J.* **1989**, *263*, 497−503.
(38) Zamakhchari, M.; Wei, G.; Dewhirst, F.; Lee, J.; Schuppan, D.; Oppenheim, F. G.; Helmerhorst, E. J. Identification of Rothia bacteria as gluten-degrading natural colonizers of the upper gastro-intestinal tract. *PLoS One* **2011**, *6*, e24455.
(39) Carlson, D. M. Proline-rich proteins and glycoproteins: expression of salivary gland multigene families. *Biochimie* **1988**, *70*, 1689−1695.
(40) Ruhl, S.; Sandberg, A. L.; Cisar, J. O. Salivary receptors for the proline-rich protein-binding and lectin-like adhesins of oral actinomyces and streptococci. *J. Dent. Res.* **2004**, *83*, 505−510.
(41) Palmerini, C. A.; Mazzoni, M.; Radicioni, G.; Marzano, V.; Granieri, L.; Iavarone, F.; Longhi, R.; Messana, I.; Cabras, T.; Sanna, M. T.; Castagnola, M.; Vitali, A. Antagonistic Effect of a Salivary Proline-Rich Peptide on the Cytosolic Ca^{2+} Mobilization Induced by Progesterone in Oral Squamous Cancer Cells. *PLoS One* **2016**, *11*, e0147925.
(42) Macias, M. J.; Wiesner, S.; Sudol, M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett.* **2002**, *513*, 30−37.
(43) Cabras, T.; Pisano, E.; Boi, R.; Olianas, A.; Manconi, B.; Inzitari, R.; Fanali, C.; Giardina, B.; Castagnola, M.; Messana, I. Age-dependent modifications of the human salivary secretory protein complex. *J. Proteome Res.* **2009**, *8*, 4126−4134.

DOI: 10.1021/acs.jproteome.8b00444
J. Proteome Res. 2018, 17, 3292−3307