

Extensive *De Novo* Sequencing of New Parvalbumin Isoforms Using a Novel Combination of Bottom-Up Proteomics, Accurate Molecular Mass Measurement by FTICR–MS, and Selected MS/MS Ion Monitoring

Mónica Carrera,^{*,†} Benito Cañas,[‡] Jesús Vázquez,[§] and José M. Gallardo[†]

Instituto de Investigaciones Marinas, IIM-CSIC, Vigo, Pontevedra, Spain, Universidad Complutense de Madrid, Madrid, Spain, and Centro de Biología Molecular Severo Ochoa, CSIC-UAM, Madrid, Spain

Received February 23, 2010

Parvalbumins (PRVB) (11.20–11.55 kDa) are considered the major fish allergens. In this work, we propose a novel strategy for extensive characterization of this group of proteins based on the integration of a classical Bottom-Up proteomics approach with accurate M_r determination by FTICR–MS of intact proteins and selected MS/MS ion monitoring (SMIM) of peptide mass gaps. For each PRVB, mass spectra obtained by LC–ESI–IT–MS/MS from two digests (trypsin, Glu-C) were *de novo* sequenced manually with help of two programs (PEAKS, DeNovoX). The deduced peptide sequences were arranged and the theoretical M_r for the resulting sequences was calculated. Experimental M_r for each PRVB was measured with high mass accuracy by FTICR–MS (0.05–4.47 ppm). The masses of several missing peptide gaps were estimated by comparing the theoretical and experimental M_r , and the MS/MS spectra corresponding to these ions were obtained by LC–ESI–IT–MS/MS in the SMIM scanning mode. Finally, all peptide sequences were combined to generate the final protein sequences. This approach allowed the complete *de novo* MS-sequencing of 25 new PRVB isoforms. These new sequences belong to 11 different species from the Merlucciidae family, organisms for which genomes remain unsequenced. This study constitutes the report accounting for the higher number of new proteins completely sequenced making use of MS-based techniques only.

Keywords: proteomics • *de novo* sequencing • mass spectrometry • FTICR • selected MS/MS ion monitoring • SMIM • parvalbumin • allergen • fish • Merlucciidae

Introduction

The primary structure of the majority of proteins is commonly inferred from the corresponding nucleotide sequence;¹ however, the genome sequence of many organisms still remains only partially elucidated, and therefore the number of entries available in protein databases for these organisms is usually scarce. In these cases, information about protein sequences is only achievable by direct *de novo* sequencing of peptide fragmentation spectra obtained by MS/MS. These are straightforward and unequivocal methods that also allow identification of possible post-translational modifications and splicing variants, which cannot be characterized by nucleic-based approaches. During several decades, the direct *de novo* sequencing of peptides and proteins was performed by N-terminal Edman degradation;² however, the development of mass

spectrometry (MS) techniques, of considerable higher speed, accuracy and sensitivity,³ has relegated the use of Edman sequencing.

Commonly, the direct *de novo* sequencing of proteins via MS analysis is carried out using a classical Bottom-Up proteomics approach.⁴ The protein of interest is separated by two-dimensional gel electrophoresis (2-DE), digested to peptides using different proteases,^{5–7} and the recovered peptides are then ionized, mass measured, isolated and fragmented in a tandem mass spectrometer.⁸ Using different computer protein database search engines, like MASCOT⁹ and SEQUEST,¹⁰ fragmentation spectra are assigned to putative peptide sequences. However, when the proteins under analysis are not included in the protein databases, the sequence of the peptides cannot be identified, and therefore the fragment spectra has to be directly interpreted to obtain the peptide sequence. For this purpose, the collision-induced dissociation (CID)¹¹ remains the fragmentation method most extensively used.^{4,12} Although there are several powerful software tools available, such as DeNovoX¹³ (Thermo Electron Co.) and PEAKS,¹⁴ which may help to interpret the information contained in peptide mass spectra, in the vast majority of cases the peptide sequence obtained by automated *de novo* sequencing must be checked

* To whom correspondence should be addressed. Mónica Carrera, Instituto de Investigaciones Marinas, IIM-CSIC, Eduardo Cabello 6, E-36208 Vigo (Pontevedra), Spain. Phone: +34 986 231930. Fax: +34 986 292762. E-mail: mcarrera@iim.csic.es.

[†] Instituto de Investigaciones Marinas, IIM-CSIC.

[‡] Universidad Complutense de Madrid.

[§] Centro de Biología Molecular Severo Ochoa, CSIC-UAM.

and/or complemented by a manual interpretation of the spectrum, which is a time-consuming procedure that requires a deep expertise. Once obtained, the *de novo* peptide sequences must be assembled and meticulously ordered, mapping the overlapping parts derived from different enzymatic digests and using, when available, the sequences of homologous proteins included in the protein databases. At present, using this approach, several proteins had been partially *de novo* sequenced;⁴ however, it is not always possible to obtain the complete sequence of a protein; in fact, only five new proteins between 7.0 and 21.0 kDa have been completely sequenced using a Bottom-Up approach only.^{12,15,16}

Alternative approaches have been developed taking advantage of the high resolving power and high mass accuracy of Fourier-transform ion-cyclotron resonance (FTICR) mass spectrometers. These instruments allow a Top-Down proteomics approach,¹⁷ where the high mass accuracy makes possible the characterization of the fragments produced by using different fragmentation techniques (CID, ECD), making it unnecessary to digest the proteins. In spite of the performance of these machines, to date only the complete *de novo* sequence of a polypeptide with a M_r of 8.5 kDa has been achieved using a hybrid strategy, based on a combination of both Bottom-Up and Top-Down methodologies.¹⁸

It is necessary to emphasize that FTICR mass spectrometers are excellent instruments to use in the determination of the exact M_r of intact proteins. This accurate information may be used to corroborate the data obtained by a complete *de novo* sequencing of a protein following a Bottom-Up approach; furthermore, as is demonstrated in the present study, it may also provide the precise masses of those peptide gaps that could not be determined using the Bottom-Up approach. The link between both approaches may then be reached using a targeted mass spectrometry procedure. In this method, the mass of peptide gaps was analyzed by selected MS/MS ion monitoring (SMIM),¹⁹ where the MS detector was programmed to perform continuous MS/MS scans on the selected precursor ion along the chromatographic gradient run. With the use of this operating mode, the MS/MS spectra from each predicted peptide mass gap was obtained, and the *de novo* sequencing of these fragment spectra allowed for the complete sequence coverage of the proteins under study.

The utility of this strategy in practice is presented in this paper, where the complete *de novo* sequence of 25 new fish parvalbumin isoforms (PRVB; 11.20–11.55 kDa) was achieved by integrating three proteomics approaches: classical Bottom-Up proteomics, accurate M_r determination of intact proteins by FTICR–MS and the monitoring of peptide gaps by SMIM. Moreover, 16 other new PRVBs isoforms were also partially sequenced (53.7–90.7% of coverage). These thermostable and allergenic proteins^{20,21} were *de novo* sequenced for all the fish species of commercial interest from the Merlucciidae family, organisms whose genomes are still waiting for sequencing. It is important to emphasize that the PRVBs are considered the major source of IgE-mediated hypersensitivities (Type-I allergy),^{20–22} and therefore the knowledge of the primary structure of these new allergens provides information that is of paramount importance for further immunological studies. To our knowledge, this work represents the highest number of new proteins that have been completely *de novo* sequenced using only MS-based techniques.

Table 1. Species and Subspecies from the Merlucciidae Family Included in the Study

species ^a /subspecies	common name	origin
<i>M. merluccius</i>	European hake	Spanish coasts
<i>M. capensis</i>	Cape hake	South Africa
<i>M. senegalensis</i>	Senegalense hake	Northwest Africa
<i>M. polli</i>	Benguela hake	Northwest Africa
<i>M. paradoxus</i>	Deep-water Cape hake	South Africa
<i>M. hubbsi</i>	Patagonian hake	South America
<i>M. gayi</i>	Peruvian or Chilean hake	South America
<i>M. australis polylepis</i>	Austral hake	South America
<i>M. australis australis</i>	Austral hake	New Zealand coasts
<i>M. productus</i>	Pacific hake	North America
<i>M. bilinearis</i>	Silver hake	North America
<i>Ma. novaezelandiae novaezelandiae</i>	Blue grenadier	New Zealand coasts
<i>Ma. novaezelandiae magellanicus</i>	Patagonian grenadier	South America

^a *M.* (*Merluccius* genus); *Ma.* (*Macruronus* genus).

Materials and Methods

1. Fish Material. All the main commercial species from the Merlucciidae family were employed in this study: ten different hake species, including two different subspecies from *Merluccius australis* and two grenadier subspecies belonging to the *Macruronus novaezelandiae* species (Table 1). Except for European hake, the specimens were frozen on board at -30°C , with special care in keeping their morphological characters in good shape, and shipped by plane to the laboratory for the analyses. The weight of every specimen studied was in the range of 3–6 kg. At least 10 fishes belonging to each different species were subjected to taxonomical study according to their anatomical and morphological features by an expert marine biologist and by genetic identification in the Food Biochemistry laboratory from the Marine Research Institute (Vigo, Pontevedra, Spain) and with the fishID Kit (Bionostra SL., Madrid, Spain). Among them, five correctly identified individuals from each of the species were considered and selected as the representative species for the study.

2. Extraction of Sarcoplasmic Proteins. Sarcoplasmic protein extraction was carried out by homogenizing 5 g of white muscle from each of the species in 10 mL of 10 mM Tris-HCl buffer, pH 7.2, supplemented with 5 mM PMFS, during 30 s in an Ultra-Turrax device (IKA-Werke, Staufen, Germany). The fish extracts were then centrifuged at $40\,000\times g$ for 20 min at 4°C (J221-M centrifuge; Beckman, Palo Alto, CA). The supernatants were then recovered, filtered using Ultrafree CL ($0.22\,\mu\text{m}$) filters (Millipore, Bedford, MA) and stored at -80°C . Protein concentration in the extracts was determined by the bicinchoninic acid method using BSA like standard (Sigma-Chemical Co., St. Louis, MO).

3. Bottom-Up Proteomics Approach. 3.1. Protein Separation by Two-Dimensional Gel Electrophoresis (2-DE). Isoelectric focusing (IEF) was performed at 10°C in a Multiphor II electrophoresis unit (Amersham Biosciences, Sweden) according to a previous report.²³ A total of $160\,\mu\text{g}$ protein extract was loaded in duplicate using a sample applicator paper on the IEF strips having a narrow pH range (4.0–6.5) (Amersham Bioscience). A mixture of standard proteins in the 2.5–6.5 pH range (low pI standard from Amersham Biosciences) was included. IEF conditions were 1500 V, 50 mA, 30 W until at least 4000 Vh were reached. IEF strips corresponding to

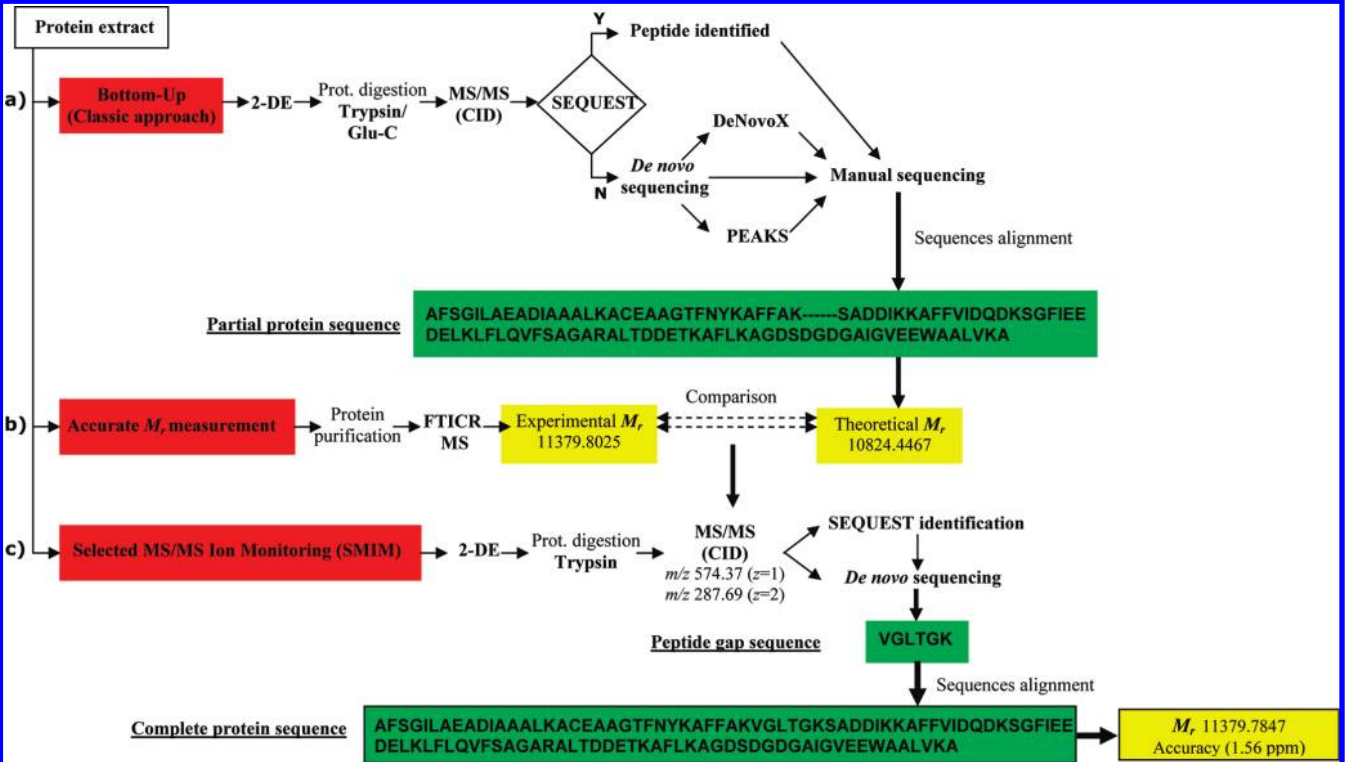


Figure 1. Analytical scheme of the three sequential proteomics approaches employed for the complete *de novo* sequencing of new proteins: (a) classical Bottom-Up proteomics approach, (b) accurate M_r determination of intact protein by FTICR–MS and (c) monitoring of peptide mass gaps by Selected MS/MS Ion Monitoring (SMIM).

Table 2. Isoelectric Point and Molecular Weight for All of the PRVB Spots Studied

species ^a /subspecies	PRVB spot number	pI	M_r (kDa)	species ^a /subspecies	PRVB spot number	pI	M_r (kDa)
<i>M. merluccius</i>	P1	4.53	11.30	<i>M. australis polylepis</i>	P21	4.20	11.20
	P2	4.19	11.38		P22	3.78	11.33
	P3	4.02	11.39		P23	4.30	11.30
<i>M. capensis</i>	P4	4.55	11.30	<i>M. australis australis</i>	P24	4.14	11.55
	P5	4.20	11.38		P25	3.98	11.33
	P6	3.95	11.39		P26	4.51	11.25
<i>M. senegalensis</i>	P7	4.55	11.30	<i>M. productus</i>	P27	4.30	11.30
	P8	4.20	11.38		P28	4.14	11.53
	P9	3.92	11.37		P29	3.98	11.33
<i>M. polli</i>	P10	4.51	11.30	<i>M. bilinearis</i>	P30	4.51	11.37
	P11	4.19	11.35		P31	4.29	11.50
	P12	3.84	11.35		P32	4.23	11.35
<i>M. paradoxus</i>	P13	4.51	11.35	<i>Ma. novaezelandiae novaezelandiae</i>	P33	4.23	11.35
	P14	4.16	11.35		P34	3.98	11.27
	P15	3.79	11.32		P35	4.51	11.25
<i>M. hubbsi</i>	P16	4.57	11.30	<i>Ma. novaezelandiae magellanicus</i>	P36	4.05	11.35
	P17	4.30	11.55		P37	3.75	11.35
	P18	4.09	11.35		P38	4.51	11.25
<i>M. gayi</i>	P19	4.56	11.30		P39	4.05	11.35
	P20	4.27	11.53		P40	3.75	11.35

^a *M.* (*Merluccius* genus); *Ma.* (*Macruronus* genus).

individual lanes were cut immediately after the run was completed and kept frozen at -80°C for the second dimension by SDS-PAGE. A duplicate IEF strip was stained with 0.1% Coomassie brilliant blue (Sigma Chemical) according to the Amersham Biosciences staining protocol. Equilibration of pH 4–6.5 IEF gel strips was carried out at room temperature as described previously.²⁴ The second dimension was run in vertical SDS-PAGE gels (10% T and 3% C) ($14 \times 13.5 \times 1.5 \text{ cm}^3$) using a Tris-Tricine buffer at 15°C in the MultiTemp III system (Amersham Biosciences). The running conditions were as

follows: 100 V, 40 mA per gel and 150 W during 16–18 h. A low molecular weight protein standard (LMW Calibration Kit, Amersham Biosciences) was used as reference. Protein spots were visualized with Coomassie brilliant blue staining (Amersham Bioscience) according to manufacturer instructions, and the PDQuest 7.1.0 software (Bio-Rad Laboratories, Ltd., U.K.) was used for image analysis.

3.2. In-Gel Protein Digestion with Trypsin. PRVB spots (11.20–11.55 kDa; 3.75–4.55 pI) were excised from the gel taking care in maximizing the protein-to-gel ratio.²³ Only the

Table 3. Summary of the Results Obtained Using SEQUEST

	PRVB spot	pI	matched peptides trypsin/Glu-C	sequence coverage (%)	homology (%)
PRVB_A		Best match with β -PRVB <i>M. merluccius</i> (P02620, PRVB_MERME)			
Complete sequences	P1	4.53	44/28	100.0	100.0
	P1b	4.53	36/26	100.0	99.0
	P4	4.55	43/32	100.0	100.0
	P7	4.55	44/29	100.0	100.0
Partial sequences	P10	4.51	38/29	96.2	95.3
	P13	4.51	36/24	90.7	89.8
	P16	4.57	40/26	96.2	95.3
	P19	4.56	38/26	96.2	95.3
	P23	4.30	37/27	93.5	92.5
	P26	4.51	37/28	93.5	92.5
	P27	4.30	37/25	93.5	92.5
	P30	4.51	30/20	85.1	85.1
	P33b	4.23	35/15	75.9	75.0
	P35	4.51	6/6	50.9	49.0
	P38	4.51	6/6	50.9	49.0
PRVB_B		Best match with β -PRVB <i>M. bilinearis</i> (P56503, PRVB_MERBI)			
Partial sequences	P2	4.19	19/5	69.4	67.5
	P5	4.20	19/8	69.4	67.5
	P8	4.20	19/9	69.4	67.5
	P11	4.19	22/9	68.5	67.5
	P14	4.16	20/10	69.4	68.5
	P17	4.30	26/8	73.1	73.1
	P20	4.27	24/10	73.1	73.1
	P21	4.20	21/6	54.6	54.6
	P24	4.14	24/5	73.1	73.1
	P28	4.14	24/5	73.1	73.1
	P31	4.29	21/4	62.9	62.9
	P33	4.23	21/6	67.5	67.5
	P36	4.05	19/7	67.5	64.8
	P39	4.05	19/11	67.5	64.8
PRVB_C		Best match with β -PRVB <i>Theragra chalcogramma</i> (Q90YK8, PRVB1_THECH)			
Partial sequences	P3	4.02	5/1	26.8	23.1
	P6	3.95	2/1	7.4	4.6
	P9	3.92	2/1	7.4	4.6
	P12	3.84	6/4	26.8	23.1
	P15	3.79	4/1	17.5	14.8
	P18	4.09	9/4	40.7	37.0
	P22	3.78	9/2	40.7	37.0
	P25	3.98	8/2	31.4	28.7
	P29	3.98	8/1	31.4	28.7
	P32	4.23	9/1	40.7	37.0
	P34	3.98	9/3	40.7	37.0
	P37	3.75	8/5	28.7	27.7
	P40	3.75	8/4	28.7	27.7

most intense stained region at center of the spot was excised to avoid extracting an excess of gel matrix. Excised pieces were subjected to *in-gel* digestion with trypsin (Promega, Madison, WI) as described.²⁵

3.3. In-Gel Protein Digestion with Endoproteinase Glu-C from *Staphylococcus aureus* V8. Other excised spots from PRVBs were subjected to *in-gel* protein digestion with *Staphylococcus aureus* V8 protease. For that, PRVB spots were excised from the gels, cut into pieces and washed with Milli-Q-water. Afterward, the pieces were dehydrated with acetonitrile and dried in a vacuum centrifuge. Gel pieces were further rehydrated for 10 min at room temperature with 1 μ g/mL of endoproteinase Glu-C from *Staphylococcus aureus* (Pierce Biotechnology, Inc., Rockford, IL) in 50 mM sodium phosphate buffer pH 7.5, to a final volume of 20 μ L. After the rehydration step, samples were digested at 37 °C for 16–24 h.

3.4. Peptide Fragmentation by MS/MS. Peptide digests were analyzed online by LC–ESI–IT–MS/MS using a LC system

model SpectraSystem P4000 (Thermo Fisher, San Jose, CA) coupled to an ion trap mass spectrometer model LCQ Deca XP Plus (Thermo Fisher). The separation was performed on a 0.18 mm $\times$ 150 mm BioBasic-18 RP column (ThermoHypersil-Keystone), using 0.5% acetic acid in Milli-Q-water and in 80% acetonitrile as mobile phases A and B, respectively. A 90 min linear gradient from 5 to 60% B, at a flow rate of 1.5–1.7 μ L/min, was used. ESI parameters were: spray voltage, 3.5 kV; N₂ flow, 10 arbitrary units; and capillary temperature, 200 °C. Peptides were detected in positive mode using the triple play method and survey scans from 200 to 2500 amu (3 μ scans), followed by a data-dependent ZoomScan (5 μ scans) and MS/MS scan (5 μ scans), using an isolation width of 3 amu and a normalized collision energy of 35%. Fragmented masses were set in dynamic exclusion for 3 min after the second fragmentation event and singly charged ions were excluded from MS/MS analysis.

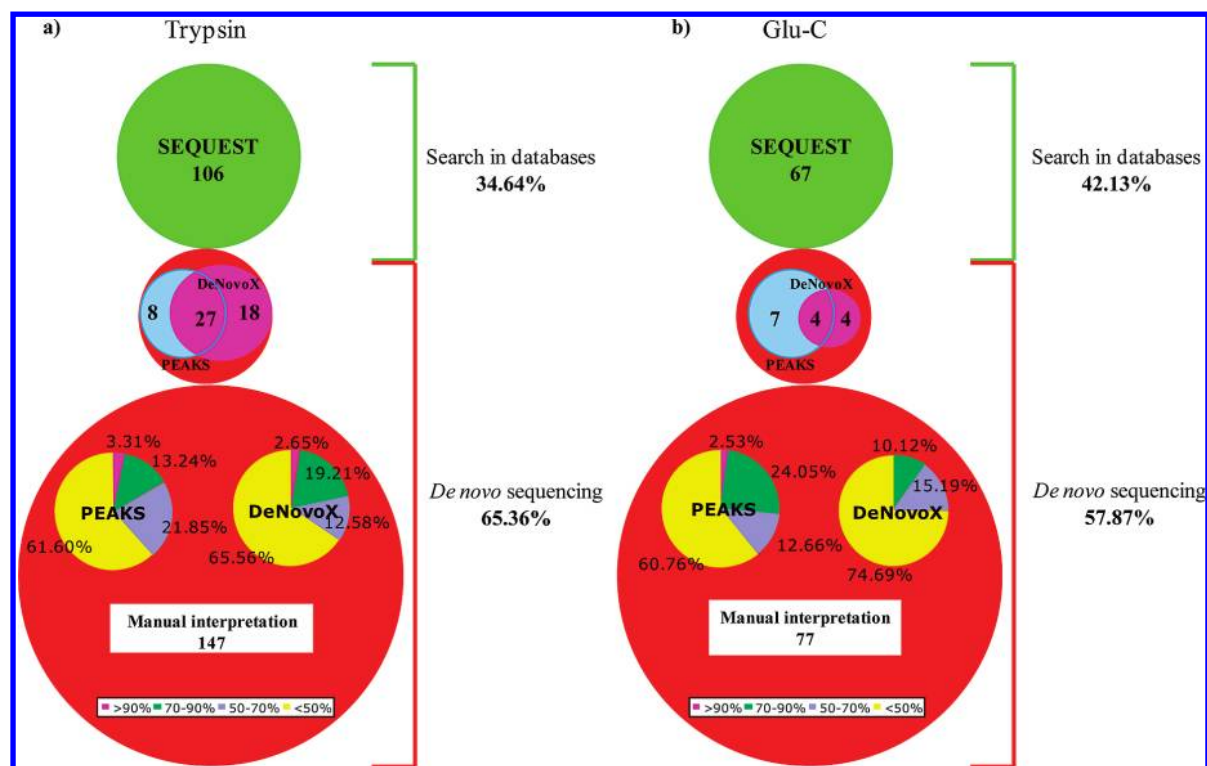


Figure 2. Comparative schematic representation of the results obtained for PRVB isoforms peptide identification and sequencing, using, respectively, database-searching and different *de novo* sequencing methods: (a) tryptic peptides and (b) Glu-C peptides. For the manual interpretation, the correct amino acid sequence assignment for each program is assigned with percentage values.

Off-line analysis by nESI-IT-MS were performed using an ion trap mass spectrometer, model LCQ Deca XP Plus from Thermo Fisher, equipped with a nanospray interface. Peptides were previously desalted and concentrated using reserve-phase ZipTip C18 minicolumns (Millipore, Bedford, MA) according to the manufacturer's recommendations and eluted with 5–10 μ L of 70% methanol/0.5% acetic acid. PicoTips emitters made with borosilicate glass needles with 1 μ m orifice (New Objective, Woburn, MA) were filled with 3–5 μ L of sample. The instrument parameters were adjusted during the analysis, using 0.8–1.2 kV and 150 $^{\circ}$ C. The analyses were performed using an isolation width of 3 amu and collision energy was tuned between 35–45% depending on the peptide fragmented. ZoomScan and MS/MS spectra were averaged for at least 1 min for every peptide.

3.5. Mass Spectrometry Data Processing by SEQUEST. MS/MS spectra were searched using SEQUEST (Bioworks 3.1 package, Thermo Fisher) against the complete and general database UniProtKB release 15.0. The following constraints were used for the searches: tryptic cleavage or Glu-C protease in sodium phosphate buffer according to the enzyme used, up to two missed cleavage sites, and tolerances ± 1.8 Da for precursor ions, and ± 0.8 Da for MS/MS fragments ions. The variable modifications allowed were methionine oxidation (Mox), carbamidomethylation of Cys (C*) and acetylation of the N-terminus of the protein (N-Acyl).

3.6. De Novo Sequencing. *De novo* sequencing was performed by manual interpretation of the ion series of the spectra with aid of the software packages: DeNovoX (Thermo Fisher) and PEAKS Studio 4.2 (Bioinformatics Solutions Inc., Waterloo, Ontario, Canada). The parameters used for both programs were as follows: selection of the peptide charge (1+, 2+, 3+), tolerances of 0.3–0.5 Da for precursor and fragments ions, use

or not of trypsin or Glu-C as proteases and three variable modifications: methionine oxidation (Mox), carbamidomethylation of Cys (C*) and acetylation of the N-terminus of the protein (N-Acyl).

All peptide sequences obtained were meticulously ordered by overlapping the results obtained with both enzymatic digests and by comparison, using BLAST, with the proteins included in the UniProtKB database. In those cases where a complete PRVB sequence was obtained, the theoretical M_r was calculated using the Molecular Weight Calculator program.

4. Accurate M_r Determination of Intact PRVBs by FTICR-MS. Intact PRVBs from each species were purified by treatment with heat. To determine the precise temperature for heat precipitation, the crude sarcoplasmic extracts from each of species were heated to different temperatures (60–90 $^{\circ}$ C, for 5 min). Afterward the extracts were centrifuged at 40 000 $\times g$ for 10 min at 4 $^{\circ}$ C (J221-M centrifuge; Beckman) and the supernatants were evaluated by IEF gels (4.0–6.5 pH) (Amersham Bioscience). Once an adequate purification temperature was selected (see Results section), 300 μ g of protein were desalted and concentrated using reversed-phase microcolumns (Vivapore C-18, Vivascience Ltd., Stonehouse, U.K.) according to the manufacturer's instructions and samples eluted in 60 μ L of 70% acetonitrile/0.1% TFA.

Measurements of M_r of intact PRVBs were performed on a 7 T FTICR mass spectrometer (APEXIII, Bruker Daltonics). The samples (60 μ L) were analyzed by direct infusion in a Apollo II ESI source, at a flow rate of 1.0 μ L/min. Ions were generated by positive ion mode using a potential of 3.7 kV, externally accumulated (2 s) in a quadrupole and transferred to the region with high vacuum ($\sim 2 \times 10^{-10}$ Torr). The ions were excited in the FTICR using a range of excite radiofrequency between 10 kHz and 3 MHz and detected in the ICR cell. For better mass

Table 4. Summary of Results Obtained after *De Novo* Sequencing^a

	PRVB spot	SEQUESTTryp/Glu-C	PEAKS Tryp/Glu-C	DeNovoX Tryp/Glu-C	manual interp.Tryp/Glu-C	sequence coverage (%)	homology (%)
PRVB_A							
Best match with β -PRVB <i>M. merluccius</i> (P02620, PRVB_MERME)							
Complete seq.	P1	44/28	2/0	3/0	13/5	100.0	100.0
	P1b	36/26	2/0	3/0	16/8	100.0	99.0
	P4	43/32	3/1	7/0	15/12	100.0	100.0
	P7	44/29	1/0	5/0	9/4	100.0	100.0
	P23	37/27	1/0	3/0	11/8	100.0	97.2
	P26	37/28	2/0	4/0	10/5	100.0	96.2
	P27	37/25	1/0	3/0	9/5	100.0	97.2
Partial seq.	P33b	35/15	1/1	4/1	8/4	100.0	96.2
	P10	38/29	2/3	3/1	14/15	96.2	95.3
	P13	36/24	2/1	2/0	13/9	90.7	89.8
	P16	40/26	2/0	6/0	13/13	96.2	95.3
	P19	38/26	1/4	3/2	8/10	96.2	95.3
	P30	30/20	2/0	4/0	8/6	96.2	93.5
	P35	6/6	8/0	8/0	13/2	90.7	80.5
PRVB_B	Best match with β -PRVB <i>M. bilinearis</i> (P56503, PRVB_MERBI)						
	P36	19/7	4/0	4/2	12/8	100.0	91.6
	P39	19/11	4/2	4/3	11/11	100.0	91.6
Partial seq.	P2	19/5	1/1	2/1	13/7	99.4	88.8
	P5	19/8	1/1	2/0	20/7	99.4	88.8
	P8	19/9	1/1	2/0	20/7	99.4	88.8
	P11	22/9	1/1	1/0	12/4	99.4	89.8
	P14	20/10	0/1	1/1	13/7	99.4	89.8
	P17	26/8	1/2	1/1	8/7	99.4	90.7
	P20	24/10	0/1	0/0	7/10	99.4	90.7
	P21	21/6	0/1	0/0	3/5	69.4	67.5
	P24	24/5	0/1	0/0	13/7	99.4	90.7
	P28	24/5	0/1	0/0	14/7	99.4	90.7
	P31	21/4	0/0	1/0	10/4	99.4	88.8
	P33	21/6	0/1	1/1	12/7	99.4	90.7
PRVB_C							
Best match with β -PRVB <i>Theragra chalcogramma</i> (Q90YK8, PRVB1_THECH)							
Partial seq.	P3	5/1	3/0	4/0	11/1	63.8	53.7
	P6	2/1	2/0	2/0	10/2	53.7	43.5
	P9	2/1	3/0	3/0	8/1	53.7	43.5
	P12	6/4	3/0	3/0	9/2	63.8	53.7
	P15	4/1	6/0	7/0	11/5	87.9	75.9
	P18	9/4	3/0	5/0	17/7	79.6	68.5
	P22	9/2	2/0	2/0	7/3	84.2	72.2
	P25	8/2	3/0	4/0	7/2	70.3	60.1
	P29	8/1	2/0	2/0	4/1	63.8	54.6
	P32	9/1	2/0	3/0	6/2	80.5	68.5
	P34	9/3	2/0	2/0	10/1	87.0	74.0
	P37	8/5	2/0	3/0	4/4	76.8	67.5
	P40	8/4	2/0	3/0	5/4	68.5	59.2

^a This table includes the number of peptides correctly assigned according to each procedure. Manual interp. (manual interpretation of the corresponding MS/MS spectrum).

accuracy, the instrument was calibrated externally before each measurement. Mass spectra were deconvoluted, and the precise M_r of each PRVB was determined using the FlexAnalysis program (Bruker).

5. Selected MS/MS Ion Monitoring (SMIM). The comparison between the theoretical M_r (Bottom-Up approach) and experimental M_r (FTICR–MS) for each of the PRVBs allowed for the calculation of masses for certain peptide gaps, whose sequences were not determined by the first approach. These gaps were analyzed by the SMIM,¹⁹ using peptides from the tryptic digestion of several PRVBs spots. A LC–ESI–IT–MS/MS setup (SpectraSystem P4000; LCQ Deca XP Plus, Thermo Fisher) was used as previously described. For this routine, the MS instrument was programmed to perform continuous MS/MS scans of the predicted singly or doubly charged precursor

ions along the complete chromatographic separation. All MS/MS spectra were analyzed by *de novo* sequencing and the peptide sequences obtained were combined with those previously obtained to complete the sequence coverage of some of the PRVBs proteins. Finally, the experimental and the theoretical M_r from each complete PRVB isoforms were compared and the new protein sequences obtained were put side by side with the homologous proteins in the UniProtKB database using the BLAST program.

Results and Discussion

1. Strategy for Complete *De Novo* Sequencing of Proteins. The strategy for the complete *de novo* sequencing of the proteins proposed in this work is summarized in Figure 1. This

Table 5. Comparison between Theoretical and Experimental M_r for those PRVB Isoforms that Presented a Sequence Coverage Higher than 90%

	PRVB spot	theoretical M_r (Bottom-Up) (Da)	experimental M_r (FTICR) (Da)	$\Delta M_r/M_r$ (ppm)	M_r gap (to SMIM) (Da)
PRVB_A					
Best match with β -PRVB <i>M. merluccius</i> (P02620, PRVB_MERME)					
Complete seq.	P1	11329.7473	11329.7460	0.11	—
	P1b	11315.7316	11315.7370	0.47	—
	P4	11329.7473	11329.7416	0.49	—
	P7	11329.7473	11329.7440	0.29	—
	P23	11302.7000	11302.7150	1.33	—
	P26	11272.7258	11272.7763	4.47	—
	P27	11302.7000	11302.7144	1.27	—
	P33b	11271.7054	11271.6845	1.85	—
Partial seq.	P10	10896.5148	11329.7890	—	433.2742
	P13	10311.2025	11355.8076	—	1044.6050
	P16	10896.5148	11329.7479	—	433.2331
	P19	10896.5148	11329.7490	—	433.2342
	P30	10940.5410	11373.7940	—	433.2530
	P35	10383.2237	11329.7278	—	946.5041
	P38	10383.2237	11329.7317	—	946.5080
PRVB_B					
Best match with β -PRVB <i>M. bilinearis</i> (P56503, PRVB_MERBI)					
Complete seq.	P36	11341.7149	11341.7286	1.21	—
	P39	11341.7149	11341.7252	0.90	—
Partial seq.	P2	10824.4467	11379.8025	—	555.3558
	P5	10824.4467	11379.8119	—	555.3652
	P8	10824.4467	11379.8196	—	555.3729
	P11	10794.4725	11349.8483	—	555.3757
	P14	10810.4310	11365.8076	—	555.3765
	P17	10799.4514	11354.8086	—	555.3571
	P20	10799.4514	11354.8095	—	555.3581
	P24	10799.4514	11354.8058	—	555.3544
	P28	10799.4514	11354.8027	—	555.3513
	P31	10755.4616	11310.8227	—	555.3611
	P33	10799.4514	11354.7956	—	555.3442

strategy integrates three different proteomics approaches: (a) classical Bottom-Up approach, (b) accurate M_r determination of intact proteins by FTICR-MS and (c) monitoring of selected peptide mass gaps by SMIM. With this strategy 25 new PRVB protein isoforms were completely sequenced. The results of *de novo* sequencing of all the PRVB isoforms from Merlucciidae family, using each of these approaches are described in the following sections.

2. Bottom-Up Proteomics Approach. 2.1. Peptides Identified by SEQUEST. Sarcoplasmic protein extracts from 5 individuals of each species/subspecies (65 specimens) were analyzed by 2-DE. To ensure the reproducibility of the experiments, 3 gels were run per individual. Thus, a total of 195 2-DE gels were processed and analyzed using the PdQuest image software. All PRVB spots (in the range 11.20–11.55 kDa and pI 3.75–4.57 units) were identified as previously described.²³ The 2-DE PRVB patterns for each species/subspecies are shown in the Figure 1 of Supplemental Data 1 (Supporting Information). Table 2 shows a compilation of the pI and M_r for all the PRVB spots studied in this work, which were designated as P1–P40.

The PRVB spots were excised from the gels, digested *in-gel* with trypsin or Glu-C proteases and the peptides produced were analyzed by LC-ESI-IT-MS/MS. To increase the reproducibility, three LC-ESI-IT-MS/MS runs were taken per spot, making a total of 600 analyses. The fragmentation spectra were analyzed by protein database search using SEQUEST.¹⁰ Unfortunately, little information is available in the databases about the PRVBs from the Merlucciidae family. In fact, only two PRVB isoforms for this family are included in the protein databases: β -PRVB from *M. merluccius* (P02620, PRVB_MERME)²⁶ and

β -PRVB from *M. bilinearis* (P56503, PRVB_MERBI).²⁷ For this reason, SEQUEST searches were performed against the general database UniProtKB release 15.0. Identifications were preliminarily considered as correct when presenting the following parametric values: $\Delta Cn \geq 0.1$ and XCorr ≥ 1.5 , ≥ 2.0 or ≥ 2.5 for singly, doubly or triply charged peptides, respectively. Afterward, all results were carefully validated by manual interpretation of the corresponding MS/MS spectra.

The complete list of 106 peptide sequences obtained after digestion with trypsin and identified by SEQUEST can be found in the Tables 1–3 of Supplemental Data 2 (Supporting Information).

To increase sequence coverage, all PRVB spots were also digested with the endoproteinase Glu-C from *Staphylococcus aureus* V8 before LC-ESI-IT-MS/MS analysis. This enzyme cleaves the peptide bonds at the carboxyl side of Glu and Asp in sodium phosphate buffer.⁶ Tables 4–6 of Supplemental Data 2 (Supporting Information) collect the 67 peptide sequences, produced by the Glu-C protease digestion, which were identified by SEQUEST and validated manually.

In agreement with previous work performed using MALDI-TOF-MS,²³ the results obtained by SEQUEST show that all the PRVB isoforms with the highest pI (4.30–4.57) presented some degree of homology with the β -PRVB from *M. merluccius* (P02620, PRVB_MERME) (Table 3). These PRVB isoforms were included into a group named as PRVB_A. Another group of isoforms, which were found to display homology with the β -PRVB from *M. bilinearis* (P56503, PRVB_MERBI), were included in the PRVB_B group. Finally, a third group, named PRVB_C, contained isoforms showing no significant homologies

with the described Merlucciidae PRVBs but some degree of homology with the β -PRVB from the fish *Theragra chalcogramma* (Q90YK8, PRVB1_THECH). A summary of the data obtained by SEQUEST is shown in the Table 3.

All the peptides identified by SEQUEST were then aligned with one of the three PRVB sequences included in the protein databases showing the highest degree of homology. These alignments are shown in Figure 1–3 in Supplemental Data 3 (Supporting Information).

With this database-search approach, only 4 PRVB isoforms were completely sequenced (spots P1, P1b, P4 and P7). The homology degree was variable in function to species and type of isoform. Thus the spots P1, P4 and P7 from PRVB_A presented 100% homology when compared with the sequence PRVB_MERME. However, the spots P6 and P9 from PRVB_C presented only a 4.6% homology with the sequence PRVB1_THECH. It is important to remark here that from the analysis of spot P1 from *M. merluccius*, besides the sequence described in the protein databases (P02620, PRVB_MERME), a new isoform of PRVB_A was found. This new isoform, denoted P1b, was completely sequenced and presented a 99.0% homology with PRVB_MERME. This polymorphism is due to a one amino acid substitution, Glu to Asp in position 100 (Figure 1 in Supplemental Data 3, Supporting Information). Regarding the PRVB_A group, it is also noteworthy that sequence P33b, corresponding to *M. bilinearis*, had a 75% homology with PRVB_MERME. This sequence was identified within spot P33 belonging to PRVB_B group, where two different PRVB isoforms were found.

The lack of total sequence coverage in the majority of PRVB sequences was indicative of amino acid substitutions in their sequences, making necessary the use of *de novo* sequencing procedures to characterize the remaining MS/MS spectra.

2.2. De Novo Mass Spectrometry Sequencing. All MS/MS spectra that were not unambiguously identified by SEQUEST were further analyzed by *de novo* peptide sequencing using two automatic programs: PEAKS and DeNovoX. The candidate peptide sequences were evaluated as a function of their probabilistic score, homology degree using BLAST and by means of an exhaustive manual interpretation of their corresponding MS/MS spectra. Those spectra producing ambiguous sequences were acquired again using an off-line nanospray ionization source setting.

Tables 1–3 in Supplemental Data 4 (Supporting Information) show the results for the *de novo* sequencing, using the program PEAKS, of the tryptic peptides produced by each of the three types of PRVBs. With this program, a total of 198 tentative sequences (PRVB_A: 85, PRVB_B: 64, PRVB_C: 49) were generated. Among them, 31.3% corresponded to peptides that presented some degree of homology by BLAST with PRVBs present in protein databases (Tables 1–3 of Supplemental Data 4, in green, Supporting Information). After manual interpretation of their fragmentation spectra, 17.7% of the sequences (35 peptides) were considered to be correct assignments (Tables 1–3 of Supplemental Data 4, in yellow, Supporting Information). The analysis by PEAKS of the spectra of peptides produced by Glu-C (Tables 4–6 in Supplemental Data 4, Supporting Information) showed a total of 92 tentative sequences (PRVB_A: 45, PRVB_B: 31, PRVB_C: 16). Among them, 29.3% of the sequences corresponded to peptides showing homology with PRVBs present in the databases. After manual verification, 11.9% (11 peptides) were considered to be correct peptide sequences.

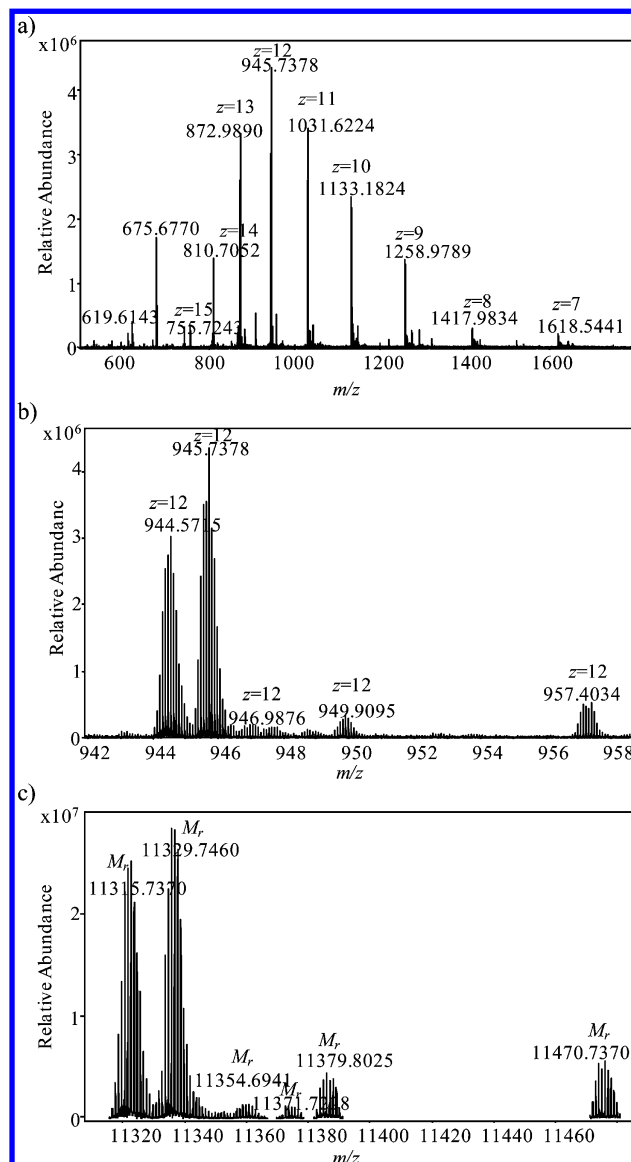


Figure 3. Analysis of intact PRVBs from *M. merluccius* by ESI-FTICR-MS: (a) distribution of the multiple charge states, (b) amplification of the most intense peak corresponding to charge ($z = 12$), showing the different PRVBs isoforms, and (c) M_r determination for each of the PRVB isoforms.

The results obtained using the program DeNovoX are compiled in Tables 1–6 in Supplemental Data 5 (Supporting Information). For tryptic peptides, from the 198 tentative peptide sequences assembled, 33.8% of the peptides corresponded to sequences that presented homology with PRVBs available in the protein databases. Among them, 22.7% (45 peptides) were considered as correct assignments. However, for the case of peptides produced by Glu-C digestion, only 19.5% of the sequences showed homology to database entries, from which 8.7% (8 peptides) were judged as correct by manual interpretation. This limited number of homologue and correct sequences obtained by DeNovoX with spectra of peptides coming from Glu-C digestion is because the available version of the program was not allowed to state Glu-C protease as a proteolytic enzyme, so that the sequencing had to be performed without indicating any protease. When Glu-C peptide spectra were excluded, the two *de novo* sequencing programs provided similar results.

Table 6. SMIM Monitoring Single and Double-Charged Tryptic Precursor Ions

	PRVB spot	M_r gap (Da)	position	residues known and mass gap	SMIM m/z (1+)	SMIM m/z (2+)	peptide sequences ^a
PRVB_A	P10	433.2742	20–25	(H) AE[433.2742] (OH)	652.37	326.68	AEGTFK
	P13	1044.6050	20–25	(H) AE[≈433.25] (OH)	652.34	326.67	AEGTFK
			39–44	(H) [≈611.35] (OH)	630.34	315.67	SPADIK
	P16	433.2331	20–25	(H) AE[433.2331] (OH)	652.33	326.66	AEGTFK
	P19	433.2342	20–25	(H) AE[433.2342] (OH)	652.33	326.66	AEGTFK
	P30	433.2530	20–25	(H) AE[433.2530] (OH)	652.35	326.67	AEGTFK
	P35	946.5041	39–44	(H) [≈645.33] or [≈671.34] (OH)	646.34 or 672.35	323.67 or 336.68	–
			84–87	(H) [≈549.31] (OH)	550.32	275.66	–
	P38	946.5080	39–44	(H) [≈645.33] or [≈671.34] (OH)	646.34 or 672.35	323.67 or 336.68	–
			84–87	(H) [≈549.31] (OH)	550.32	275.66	–
PRVB_B	P2	555.3558	33–38	(H) [555.3558] (OH)	574.37	287.69	VGLTGK
	P5	555.3652	33–38	(H) [555.3652] (OH)	574.38	287.69	VGLTGK
	P8	555.3729	33–38	(H) [555.3729] (OH)	574.39	287.69	VGLTGK
	P11	555.3757	33–38	(H) [555.3757] (OH)	574.39	287.70	VGLTGK
	P14	555.3765	33–38	(H) [555.3765] (OH)	574.39	287.70	VGLTGK
	P17	555.3571	33–38	(H) [555.3571] (OH)	574.37	287.69	VGLTGK
	P20	555.3581	33–38	(H) [555.3581] (OH)	574.37	287.69	VGLTGK
	P24	555.3544	33–38	(H) [555.3544] (OH)	574.37	287.69	VGLTGK
	P28	555.3513	33–38	(H) [555.3513] (OH)	574.36	287.68	VGLTGK
	P31	555.3611	33–38	(H) [555.3611] (OH)	574.37	287.69	VGLTGK
	P33	555.3442	33–38	(H) [555.3442] (OH)	574.36	287.68	VGLTGK

^a (–) Not determined.

Manual interpretation was, nevertheless, the method of choice for *de novo* sequencing; although it is a slow process, a total of 147 MS/MS spectra obtained after digestion with trypsin and 77 MS/MS spectra of Glu-C digests were successfully sequenced. Sometimes, the partial sequences obtained with PEAKS and/or DeNovoX were extremely useful to accelerate the manual procedure to acquire tentative sequences. All peptide sequences that were manually interpreted were later corroborated by BLAST, showing in all cases homologies with other PRVBs present in the databases.

In all cases, the assignment of Leu/Ile residues was performed by homology using BLAST against the PRVB sequences available in the protein databases. A similar procedure was used for Gln/Lys assignment, except in the C-terminal of tryptic peptides, where Lys was assumed to be correct.

Tables 1–6 in Supplemental Data 6 (Supporting Information) contain the complete list of peptide sequences obtained after SEQUEST identification and *de novo* sequencing for all PRVB isoforms. As shown, some modified amino acids such as oxidized Met and Trp residues, methylated Lys and Arg and formylated Asp were detected.

After all these analyses, from a total of 306 tryptic peptides, 106 (34.64%) were identified by SEQUEST, while the rest, 200 peptides (65.36%), were identified by *de novo* sequencing (Figure 2a). From the 159 Glu-C peptide sequences assigned, 42.13% were identified by database searching and 57.87% by *de novo* sequencing (Figure 2b).

2.3. Assembling of Peptide Sequences Obtained by the Bottom-Up Approach. The peptide sequences from each of the PRVB isoforms (P1–P40) obtained by the Bottom-Up approach were finally assembled by overlapping those produced by each of the enzymatic digest procedures and aligned with the sequences in the protein databases presenting higher homology: PRVB_MERME, PRVB_MERBI and PRVB1_THECH (Figures 1–3 of Supplemental Data 7, Supporting Information).

With this Bottom-Up approach, a total of 10 PRVB isoforms were completely sequenced: 8 PRVB_A (P1, P1b, P4, P7, P23, P26, P27, P33b) and 2 PRVB_B (P36, P39) isoforms (Figure 1–3

of Supplemental Data 7, Supporting Information). A summary of the data obtained at this state is presented in the Table 4.

The theoretical protein M_r for all, complete or not, PRVB isoforms was then calculated (Table 5). All the isoforms had lost the initial Met residue, and the new N-termini residue was found to be acetylated in all cases. These results were validated by the TerminiNator prediction program²⁸ and taken into account in the calculation of the theoretical M_s .

3. Acquisition of Accurate M_r for PRVB Isoforms using FTICR–MS. The next step in the proposed strategy was to determine the intact protein mass for each of the PRVB isoforms. PRVBs were purified from the sarcoplasmic aqueous extract, taking advantage of their thermostability.²⁹ Contaminant proteins were precipitated by heating the sarcoplasmic extract for 5 min. The most appropriate temperature for purification was determined by testing different temperatures between 60 and 90 °C and evaluating the supernatants obtained after centrifugation by IEF electrophoresis (pH 4.0–6.5). Figures 1–2 in Supplemental Data 8 (Supporting Information) show the IEF gels obtained for each of the sarcoplasmic extracts before and after the different heat treatments. The temperature selected for the purification was 70 °C, which produced the best results without an appreciable alteration of PRVBs.

Purified supernatants, containing PRVBs as the most abundant proteins, were then analyzed by FTICR–MS. The high-resolution MS spectra allowed for the determination of the different charge states from each PRVB. Figure 3a shows the distribution of the multiple charge states for a representative case (PRVBs isoforms from *M. merluccius*); as shown, a total of nine different charge states (from $z = 7$ to $z = 15$) could be observed. Figure 3b shows the most intense peaks corresponding to the $z = 12$ ions. Thus, monoisotopic mass for each of the PRVB isoforms was calculated from at least 5 different charge states (Figure 3c).

As shown in Table 5, an excellent agreement was found between theoretical and experimental masses for the 10 PRVBs that were completely sequenced by the Bottom-Up approach, yielding accuracies in the M_r measurements between 0.11 and

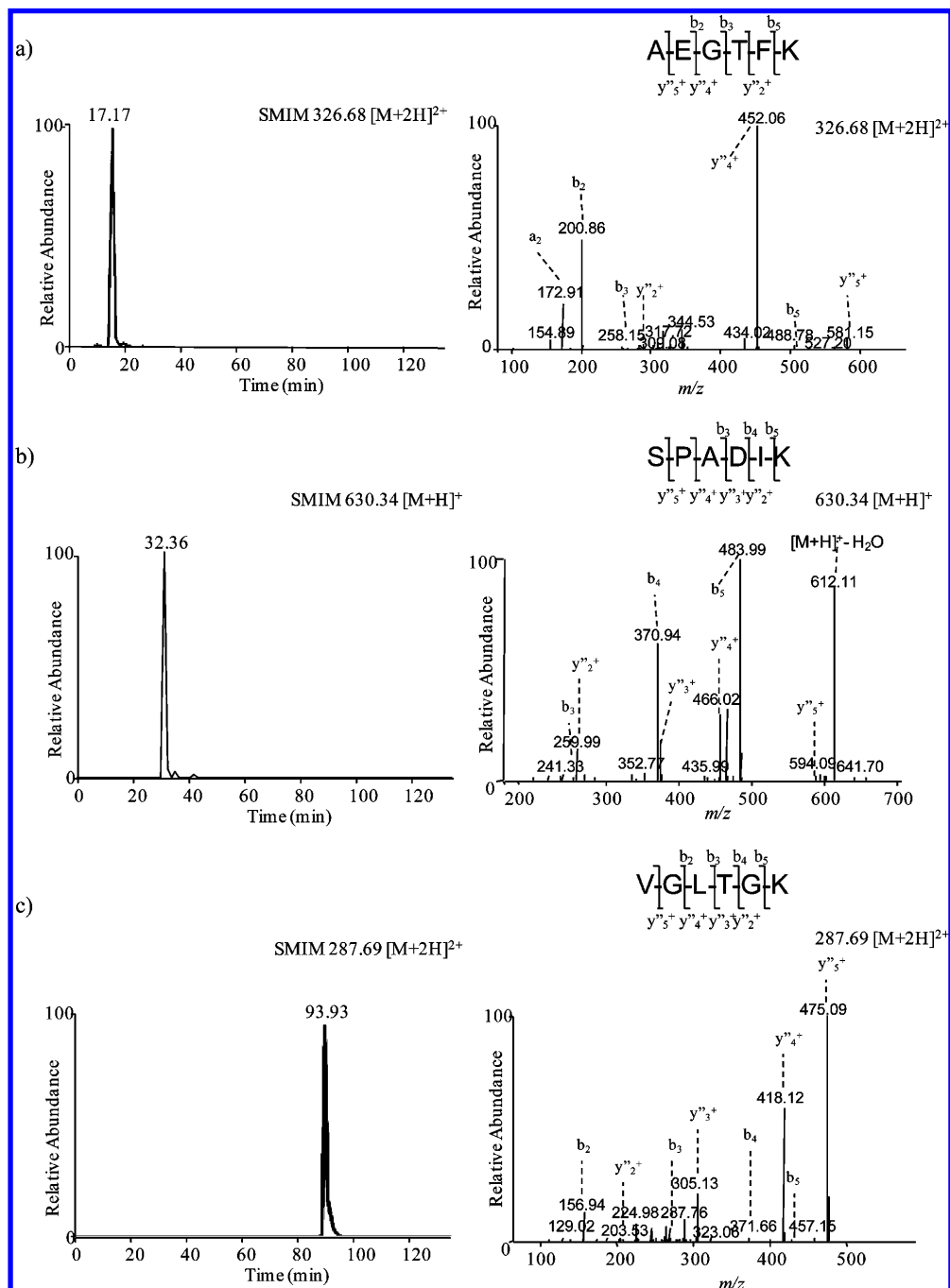


Figure 4. Selected MS/MS ion monitoring (SMIM) of different predicted peptide gaps in the tryptic digests.

4.47 ppm (Table 5). This degree of accuracy allowed the unequivocal confirmation of all amino acid assignments, thus finally demonstrating that these 10 PRVBs were correctly sequenced.

4. Selected MS/MS Ion Monitoring (SMIM) of Peptide Mass Gaps. The comparison between theoretical and experimental M_r was also very useful to determine the mass gaps in those PRVBs for which the bottom-up approach produced an incomplete but higher than 90% coverage (Table 5). These peptide mass gaps were used to predict the mass of putative tryptic peptides that should be present in the different isoforms for these proteins to have the measured M_r . These masses were used to generate a list of singly or doubly charged precursor ions in each case. These ions were monitored and fragmented

by SMIM method¹⁹ (3 replicates per sample) (Table 6 and Figure 4). The SMIM is a targeted mass spectrometry approach similar to MRM in a triple quadrupole machine but produces complete MS/MS spectra and is particularly suitable to be performed in an ion trap MS. It should be noted that in some cases precursor ion masses were calculated taking into account the amino acid residues previously observed by the Bottom-Up approach. Thus, in the case of P10 from *M. polli*, the monoisotopic singly (m/z : 652.37) and doubly charged ions (m/z : 326.68), corresponding to the tryptic peptide AE-[433.2742], were monitored. The MS/MS spectra obtained by SMIM were analyzed by automatic and manual *de novo* sequencing, as described above, and the final peptide se-

Table 7. Theoretical and Experimental M_r for PRVB Isoforms^a

	PRVB spots	theoretical M_r (Bottom-Up) (Da)	experimental M_r (FTICR) (Da)	$\Delta M_r/M_r$ (ppm)
PRVB_A				
Complete seq.				
Best match with β -PRVB <i>M. merluccius</i> (P02620, PRVB_MERME)				
	P1	11329.7473	11329.7460	0.11
	P1b	11315.7316	11315.7370	0.47
	P4	11329.7473	11329.7416	0.49
	P7	11329.7473	11329.7440	0.29
	P10	11329.7473	11329.7890	3.67
	P13	11355.7629	11355.8076	3.93
	P16	11329.7473	11329.7479	0.05
	P19	11329.7473	11329.7490	0.15
	P23	11302.7000	11302.7150	1.33
	P26	11272.7258	11272.7763	4.47
	P27	11302.7000	11302.7144	1.27
	P30	11373.7735	11373.7940	1.80
	P33b	11271.7054	11271.6845	1.85
PRVB_B				
Complete seq.				
Best match with β -PRVB <i>M. bilinearis</i> (P56503, PRVB_MERBI)				
	P2	11379.7847	11379.8025	1.56
	P5	11379.7847	11379.8119	2.39
	P8	11379.7847	11379.8196	3.07
	P11	11349.8105	11349.8483	3.32
	P14	11365.7690	11365.8076	3.39
	P17	11354.7894	11354.8086	1.68
	P20	11354.7894	11354.8095	1.76
	P24	11354.7894	11354.8058	1.44
	P28	11354.7894	11354.8027	1.17
	P31	11310.7996	11310.8227	2.04
	P33	11354.7894	11354.7956	0.54
	P36	11341.7149	11341.7286	1.21
	P39	11341.7149	11341.7252	0.90

^a All these isoforms were completely sequenced.

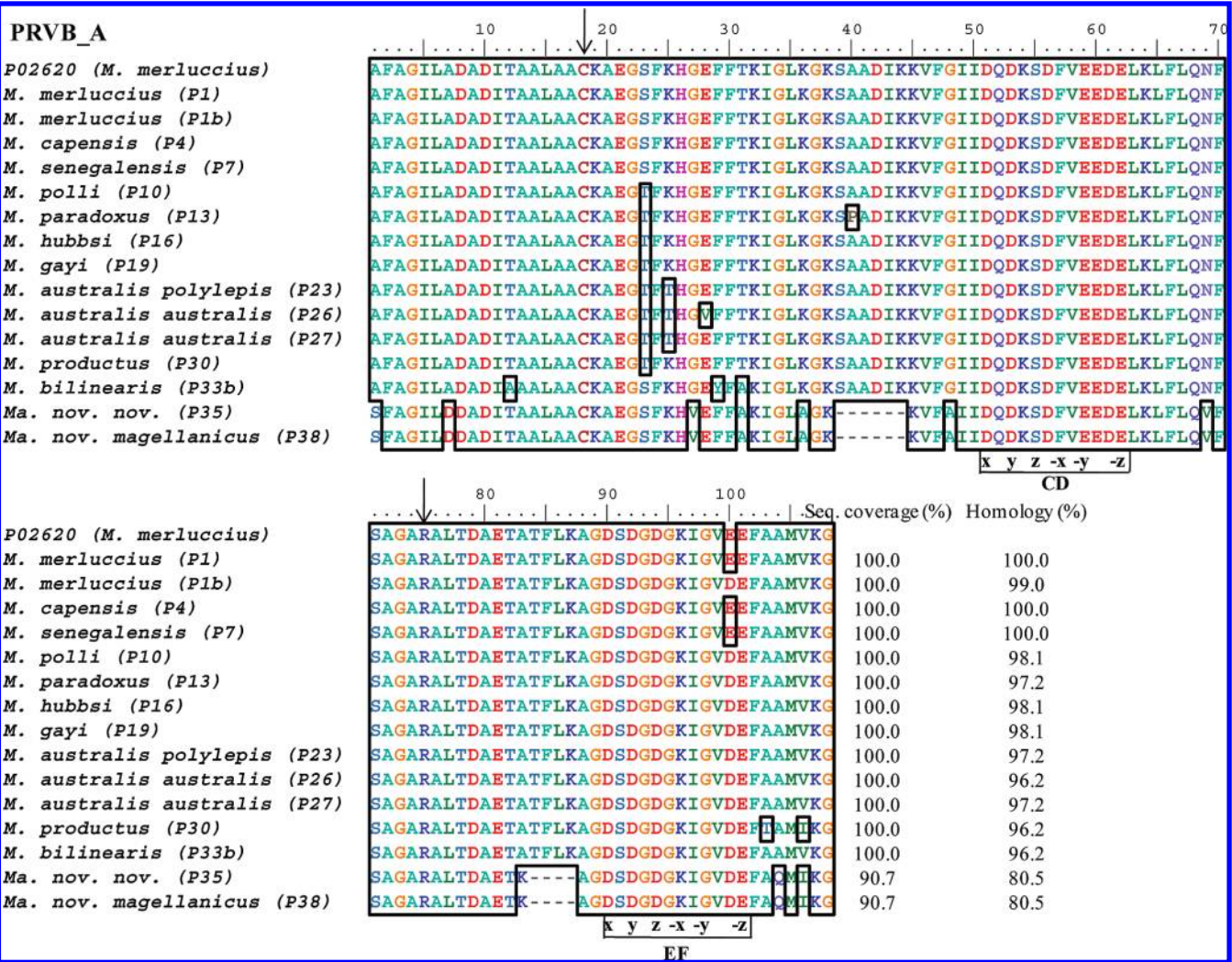


Figure 5. Final alignment of all PRVB_A amino acid sequences for all species studied. Seq. coverage (%), percentage of sequence coverage; (—), amino acid not determined.

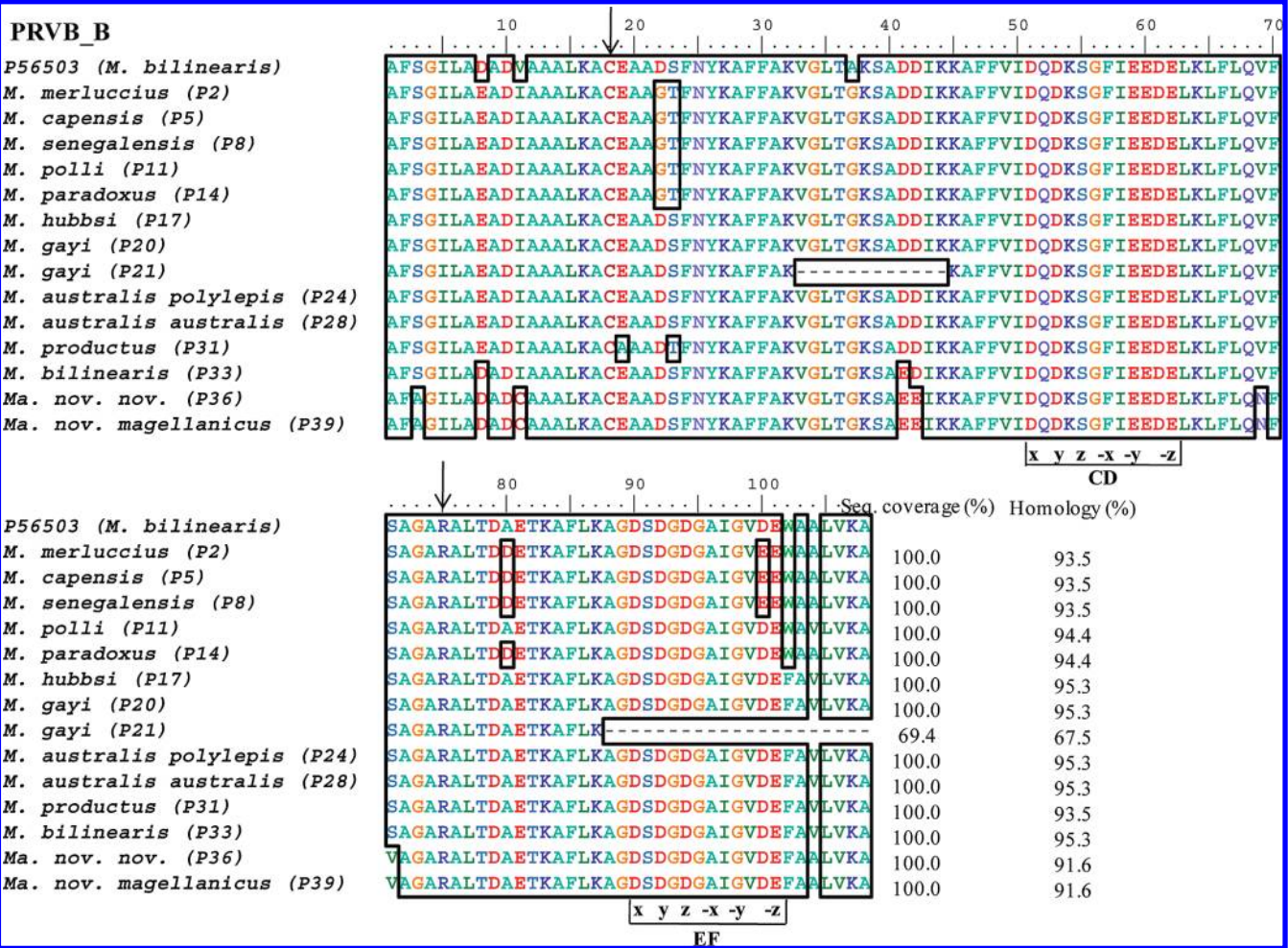


Figure 6. Final alignment of all PRVB_B amino acid sequences for all species studied. Seq. coverage (%), percentage of sequence coverage; (–), amino acid not determined.

quences were combined and aligned to complete the sequence coverage of these proteins (Table 6).

With the information produced by this targeted mass spectrometry approach, the complete sequence of an additional set of 16 new PRVB isoforms, 5 from PRVB_A (P10, P13, P16, P19, P30) and 11 from PRVB_B (P2, P5, P8, P11, P14, P17, P20, P24, P28, P31, P33), could be obtained.

5. Alignment of All the PRVB Sequences. By integrating the various strategies described in the preceding paragraphs, the complete *de novo* MS-sequencing of 25 new PRVB isoforms was obtained. The majority of these sequences corresponded to PRVB_A and PRVB_B isoforms. The accuracy of the agreement between the experimental and the theoretical *M_r* for the 25 new complete PRVBs isoforms (errors in the range between 0.05–4.47 ppm) guaranteed the correct *de novo* interpretation of fragmentation spectra (Table 7). These sequences, together with the other 16 PRVB isoforms that were partially sequenced (53.7–90.7% of coverage), were aligned and compared by BLAST with the PRVBs described in the protein databases presenting the highest homology degree. These final alignments are shown in Figures 5–7.

The degree of homology between the sequences found and those contained in the databases varied in function of the species and the isoform types. Thus, the homology among the PRVB_A studied and the β -PRVB from the *M. merluccius* (PRVB_MERME) varied between 80.5 and 100%, being higher

(100%) for the spots P1, P4 and P7 from *M. merluccius*, *M. capensis* and *M. senegalensis*, respectively. Taking into account only the common amino acid positions completely identified in all of the spots, the analysis of PRVB_A sequences showed the presence of 18 polymorphic sites with two variants each. Homology between PRVB_B isoforms and β -PRVB from *M. bilinearis* (PRVB_MERBI) varied in the range of 67.5% (P21) to 95.3% (P17, P20, P24, P28, P33). None of the PRVB_B isoforms sequences presented a 100% identity, even with the P33 spot sequenced for the same *M. bilinearis* species. As noted previously, both isoforms, the newly sequenced in this work and the one described in the databases, are different. In this group, a total of 16 polymorphic sites were identified. Among those, 13 presented two variants, while the other 3 presented three possible variants. Finally, when PRVB_C isoforms were aligned with β -PRVB from *Theragra chalcogramma* (PRVB1_THECH), the degree of homology was very variable in the range between 43.5% (P6, P9) and 75.9% (P15). Thus, 23 polymorphic sites were found in the sequenced regions, 18 of them with two variants and 3 with three possible variants.

All of the isoforms sequenced had the same number of amino acids, 108, characteristic of the β -lineage from PRVBs. Previously known conserved residues in the β -lineage were also identified in this work:³⁰ Cys18 and Arg75 (Figures 5–7).

PRVBs constitute one of the 32 subclasses within the EF-hand superfamily, a group of proteins with several highly

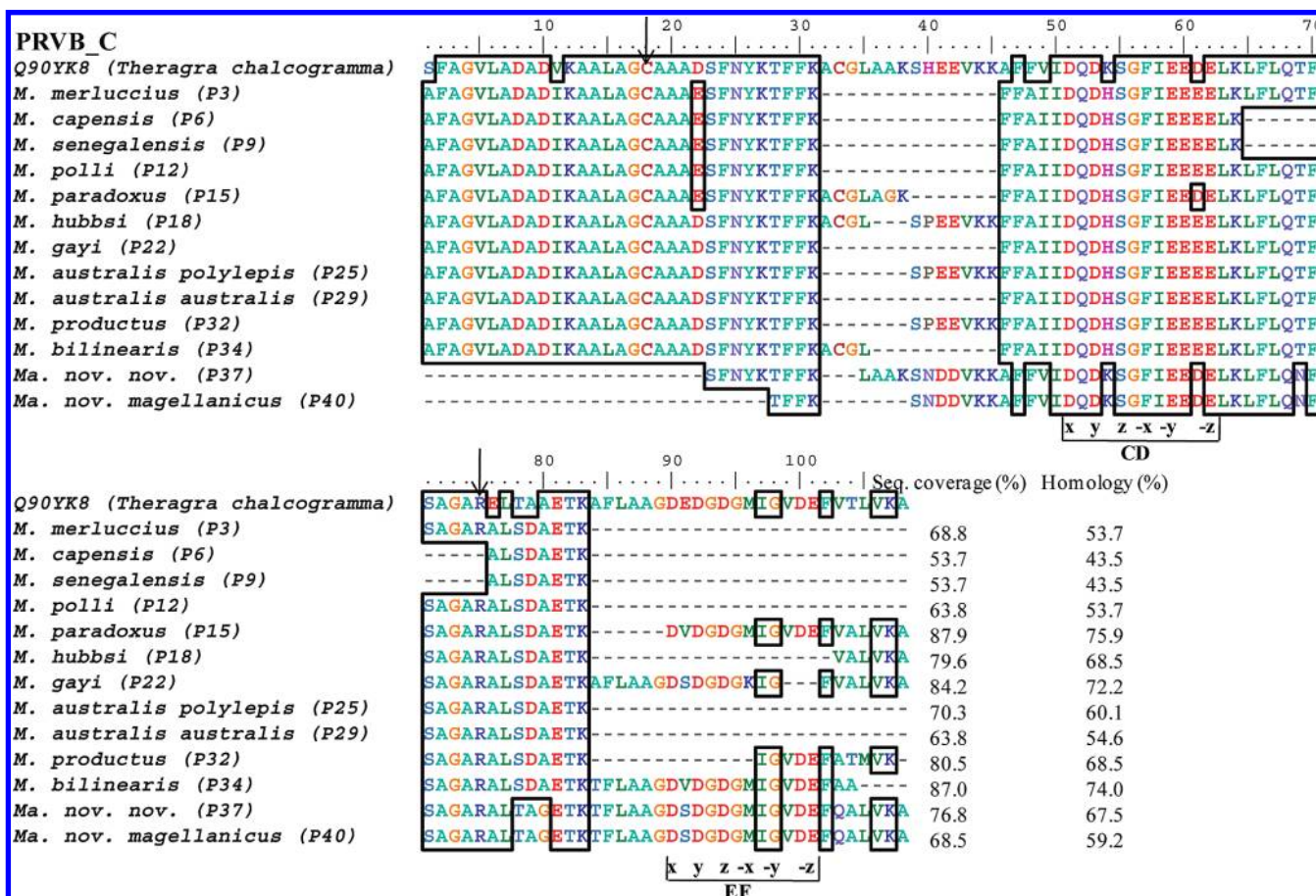


Figure 7. Final alignment of all PRVB_C amino acid sequences for all species studied. Seq. coverage (%), percentage of sequence coverage; (—), amino acid not determined.

conserved helix–loop–helix (EF-hand) motifs,³¹ which bind both Ca^{2+} and Mg^{2+} . PRVBs comprise three such regions, known as AB, CD, and EF. Among them, only two (CD and EF) are functional for chelating Ca^{2+} , containing a central 12-residue Ca^{2+} -binding loop with the flanking two α -helix positioned roughly perpendicular to each other. The Ca^{2+} ion is coordinated by conserved residues located in position (x, y, z, -x, -y, -z) of the binding loop.³² The *de novo* sequencing of all of these new PRVB isoforms allowed for the discovery of several conserved residues in both EF-hand motifs (Figures 5–7). Thus, in the CD domain, all residues localized in the Ca^{2+} -binding positions were conserved (x: Asp, y: Asp, z: Ser, -x: Phe, -y: Glu, -z: Glu). For the EF domain, all Ca^{2+} -binding positions were also conserved: (x: Asp, y: Asp, z: Asp, -y: Gly, -z: Glu), expect for the position (-x), in which three possible different residues (Lys, Ala, Met) were found.

All of these new sequences will be deposited in the UniProtKB protein database.

Conclusions

In this study, a novel strategy for the complete *de novo* sequencing of proteins is described. It is based on the integration of three different proteomics approaches: the classical Bottom-Up, the accurate M_r determination of intact proteins by FTICR–MS and the monitoring of peptide mass gaps by SMIM. In comparison with other hybrid strategies,¹⁸ the inclusion of a targeted proteomic approach by SMIM¹⁹ allowed the completion of the sequence of a larger number of new

proteins. With this novel strategy a total 25 new PRVB isoforms (11.27–11.38 kDa) from all of the important commercial fish species from the Merlucciidae family were unequivocally and completely sequenced. The accuracy of the agreement between the theoretical and the experimental M_r for the 25 new complete PRVBs isoforms (0.05–4.47 ppm) guarantee their correct *de novo* sequencing. In addition, 16 other new PRVB isoforms were partially sequenced (53.7–90.7% of sequence coverage). At present, there are no references describing the complete *de novo* sequencing of such a large number of proteins using different MS approaches. Therefore, the information produced by this work, 41 new PRVB sequences (25 complete and 16 partial sequences), constitute a considerable increase in the number of proteins characterized for this family.

Acknowledgment. We express our gratitude to Mrs. Lorena Barros for her excellent technical assistance and to Dr. Manuel Marcos (CACTI, University of Vigo) for his helpful assistance and suggestions in the FTICR–MS analysis. We also acknowledge Freiremar SA and CETMAR for their assistance in the collection of hakes and grenadiers used in this study. This work was supported by the Comisión Interministerial de Ciencia y Tecnología (CICyT) (Project AGL2000-0440-P4-02).

Supporting Information Available: Supplemental Data 1: 2-DE gels showing the PRVB spots from all species studied. Supplemental Data 2: Peptides identified by SEQUEST. Supplemental Data 3: Alignment of PRVB peptide sequences

identified by SEQUEST. Supplemental Data 4: PEAKS results. Supplemental Data 5: DeNovoX results. Supplemental Data 6: Complete peptide lists after SEQUEST and *de novo* analysis. Supplemental Data 7: Alignment of PRVB peptide sequences identified by SEQUEST and *de novo*. Supplemental Data 8: IEF gels of sarcoplasmic extracts heated to different temperatures. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Lodish, H.; Berk, A.; Zipursky, L. S.; Matsudaira, P.; Baltimore, D.; Damell, K. *Molecular and Cellular Biology*; WH Freeman & Company: New York, 2000.
- (2) Edman, P. A method for the determination of amino acid sequence in peptides. *Arch. Biochem.* **1949**, *22* (3), 475.
- (3) Standing, K. G. Peptide and protein *de novo* sequencing by mass spectrometry. *Curr. Opin. Struct. Biol.* **2003**, *13* (5), 595–601.
- (4) Carrera, M.; Cañas, B.; Piñeiro, C.; Vázquez, J.; Gallardo, J. M. *De novo* mass spectrometry sequencing and characterization of species-specific peptides from nucleoside diphosphate kinase B for the classification of commercial fish species belonging to the family Merlucciidae. *J. Proteome Res.* **2007**, *6* (8), 3070–3080.
- (5) Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **2004**, *5* (9), 699–711.
- (6) Sørensen, S. B.; Sørensen, T. L.; Breddam, K. Fragmentation of proteins by *S. aureus* strain V8 protease. Ammonium bicarbonate strongly inhibits the enzyme but does not improve the selectivity for glutamic acid. *FEBS Lett.* **1991**, *294* (3), 195–197.
- (7) Boersema, P. J.; Taouatas, N.; Altelaar, A. F.; Gouw, J. W.; Ross, P. L.; Pappin, D. J.; Heck, A. J.; Mohammed, S. Straightforward and *de novo* peptide sequencing by MALDI-MS/MS using a Lys-N metalloendopeptidase. *Mol. Cell. Proteomics* **2009**, *8* (4), 650–660.
- (8) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.
- (9) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.
- (10) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.
- (11) Wells, J. M.; McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* **2005**, *402*, 148–185.
- (12) Branca, R. M.; Bodó, G.; Bagyinka, C.; Prokai, L. *De novo* sequencing of a 21-kDa cytochrome c4 from *Thiobacillus roseopersicina* by nanoelectrospray ionization ion-trap and Fourier-transform ion-cyclotron resonance mass spectrometry. *J. Mass Spectrom.* **2007**, *42* (12), 1569–1582.
- (13) Scigelova, M.; Maroto, F.; Dufresne, C.; Vázquez, J. High-throughput *de novo* sequencing. 2007, June 12th <http://www.thermo.com/>.
- (14) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–2342.
- (15) Hopper, S.; Johnson, R. S.; Vath, J. E.; Biemann, K. Glutaredoxin from rabbit bone marrow. Purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *J. Biol. Chem.* **1989**, *264* (34), 20438–20447.
- (16) Whiteaker, J. R.; Warscheid, B.; Pribil, P.; Hathout, Y.; Fenselau, C. Complete sequences of small acid-soluble proteins from *Bacillus globigii*. *J. Mass Spectrom.* **2004**, *39* (10), 1113–1121.
- (17) Kelleher, N. L. Top-down proteomics. *Anal. Chem.* **2004**, *76* (11), 197A–203A.
- (18) Ma, M.; Chen, R.; Ge, Y.; He, H.; Marshall, A. G.; Li, L. Combining bottom-up and top-down mass spectrometric strategies for *de novo* sequencing of the crustacean hyperglycaemic hormone from *Cancer borealis*. *Anal. Chem.* **2009**, *81* (1), 240–247.
- (19) Jorge, I.; Casas, E. M.; Villar, M.; Ortega-Pérez, I.; López-Ferrer, D.; Martínez-Ruiz, A.; Carrera, M.; Marina, A.; Martínez, P.; Serrano, H.; Cañas, B.; Were, F.; Gallardo, J. M.; Lamas, S.; Redondo, J. M.; García-Dorado, D.; Vázquez, J. High-sensitivity analysis of specific peptides in complex samples by selected MS/MS ion monitoring and linear ion trap mass spectrometry: application to biological studies. *J. Mass Spectrom.* **2007**, *42* (11), 1391–1403.
- (20) Elsayed, S.; Bennich, H. The primary structure of allergen M from cod. *Scand. J. Immunol.* **1975**, *4* (2), 203–208.
- (21) Bugajska-Schretter, A.; Elfman, L.; Fuchs, T.; Kapiotis, S.; Rumpold, H.; Valenta, R.; Spitzauer, S. Parvalbumin, a cross-reactive fish allergen, contains IgE-binding epitopes sensitive to periodate treatment and Ca²⁺ depletion. *J. Allergy Clin. Immunol.* **1998**, *101* (1), 67–74.
- (22) Sicherer, S. H.; Sampson, H. A. 9. Food allergy. *J. Allergy Clin. Immunol.* **2006**, *117* (2), S470–S475.
- (23) Carrera, M.; Cañas, B.; Piñeiro, C.; Vázquez, J.; Gallardo, J. M. Identification of hake and grenadier species by proteomic analysis of the parvalbumin fraction. *Proteomics* **2006**, *6* (19), 5278–5287.
- (24) Piñeiro, C.; Barros-Velázquez, J.; Sotelo, C. G.; Pérez-Martín, R. I.; Gallardo, J. M. Two-dimensional electrophoretic study of the water-soluble protein fraction in white muscle of Gadoid fish species. *J. Agric. Food Chem.* **1998**, *46* (10), 3991–3997.
- (25) Jensen, O. N.; Wilm, M.; Shevchenko, A.; Mann, M. Sample preparation methods for mass spectrometric peptide mapping directly from 2-DE gels. *Methods Mol. Biol.* **1999**, *112*, 513–530.
- (26) Capony, J. P.; Rydén, L.; Demaille, J.; Pechère, J. F. The primary structure of the major parvalbumin from hake muscle. Overlapping peptides obtained with chemical and enzymatic methods. The complete amino-acid sequence. *Eur. J. Biochem.* **1973**, *32* (1), 97–108.
- (27) Revett, S. P.; King, G.; Shabanowitz, J.; Hunt, D. F.; Hartman, T. M.; Nelson, D. J. Characterization of a helix-loop-helix (EF hand) motif of silver hake parvalbumin isoform B. *Protein Sci.* **1997**, *6* (11), 2397–2408.
- (28) Meinel, T.; Peynot, P.; Giglione, C. Processed N-termini of mature proteins in higher eukaryotes and their major contribution to dynamic proteomics. *Biochimie* **2005**, *87* (8), 701–712.
- (29) Kawai, Y.; Uematsu, S.; Shinano, H. Effect of heat-treatment on some physicochemical properties and emulsifying activity of carp sarcoplasmic protein. *Nippon Suisan Gakkai.* **1992**, *58* (7), 1327–1331.
- (30) Kretsinger, R. H. Structure and evolution of calcium-modulated proteins. *CRC Crit. Rev. Biochem.* **1980**, *8* (2), 119–174.
- (31) Ikura, M. Calcium binding and conformational response in EF-hand proteins. *Trends Biochem. Sci.* **1996**, *21* (1), 14–17.
- (32) Kawasaki, H.; Kretsinger, R. H. Calcium-binding proteins. 1: EF-hands. *Protein Profile* **1994**, *1* (4), 343–517.

PR100163E

SUPPLEMENTAL DATA 1

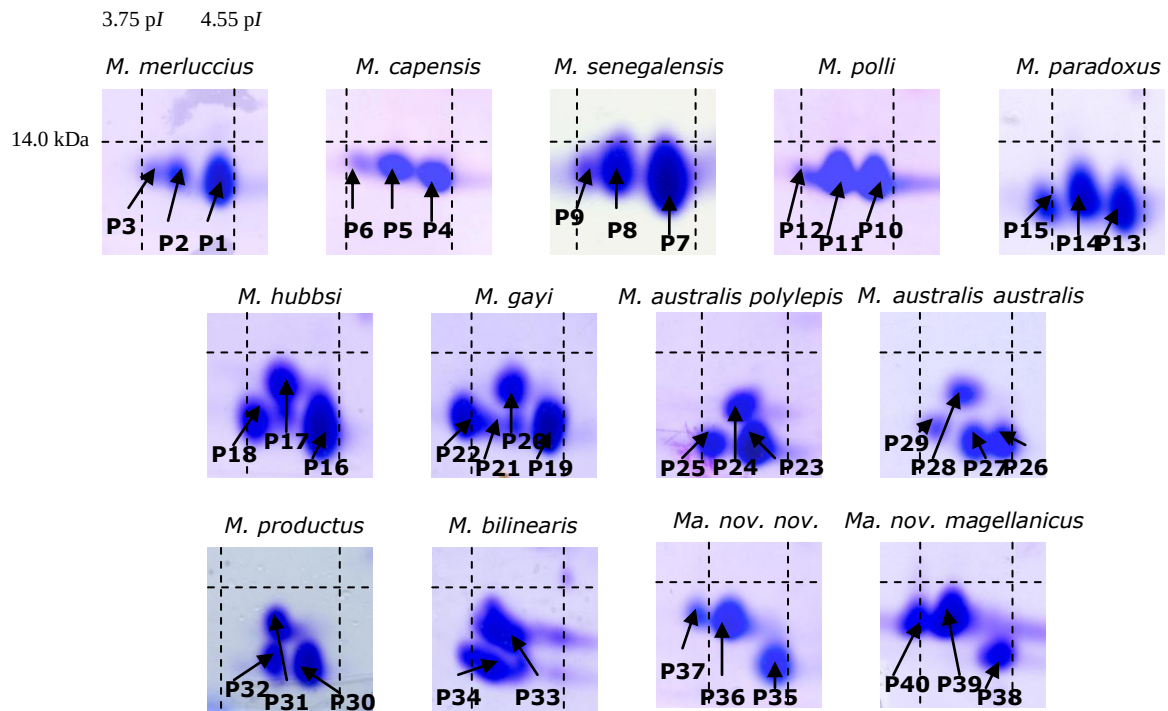
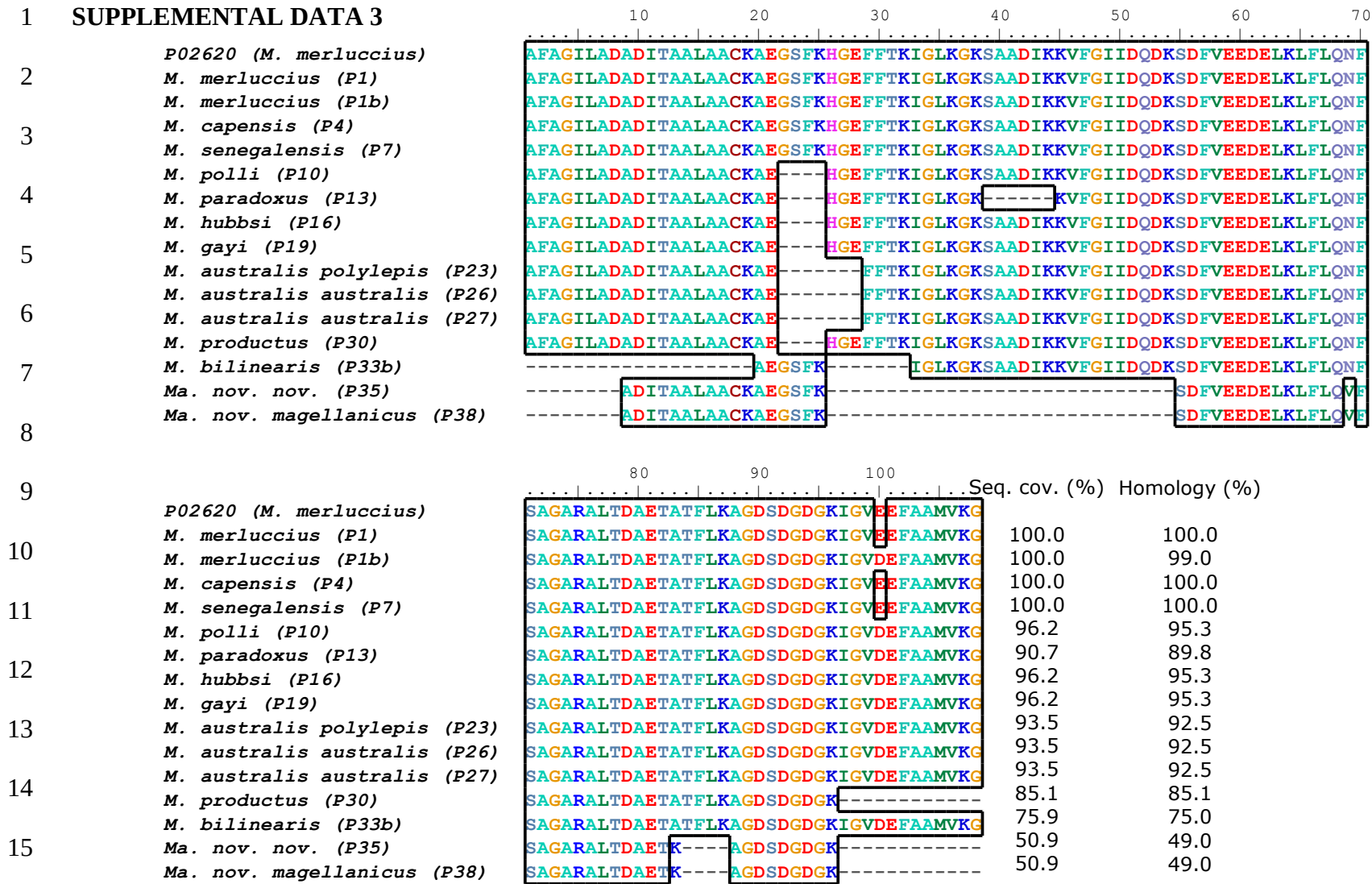
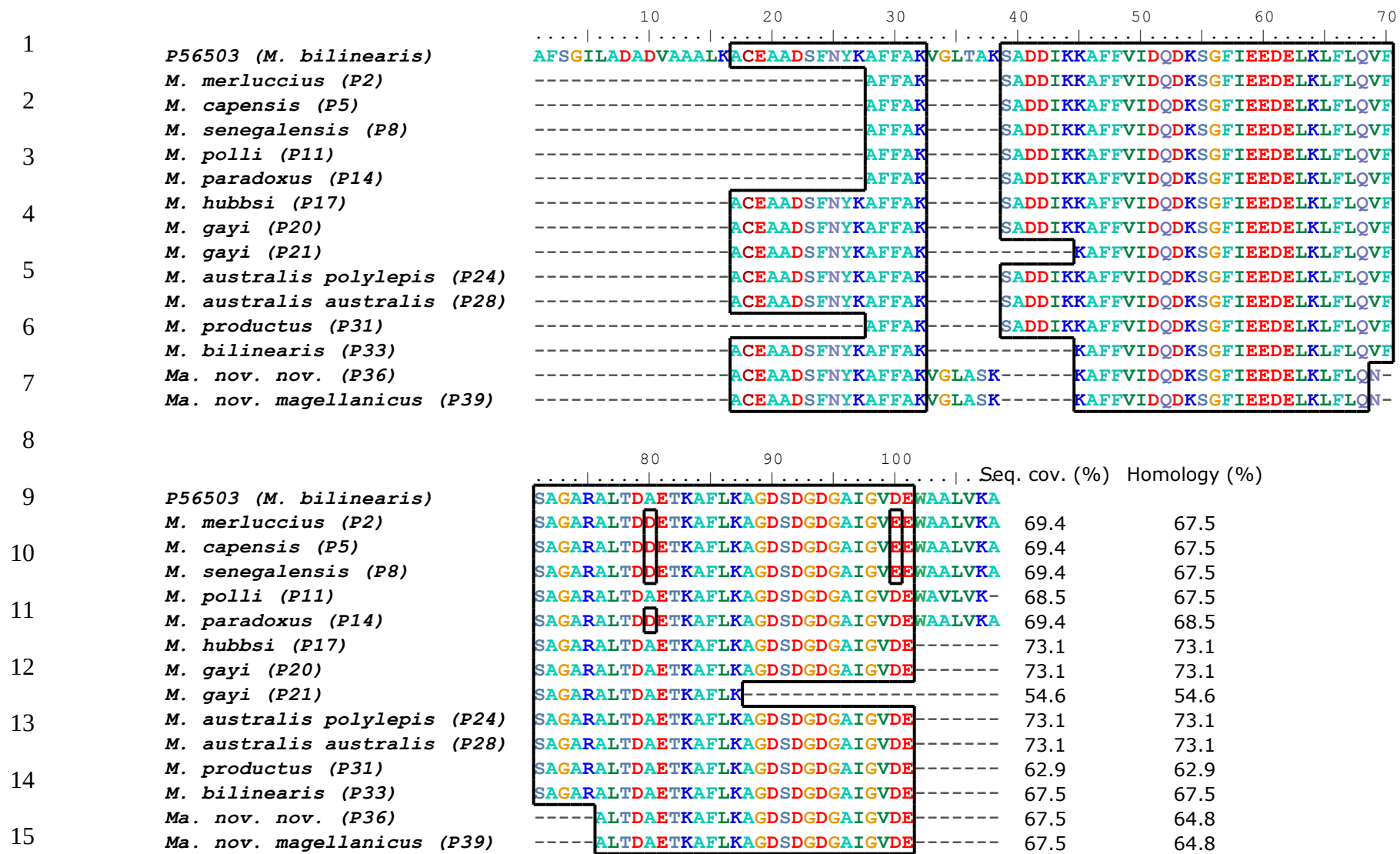


Figure 1. Cropped regions of representative Coomassie brilliant blue-stained 2-DE gels showing the separated PRVB spots from all species studied. The different isoforms from PRVBs are indicated with different Arabic numbers. *M.* (*Merluccius* spp.); *Ma. nov.* (*Macruronus novaezelandiae*).

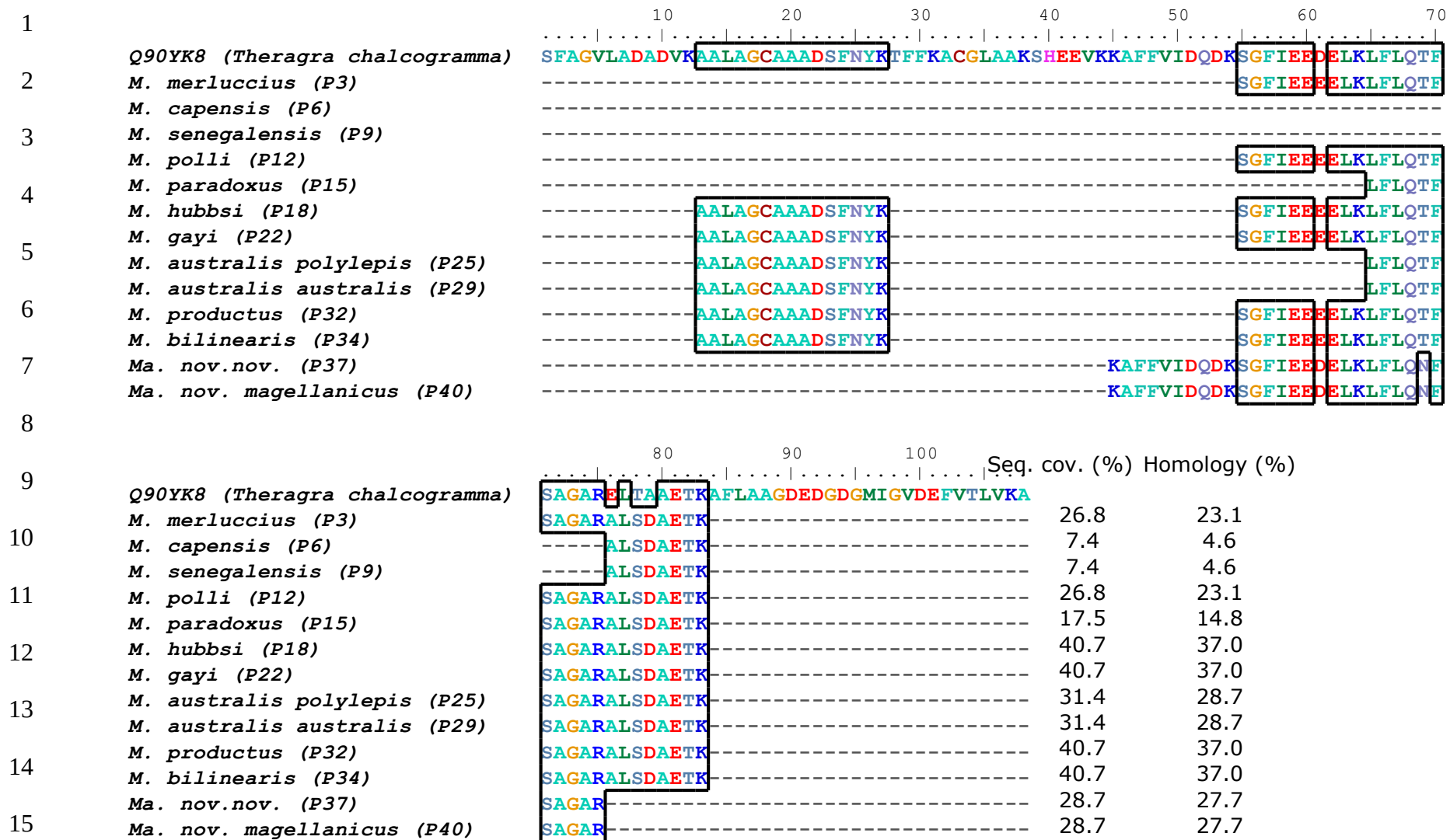


16 **Figure 1.** Alignment of all PRVB_A amino acid sequences obtained by LC-ESI-IT-MS/MS and Sequest search. The sequences were aligned

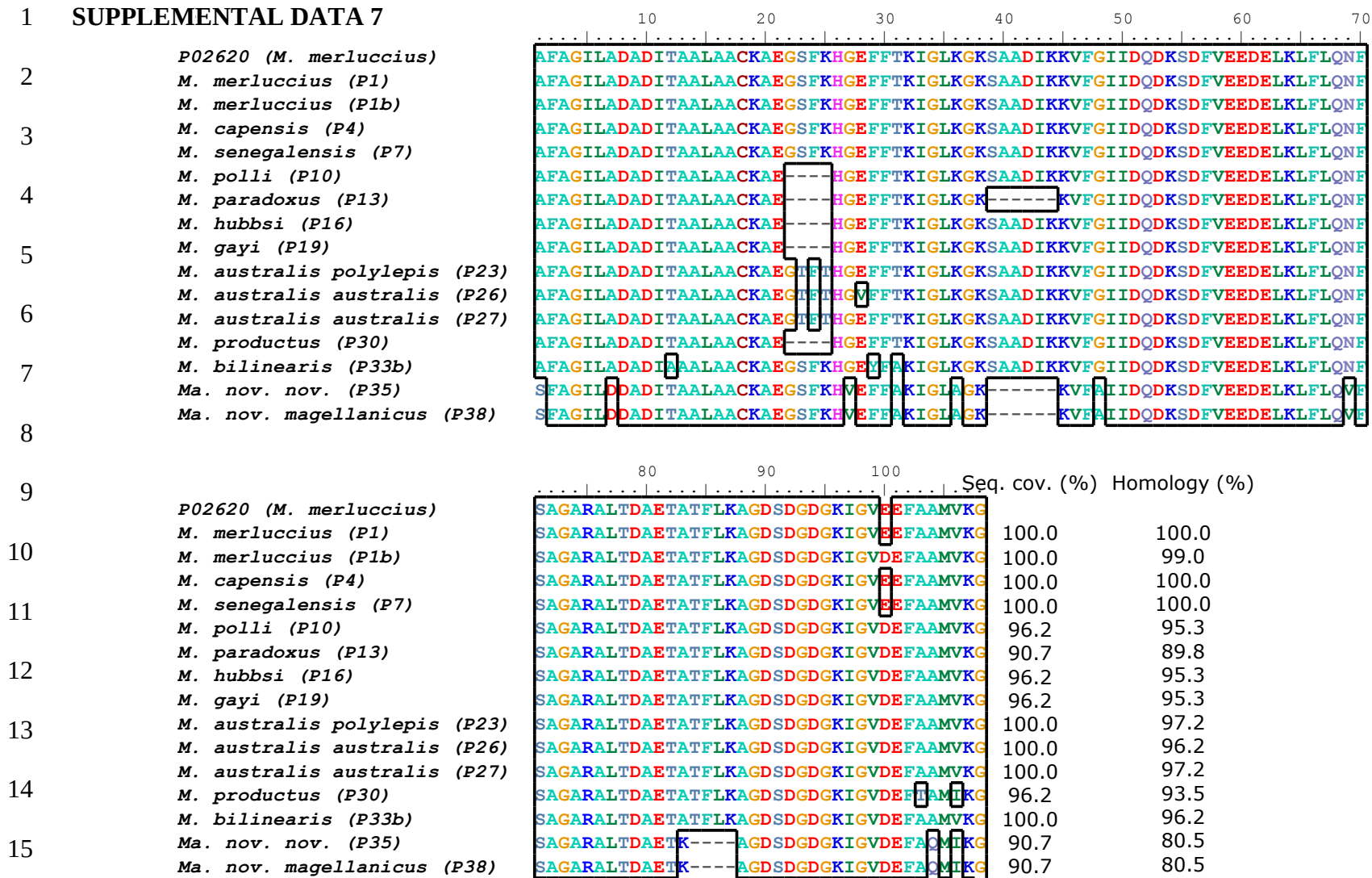
17 with the β -PRVB from *M. merluccius* (P02620, PRVB_MERME). (-) no determined amino acid. Seq. cov. (sequence coverage).



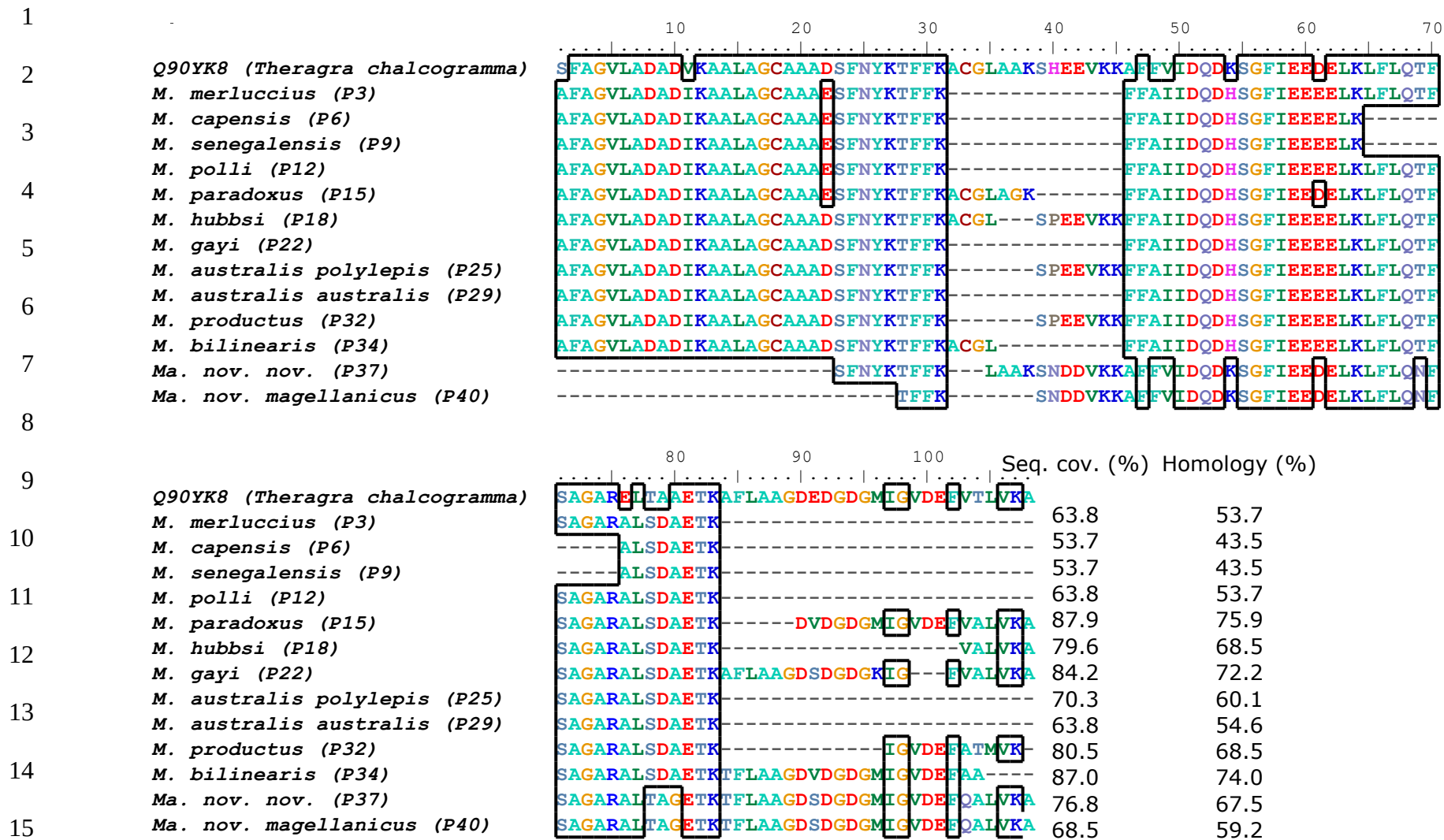
16 **Figure 2.** Alignment of all PRVB_B amino acid sequences obtained by LC-ESI-IT-MS/MS and Sequest search. The sequences were aligned
 17 with the β -PRVB from *M. bilinearis* (P56503, PRVB_MERBI). (-) no determined amino acid. Seq. cov. (sequence coverage).



16 **Figure 3.** Alignment of all PRVB_C amino acid sequences obtained by LC-ESI-IT-MS/MS and Sequest search. The sequences were aligned
 17 with the β -PRVB from *Theragra chalcogramma* (Q90YK8, PRVB1_THECH). (-) no determined amino acid. Seq. cov. (sequence coverage).



16 **Figure 1.** Alignment of all PRVB_A amino acid sequences obtained by the Bottom-Up approach (Sequest and *de novo*). The sequences were
17 aligned with the β -PRVB from *M. merluccius* (P02620, PRVB_MERME). (-) no determined amino acid. Seq. cov. (sequence coverage).



16 **Figure 3.** Alignment of all PRVB_C sequences obtained by the Bottom-Up approach (Sequest and *de novo*). The sequences were aligned with
 17 the β -PRVB from *Theragra chalcogramma* (Q90YK8, PRVB1_THECH). (-) no determined amino acid. Seq. cov. (sequence coverage).

SUPPLEMENTAL DATA 8

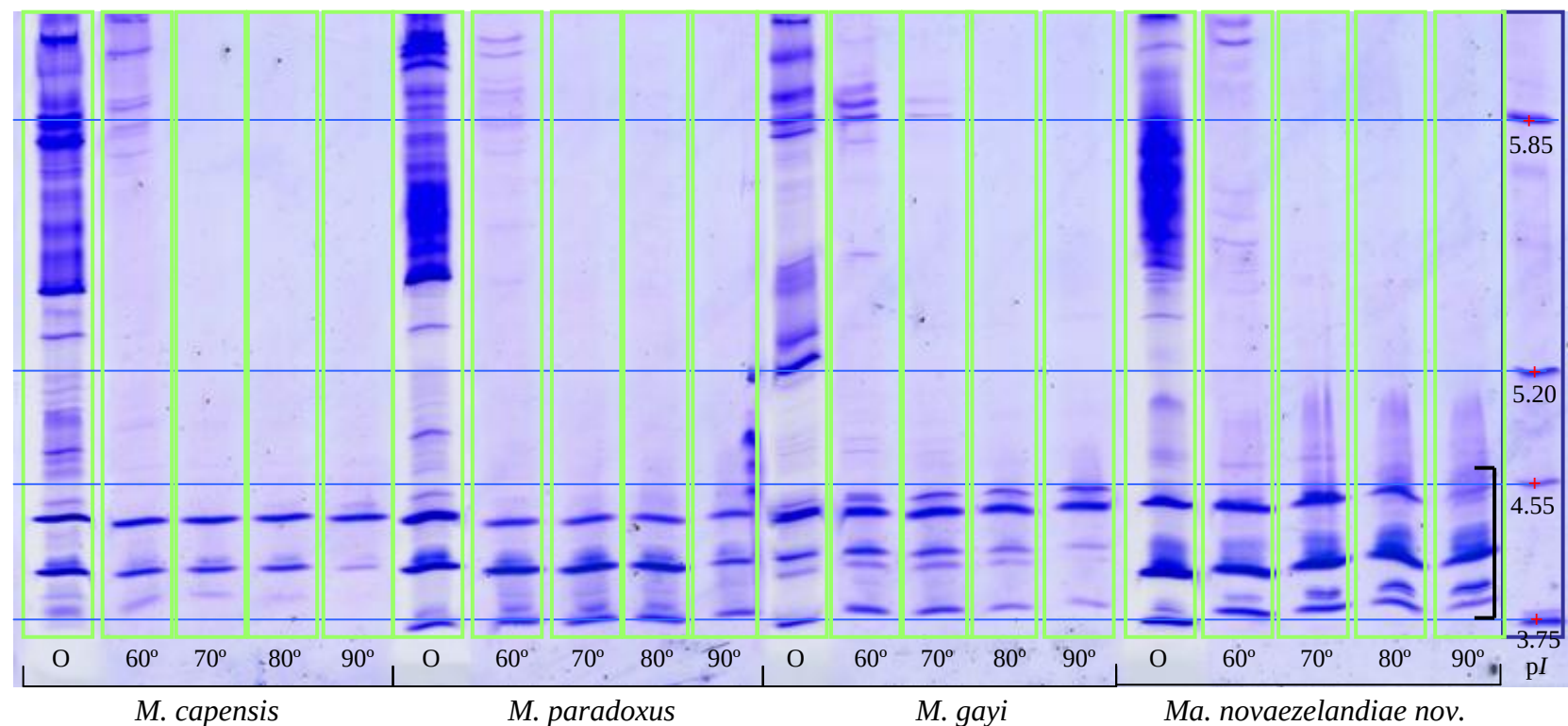


Figure 1. Comparison by IEF (4.0-6.5 pH) of sarcoplasmic extracts heated to different temperatures (60°C-90°C). O (original sarcoplasmic extract). The bar indicates the region where are resolved the PRVBs.

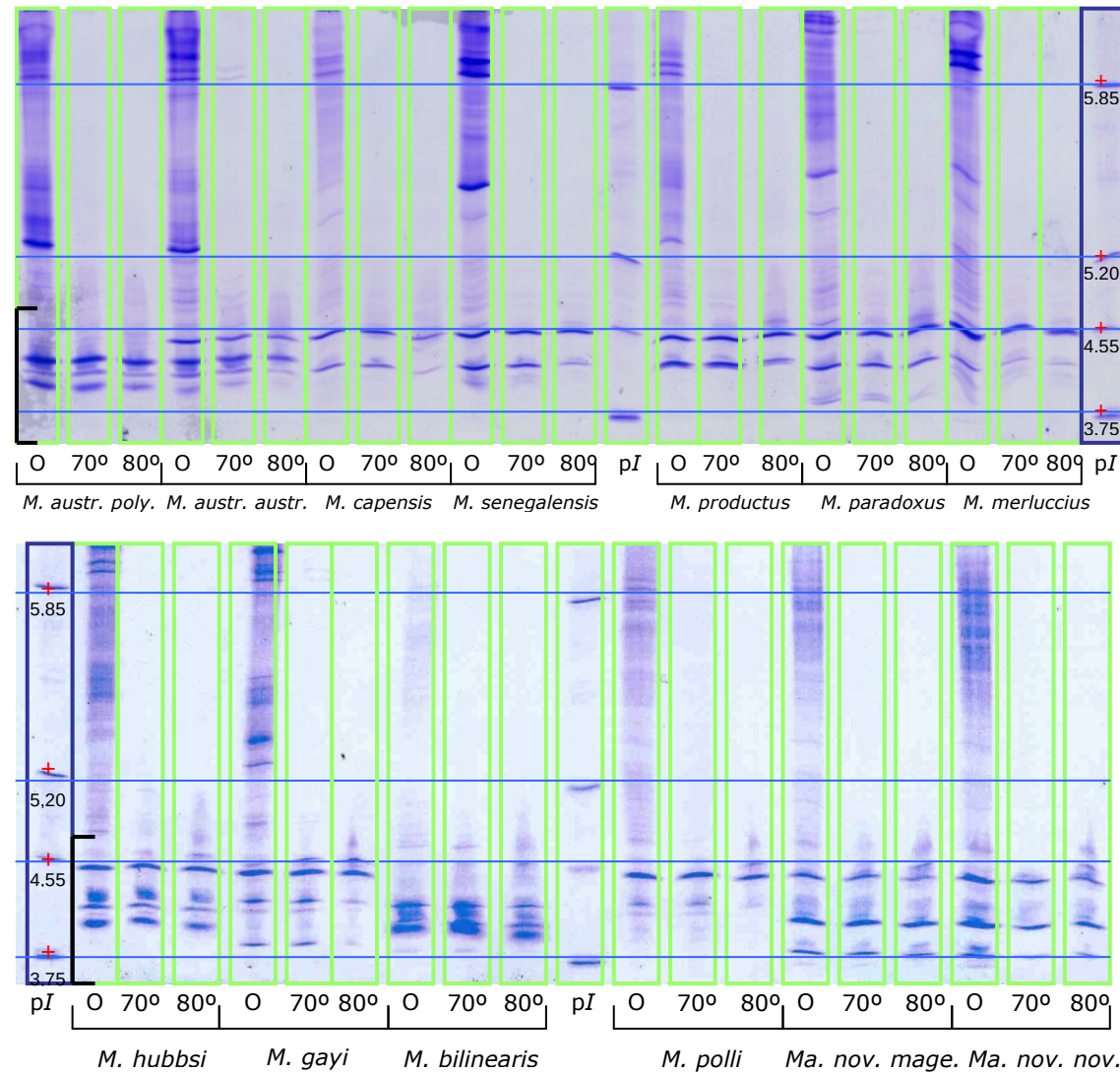


Figure 2. Comparison by IEF (4.0-6.5 pH) of sarcoplasmic extracts heated to different temperatures (70°C-80°C). O (original sarcoplasmic extract). The bar indicates the region where are resolved the PRVBs.