Complete sequence and characterization of the major sperm nuclear basic protein from *Mytilus trossulus*

Corinne Rocchini, Philip Rice, Juan Ausio*

Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Victoria, BC, V8W 3P6, Canada

Received 27 January 1995; revised version received 7 March 1995

Abstract **We have characterized the major protamine-like protein (PL-III) from the sperm of the mussel *Mytilus trossulus*. The molecular mass of the main protein component of PL-III, determined by mass spectrometry using fast atom bombardment and matrix-assisted laser desorption ionization was 11304 ± 6 Da. The complete protein sequence was established by Edman degradation and the molecular mass derived from this sequence coincides with that experimentally determined. The protein has a sedimentation coefficient $s_{20,w}$ = 1.15 ± 0.5 S which is consistent with a random coil of radius of gyration 34.3 ± 1.4 Å.**

Key words: Sperm nuclear basic protein; Protein characterization; *Mytilus*

1. Introduction

In 1969, David Bloch published the first thorough classification of the sperm nuclear basic proteins (SNBP) [1]. In this early classification, SNBPs were arranged in five major types: Crab type, consisting of relatively non-basic proteins (such as those found in the sperm of several crustaceans), Rana type, consisting of somatic-like sperm histones, Salmon type (Monoprotamines), consisting of protamines lacking cysteine, Mouse type (Stable protamines), consisting of the cysteine containing protamines found in the sperm of eutherian mammals and Mytilus type, consisting of sperm proteins with an amino acid composition intermediate between protamines and histones.

Although an extensive sequence analysis has been carried out on both the monoprotamine and stable protamine groups and to lesser extent on the sperm-specific histone group, there is still very little information available on the primary structure of the compositionally intermediate SNBPs of the Mytilus type.

In 1973 Subirana and co-workers [2] carried out a rather comprehensive analysis of the SNBPs from different molluscs including *Mytilus*. The nuclear sperm protein composition of this later organism was shown to consist of three major proteins: ϕ2B, ϕ1, and ϕ3. The nomenclature of these proteins was adopted following that of somatic histones and based on their similar extractability with the organic solvents used at that time to fractionate histones [3]. The letter H of histones was replaced by ϕ to indicate the sperm origin of these proteins. A few years ago, based on a more thorough biochemical and structural characterization, we adopted a protamine-like (PL) nomenclature for these proteins PL-II* (ϕ2B), PL-III (ϕ1), and PL-IV (ϕ3) [4]. Recently we have published the complete sequence of PL-II* and PL-IV [5,6]. In the present paper, we report the complete sequence of the PL-III component which is the most abundant chromosomal protein of the sperm of *Mytilus* [7].

2. Materials and methods

2.1. Living organisms

Specimens of *Mytilus trossulus*, Gould, 1850, were collected at Esquimalt Lagoon (Victoria, BC).

2.2. Gel electrophoresis

Acetic acid–urea–Triton (AUT) polyacrylamide gels were prepared as described elsewhere [6].

2.3. Sperm collection and protein isolation

Sperm collection, nuclei isolation and HCl extraction of SNBPs was carried out as described previously [4]. The crude HCl protein components were fractionated into their individual PL protein components by conventional ionic exchange chromatography on CM–Sephadex C-25 [5] or using reverse phase HPLC on a VYDAC C_{18} (5 μm)(15 × 0.46 cm) column (see Fig. 1) [5].

2.4. Proteolytic digestions

Protein PL-III at a concentration of ≈ 2 mg/ml in 100 mM ammonium bicarbonate pH 8.0 was digested at room temperature with *Astacus fluviatilis* protease(EC 3.4.99.6) (Serva) (E/S = 1/500, 30 min) or Elastase (EC 3.4.21.36) (Worthington) (E/S = 1/100, 7 min). The protein was also digested with endoproteinase Lys-C (EC 3.4.21.50) (Boehringer) at 1.5 mg/ml in 35 mM Tris-HCl (pH 8.0) (E/S = 1/100, 20 min) at room temperature. Cleavage at Valine was achieved by digestion with thermolysin (EC 3.4.24.4) (type X, Sigma) (E/S = 1/200) in 100 mM Ammonium bicarbonate pH 8.0 buffer at a protein concentration of 1.4 mg/ml for 60 min at room temperature.

Immediately after digestion the resulting peptide mixtures were injected directly onto a Vydac C_{18} column and the peptides were fractionated using a linear 0–20% (120 min) acetonitrile gradient in 0.1% trifluoroacetic acid, at a flow rate of 1 ml/min.

2.5. Amino acid analysis and protein sequencing

Amino acid analyses were carried out on an ABI Model 420A derivatizer analyzer system, and the peptide sequencing was performed on an ABI Model 470A gas-phase protein sequenator as described elsewhere [5]. Both types of analyses were carried out at the Protein Microchemistry Center of the University of Victoria, BC.

2.6. Mass spectrometry

The molecular mass of PL-III was determined by matrix-assisted laser desorption ionization using a Kratos Kompact MALDI III V2.0.1 (Kratos Analytical) time of flight spectrometer with a cyano-4-hydroxyciannamic acid matrix [8]. The molecular mass was also determined by fast atom bombardment using a Profile Kratos (Kratos Analytical). In the latter case the sample was dissolved in a solution containing 50% acetonitrile 1% acetic acid in water.

2.7. Sedimentation analysis

Analytical ultracentrifuge analysis was carried out in either a Beckman Model E or a Beckman XL-A analytical ultracentrifuge.

In some instances, sedimentation velocity analysis was carried out on the model E using Schlieren optics and double sector synthetic bound-

*Corrresponding author. Fax: (1) (604) 721 8855.
E-mail: jausio@uvvm.uvic.ca

Abbreviations: SNBP, sperm nuclear basic protein.

SSDI 0014-5793(95)00275-8

ary (Aluminum-filled epon) centerpieces in an An E rotor. Alternatively, the analysis was carried out in the XL-A ultracentrifuge and the sedimenting boundaries were scanned at 230nm. In this later instance double sector (aluminum-filled epon) centerpieces were used in an An-60 Ti rotor. The scans were analyzed by the method of Van Holde and Weischet [9] using the Ultrascan data analysis software (Borries Demeler, San Antonio, TX). Apparent sedimentation coefficients were converted to standard conditions $s_{20,w}$ as described elsewhere [8]. Partial specific volumes were estimated from the amino acid composition [10]. The scans were analysed using the Ultrascan software described above. Conformational analysis was carried out according to Tanford [11]. The buffer used in all these experiments was 150 mM NaCl 10 mM Tris-HCl (pH 7.5).

2.8. Secondary structure prediction

Secondary structure prediction from the primary structure of PL-III was carried out according to Chou and Fasman [12,13] and to Garnier et al. [14] using a MacVector programme (IBI).

3. Results

Fig. 1 shows the electrophoretic pattern and an HPLC profile of the nuclear basic proteins of the sperm from *Mytilus trossulus*. Three major sperm-specific components can be distinguished, PL-II*, PL-III and PL-IV which coexist with a limited amount (ca.10–15%) of somatic-like histones [4]. As it can be seen in this figure, PL-III is the most abundant PL in the genus *Mytilus* and it represents about 60–70% of the PL protein complement [7]. Despite being the most abundant of the three PLs the primary structure of this protein has remained elusive. To date, only partial sequence information was available on PL-III from different *Mytilus* species [5,15].

Fig. 2 shows the proteolytic-cleavage strategy followed to ascertain the sequence of the protein. Several of the peptides

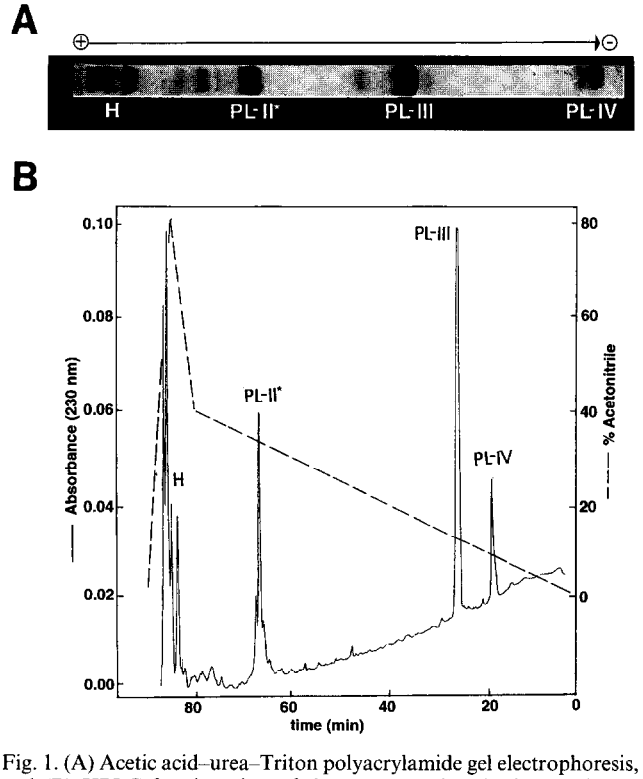
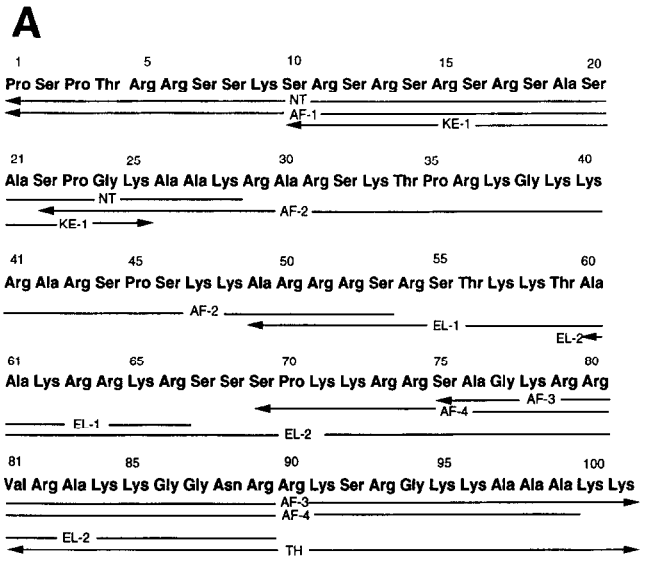


Fig. 1. (A) Acetic acid–urea–Triton polyacrylamide gel electrophoresis, and (B) HPLC fractionation of the sperm nuclear basic proteins of *Mytilus trossulus*. H = histones.



B

M.t. PSPTRRSSKSRSRSRSRS**ASASPGKAAKRARSKTP

M.c. PSPTRRS*KSRSKSRSRSRSASA**GKAAKRAKSKTG

M.e. PSPTRRSSKSRSKSRSRSRSASASPGKAAKRARSKTP

Fig. 2. (A) Complete amino acid sequence of PL-III. AF = peptides obtained by digestion with *Astacus fluviatilis* protease. KE = peptides obtained by digestion with endoproteinase Lys-C. EL = peptides obtained by digestion with elastase. TH = peptides obtained by digestion with thermolysin. (B) Sequence comparison of the N-terminal region of PL-III from: MC = *M. californianus* from reference [5] (partial sequence), ME = *M. edulis* from reference [15] (incomplete sequence) and MT = *M. trossulus*.

generated in this way were then subjected to automated Edman-degradation sequence analysis to generate the complete sequence of the proteins shown in Fig. 2A.

The average molecular mass deduced form this sequence 11305 (11299 monoisotopic) is in very good agreement with that of 11304 ± 6 experimentally determined by mass spectrometry for the main protein component of PL-III (see Fig. 3). Refined fast atom bombardment analysis of the PL-III from *M. trossulus* (data not shown) indicated the presence, in addition, of a minor protein component of molecular mass 11333 ± 4. This molecular mass is also in good agreement with the experimental values previously reported using sedimentation equilibrium and electrophoretic techniques. In fact the number of amino acids of the sequence (101) agrees very well with the number established (100) with these experimental techniques [16].

PL-III has been shown to exhibit some interspecific variability [17]. Indeed, mass spectroscopy analysis revealed that the molecular mass of the major PL-III component of *M. trossulus* is smaller than that of *M. californianus* (11350 ± 10) or *M. edulis* (11580 ± 10).

Fig. 2B shows the partial N-terminal sequence of this protein for the three species. As it can be seen there this region is extremely variable, with most of the variability arising form either deletion or conserved amino acid replacements. Indeed

most of the mass difference between PL-III from *M. trossulus* and *M. edulis* can be accounted for by the compositional differences observed in this region. As a matter of fact comparison of the incomplete amino acid sequence available for *M. edulis* [15] with that of *M. trossulus* presented here shows an identical sequence beyond the first 20 N-terminal amino acids with only one K→R substitution at position 39 (from the N terminus).

Secondary structure prediction analyses [12–14] show considerable discrepancies (especially on the α helix assignment) depending on the algorithm used. This most likely reflects the difficulties encountered by the prediction methods [12–14] (but see discussion) with arginine rich proteins such as PL-III [18].

Hydrodynamic analysis (Table 1) allowed us to model the protein as a random coil of radius of gyration 34.3 ± 1.4 Å in solution (see Table 1).

4. Discussion

Very recently Ruiz-Lara et al. [15], using an oligo(dT)-primed cDNA library from gonadal poly(A)$^+$ RNA, published the sequence of a cDNA for the protein PL-III (ϕ1) from the sperm of the mussel *Mytilus edulis*. The cDNA, obtained in this way, spanned an open reading frame of 91 amino acids. As discussed by these authors, this number closely resembled that of 100 previously determined by electrophoretic and sedimentation equilibrium analysis [16]. However, the Mass spectroscopy of PL-III from different species of *Mytilus* (including *edulis*) and Edman degradation sequence analysis of PL-III of *M. trossulus* reported in the preceding section, while in good agreement with most of the cDNA sequence, clearly confirm the early suspicion [19] that the cDNA sequence could have been incomplete.

The extremely adenine rich region (arising from the lysine codons) found towards the 3′ end of the cDNA isolated by Ruiz-Lara et al. [15] (which did not contain termination signal [19]) and the priming strategy used, most likely account for the incomplete protein sequence determined by this method. Nevertheless, as we have mentioned earlier the first 87 amino acids of partial sequence from the cDNA from *M. edulis* [15] are in very good agreement with the protein sequence of PL-III from *M. trossulus* determined by Edman degradation sequencing (Fig. 2A).

Table 1
Conformational parameters of PL-III

Molecular mass[a]	11,300
$\bar{\upsilon}_2$, cm^3/g	0.736
$s_{20,w}$, S	1.15 ± 0.05
ξ_1, g H_2O/g protein	0.41
f/f_0	1.32 ± 0.05
R_0, Å	17.3
R_S, Å	22.8 ± 1
R_G[b]	34.3 ± 1.4

[a] Molecular mass determined by mass spectrometry
$\bar{\upsilon}_2$ partial specific volume ($\bar{\upsilon}_2$) estimated from the amino acid composition of the protein as described in [8].
ξ_1 preferential hydration parameter estimated from the amino acid composition according to Kuntz [25].
f/f_0 frictional ratio [11].
R_0 radius of an equivalent sphere [11].
R_s Stokes radius.
R_G[b] radius of gyration assuming a random coil conformation. [$R_G = R_s/0.665$] [11].

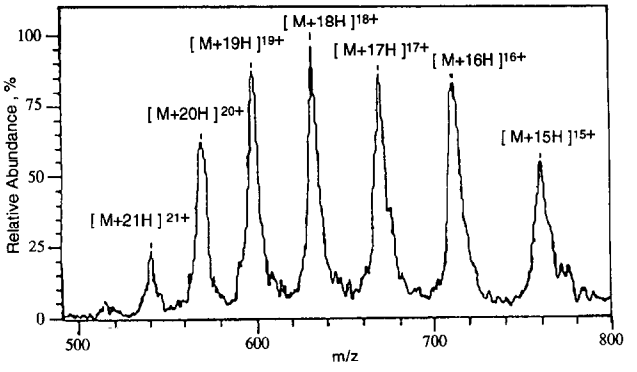


Fig. 3. Fast atom bombardment spectrum of PL-III.

One of the most unique features of PL-III sequence (see Fig. 2) is the presence of a long N-terminal stretch of alternating S/R(K) residues which spans over the first 20 N-terminal amino acids. This sequence has been found in the N-terminal region of the SNBP of other bivalve molluscs form which partial sequences are also available [20]. Also, a highly 'homologous' N-terminal region is present in PL-II* (see Fig. 1) (an H1-like protein) found in the sperm nucleus of *Mytilus* [5]. At this point the possibility (not proven) of alternative splicing mechanism such as it is also observed in some vertebrate protamines [21] becomes very suggestive. Interestingly enough the repetitive S/R motif is strongly reminiscent of the S/R domain of the pre-mRNA splicing and processing regulators from *Drosophila* [22].

The sequence also contains a unique valine residue which, from the amino acid analysis composition, seems to be conserved throughout all the species analyzed thus far [17]. This valine is found within the sequence ...PKKRRsagkrRV.... This sequence strongly resembles the bipartite motifs of the nuclear targeting signals of proteins that follow the model sequence (PKKKRKV) of the SV40 large T antigen [23].

The secondary structure prediction analysis for PL-III, based on the primary structure of PL-III, varies enormously depending on the prediction method used. The protein seems to consist mainly of turns and α helices ~10% (Chou and Fasman [12,13]) or ~40% (Garnier and Robson [14]). The latter value is closer to the values (26–52%) experimentally determined for this protein in the presence of helicogenic solvents [18]. None of the prediction methods used indicate any presence of β-pleated sheet conformation.

Under the buffer conditions used in the sedimentation velocity experiments (150 mM NaCl, 10 mM Tris-HCl pH 7.5), the protein adopts a mainly random coil configuration [16]. The radius of gyration (34.3 ± 1.4 Å) determined from the sedimentation coefficient is in good agreement with the radius of gyration R_G = 34.5 Å previously determined from viscosity measurements under the assumption of a random coil conformation [16]. These values compare very well with the value of 34 Å established for the radius of gyration of the nonhistone protein HMG-17 which consists of 89 amino acids and is also devoid of organized folding [24]. These results do not preclude the possibility that, upon binding to DNA, PL-III might adopt a more rigid structure. This could be the result of secondary structure organization of the molecule induced by the neutralization of the positive charges of the lysine and arginine side chains by the DNA phosphates [18]. However, as discussed

above, the exact extent of secondary structure involved remains yet to be established.

Acknowledgements: We would like to thank Nieves Forcada for her skillful assistance in preparing some of the Figures. We are also very grateful to Maree Roome for her careful typing of the manuscript. This work was supported by NSERC Grant OGP0046399.

References

[1] Bloch, D.P. (1969) Genetics (Suppl.) 61, 93–110.
[2] Subirana, J., Cozcolluela, C., Palau, J. and Unzeta, M. (1973) Biochim. Biophys. Acta 317, 364–379.
[3] Johns, E.W. (1964) Biochem. J. 92, 55–59.
[4] Ausio, J. (1986) Comp. Biochem. Physiol. 85B, 439–449.
[5] Carlos, S., Jutglar, L, Borrell, I., Hunt, D.R. and Ausio, J. (1993) J. Biol. Chem. 268, 185–194.
[6] Carlos, S., Hunt, D.F., Rocchini, C., Arnott, D.P. and Ausio, J. (1993) J. Biol. Chem. 268, 195–199.
[7] Ausio, J. and Subirana, J.A. (1982a) Exp. Cell. Res. 141, 39–45.
[8] Ausio, J., van der Goot, G.F. and Buckley, J.T. (1993) FEBS Lett. 333, 296–300.
[9] Van Holde, K. and Weischet, W.O. (1978) Biopolymers 17, 1387–1403.
[10] Perkins, S.J. (1986) Eur. J. Biochem. 157, 169–180.
[11] Tanford, C. (1961) Physical Chemistry of Macromolecules, John Wiley and Sons, NY, 452 pp.
[12] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211–222.
[13] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 223–245.
[14] arnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[15] Ruiz-Lara, S., Prats, E., Casas, M.T. and Cornudella, L. (1993) Nucleic Acids Res. 21, 2774.
[16] Ausio, J. and Subirana, J.A. (1982b) Biochemistry 21, 5910–5918.
[17] Mogensen, C., Carlos, S. and Ausio, J. (1991) FEBS Lett. 282, 273–276.
[18] Verdaguer, N., Perell–, M., Palau, J. and Subirana, J.A. (1993) Eur. J. Biochem. 214, 879–887.
[19] Ruiz-Lara, S. (1993) Ph.D. Thesis. Universitat Politécnica de Catalunya. Barcelona, Spain.
[20] Giancotti, V., Buratti, E., Santucci, A., Neri, P. and Crane-Robinson, C. (1992) Biochim. Biophys. Acta 1119, 296–302.
[21] Kremling, H., Reinhart, N., Schlösser, M. and Engel, W. (1992) Biochim. Biophys. Acta 1132, 133–139.
[22] Li, H. and Bingham, P.M. (1991) Cell 67, 335–342.
[23] Dingwall, C. and Laskey, R.A. (1991) Trends Biochem. Sci. 16, 478–481.
[24] Abercrombie, B.D., Kneale, G.G., Crane-Robinson, C., Bradbury, E.M., Goodwin, G.H., Walker, J.M. and Johns, E.W. (1978) Eur. J. Biochem. 84, 173–177.
[25] Kuntz, I.D. Jr. and Kauzmann, W. (1974) Adv. Protein Chem. 28, 239–345.