PL-I of *Spisula solidissima*, a Highly Elongated Sperm-Specific Histone H1[†]

John D. Lewis,[‡] Reginald McParland,[§] and Juan Ausió*,[‡]

Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada, V8W 3P6, and Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331

Received November 15, 2003; Revised Manuscript Received March 30, 2004

ABSTRACT: The major chromosomal protein of the mature sperm of the surf clam, *Spisula solidissima*, is a histone H1-related protamine-like (PL-I) protein of low electrophoretic mobility. We report here the complete sequence of two isoforms of its encoding genes. These genes encode a protein of 453 and 454 amino acids, respectively. The predicted mass of the larger isoforms (51 437 Da) was confirmed using electrospray ionization mass spectrometry. The amino-terminal tail of the *S. solidissima* PL-I is greatly elongated because of the presence of 39 tandem hexapeptide repeats of the motif (K/R)KRSAS with a few semiconservative amino acid substitutions. These repeats are very closely mirrored by their encoding DNA sequence, which indicates that an expansion because of sequence duplication most likely occurred. The C-terminal domain consists of a histone H1-related core with a predicted winged-helix tertiary structure, which is followed by an unstructured lysine-rich tail. This information provides additional molecular support for the classification and underlying evolution of sperm nuclear basic proteins in bivalve molluscs.

During the final stages of spermatogenesis, the DNA of sperm in most organisms is compacted, as germinal histones are replaced with sperm nuclear basic proteins (SNBPs).[1] The SNBPs can be grouped into three major types: histone (H type), protamine-like (PL type), and protamines (P type) (*1*). Protamine-like (PL) proteins are a structurally heterogeneous group of chromosomal proteins with a rather homogeneous amino acid composition intermediate to that of histones and protamines (*1*). They are referred to as PL because, like protamines, they replace a significant amount of the germinal histones during spermiogenesis.

The PL-I sperm nuclear protein from *Spisula solidissima* was first isolated in 1982. Analysis of its amino acid composition at that time showed that it contained similarly high amounts of lysine and arginine (24.8 and 23.1%, respectively) (*2*). Structural analysis of PL-I revealed a tripartite structure, consisting of N- and C-terminal "tails" flanking a globular, trypsin-resistant core of 75 amino acids (*3*). The trypsin-resistant globular core of *Spisula* PL-I was fully sequenced and found to be very similar to the globular "winged-helix" motif, a defining characteristic of the highly heterogeneous family of histone H1 (*4*, *5*). The winged-helix motif is found in a variety of proteins, including histone H5 (*6*, *7*) and a number of transcription factors such as HNF3 (*8*) and the FOXn1 family of oncogenes (*9*). A significant sequence similarity was found with the globular core of the chromatin-condensing protein histone H5, which condenses the chromatin of terminally differentiated chicken erythrocytes. The discovery of a histone H1-like core in the PL-I of *S. solidissima* (*3*) established a connection between the histone H1 family of chromosomal proteins and PL proteins (*1*, *10*).

The evolution of SNBPs has been the subject of much debate in recent years. From the time that these nuclear proteins were first characterized, it was suggested that protamines of germ cells and somatic histones were evolutionarily related. On the basis of compositional amino acid analysis, it has been hypothesized that protamines had evolved from a primitive somatic-like histone precursor via a PL intermediate by a mechanism of vertical evolution (*10*, *11*). Anecdotal evidence of this link has been provided by the identification of histone H1-like sperm nuclear proteins in a diverse range of organisms, including marine invertebrates, amphibians (*12*, *13*), and fish (*14*, *15*). Recently, an evolutionary mechanism has been identified for the direct conversion of a sperm-specific histone H1 to an arginine-rich protamine via frameshift mutations in the carboxyl-terminal tail in primitive chordates (*16*). The H1-like SNBPs of the bivalve molluscs, however, are distinctive in that they typically show the expansion of the amino-terminal tail of the sperm-specific H1 (*17*), similar to what is observed in certain fish (*18*). In many cases, these elongated proteins undergo posttranslational cleavage to produce shorter mature sperm proteins (*17*, *19*). The sperm of *Mytilus californianus* express an additional SNBP (PL-III) (*20*, *21*), which resembles the H1-related *M. californianus* sperm proteins (PL-II and PL-IV) but appears to have undergone genetic divergence (Lewis and Ausió, unpublished). The *S. solidissima* PL-I is unique in the regard that it does not undergo posttranslational cleavage to produce smaller SNBPs and may

[†] This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Grant OGP 0046399.

* To whom correspondence should be addressed: Department of Biochemistry and Microbiology, University of Victoria, Petch Building, Room 220, Victoria, B.C., Canada V8N 5Y2. Tel: 250-721-8863. Fax: 250-721-8855. E-mail: jausio@uvic.ca.

[‡] University of Victoria.

[§] Oregon State University.

[1] Abbreviations: SNBP, sperm nuclear basic protein; PL, protamine-like; ESIMS/MS, electrospray ionization quadrupole time-of-flight mass spectrometry; NBS, *N*-bromosuccinimide; UTR, untranslated region; SDS, sodium dodecyl sulphate; CNBr, cyanogen bromide; EL, elastase; CHY, chymotrypsin.

10.1021/bi0360455 CCC: $27.50
Published on Web 05/25/2004

therefore provide insight into the intermediate evolutionary step between histone H1 and the smaller PL proteins of bivalve molluscs.

Here, we set out to characterize the PL-I SNBP of *S. solidissima* and its encoding gene with the hopes of gaining insight into the mechanism of histone H1-like PL expansion and evolution. We also identify for the first time some of the putative regulatory regions of a SNBP gene from bivalve molluscs. Analysis of the coding region, as well as the promoter and 3′ UTR regions are discussed herein with these ideas in mind.

MATERIALS AND METHODS

Living Organisms. Specimens of *S. solidissima* were obtained from the Department of Marine Resources at the Marine Biological Laboratory (Woods Hole, MA). Specimens from *Ostrea gigas*, *M. californianus*, and *Macoma nasuta* were collected from various locations around Vancouver Island (B.C., Canada). *Ensis ensis* was obtained from commercial suppliers in Barcelona (Spain).

Preparation and Isolation of the SNBPs. Sperm nuclear basic proteins were routinely extracted with 0.4 N HCl following the procedures described previously (*22*). Buffers used during the isolation of proteins contained complete protease inhibitor cocktail tablets (Boehringer). The dried pellets obtained by acetone precipitation were stored at −80 °C.

Protein and Peptide Fractionation. Reverse-phase HPLC was performed on a 5-mm Vydac C_{18} column (25 × 3 × 0.46 cm) with 0.1% trifluoroacetic acid as the eluant with varying acetonitrile gradients (*4*). After fractionation, aliquots of each eluted peak were dried and resuspended in 5 μL of sterile water and analyzed on urea/acetic acid polyacrylamide gels.

Proteolytic Digestions: Cleavage at Aspartic Acid Using 2% Formic Acid. PL-I at 2 mg/mL in 2% formic acid was hydrolyzed for 2 h at 110 °C under vacuum and neutralized by the addition of NaOH.

Cleavage at Tyrosine/Tryptophan Using N-bromosuccinimide (NBS). PL-I at 10 mg/mL in 5% acetic acid and 8 M urea was made in 2 mg/mL NBS by addition of a 100-fold concentrated NBS solution in the same buffer, and the reaction was allowed to proceed for 15 min at room temperature. The reaction was stopped by the addition of free tyrosine in a 20-fold molar excess over NBS.

Cleavage at Methionine Using Cyanogen Bromide (CNBr). PL-I at 10 mg/mL in 100 mM HCl was hydrolyzed with CNBr using a 25-fold molar excess over the estimated (from amino acid analysis and molecular weight) molar amount of methionine.

Pepsin Digestion. PL-I at 10 mg/mL in 5% acetic acid and 8 M urea was digested at room temperature with pepsin (EC 3.4.23.1) (Sigma) [E/S 1:50 (w/w) for 30 min]. The mixture was finally diluted with 10 volumes of water and quickly loaded onto a reverse-phase HPLC column.

Elastase Digestion. PL-I at a concentration of 2 mg/mL in 100 mM ammonium bicarbonate (pH 8.00) was digested at room temperature with elastase (EC 3.4.21.36) (Worthington) [E/S 1:100 (w/w) for 30 min].

Chymotrypsin Digestion. PL-I at 10 mg/mL in 50 mM Tris-HCl (pH 8.0) was digested at room temperature with chymotrypsin (EC 3.4.21.1) (Sigma) [E/S 1:500 (w/w) for 90 min]. The sample was then brought to 3% acetic acid and loaded directly onto a C_8 Vydac column.

Protein Microsequencing. Protein and peptide sequencing were carried out on an Applied Biosystems model 470A gas-phase protein sequencer as described elsewhere (*3*, *23*).

Protein Gel Electrophoresis. Acetic acid (5%)/urea (2.5 M) polyacrylamide gels were prepared as described in (*24*). Like protamines, PL proteins exhibit very low or no solubility on SDS, and therefore SDS−PAGE cannot be used for the analysis of the proteins.

Mass Spectrometry. The HPLC-purified sample was desalted over a POROS R2 resin and eluted with 5 μL of 60% methanol/3% formic acid that was coupled directly with an ESI quadrupole time-of-flight instrument (QStar Pulsar I). Data were deconvoluted for proteins in the 20 000−70 000 molecular weight range. Results obtained represent the uncharged average mass.

Hybridization Probe Preparation. Because of the highly repetitive nature of the region 5′ to the globular region of the PL-I gene, it was necessary to create a probe of 306 bp corresponding to the globular region and extending toward the 3′ end of the gene. This was accomplished by PCR amplification of the longer genomic clone using the primers SPISF3 (5′-ATGATGAGCATGGTCGCTGCAGCCATTG-3′) and SPISR2 (5′-CATCGTCTTCTTTGTCTTCTTTGTG-GTC-3′).

Southern Blot. A horizontal 0.7% agarose gel was run as described above with each lane containing 10 μg of genomic DNA digested overnight with *Spe* I, *Eco*R I, or *Xho* I. The marker was a λDNA-*Bst*E II marker (New England Biolabs). The gel was blotted using the VacuGene XL Vacuum Blotting System (Pharmacia Biotech) following the instructions of the manufacturer. The blotting membrane used was Zeta-Probe GT (BioRad). The gel was first depurinated for 20 min in 0.2 N HCl, denatured for 20 min in 0.5 M NaOH and 1.5 M NaCl, neutralized for 20 min in 1 M Tris-HCl at pH 7.5 and 1.5 M NaCl, and then transferred for 1 h to a membrane with 20× SSC. The blot was then washed for 5 min in 20× SSC to remove any agarose, air-dried for 30 min, and vacuum-dried for 30 min at 80 °C. The double-stranded 306-bp insert was labeled by nick translation according to ref *25*. The labeled probe was purified from the free label using a microcon 10 (Amicon) following the protocol of the manufacturer. The hybridization was performed in a Hybaid Hybridization Oven (Interscience). The blot was first prehybridized in 10 mL of 0.5 M Na_2HPO_4 (pH 7.2) and 7% SDS at 65 °C for 30 min. The probe was heat-denatured at 95 °C for 5 min and added to the hybridization solution, 5 mL of 0.5 M Na_2HPO_4 (pH 7.2) and 7% SDS. Hybridization was carried out for 18 h at 65 °C.

The blot was washed twice for 45 min in 40 mM Na_2HPO_4 (pH 7.2) and 5% SDS and twice for 30 min in 40 mM Na_2HPO_4 (pH 7.2) and 1% SDS. It was then wrapped in Saran wrap and exposed and visualized using the PhosphorImager System (Molecular Dynamics) at room temperature for 24 h.

Degenerate PCR. Degenerate primers for PCR were created based on the determined amino acid sequence of the globular core of PL-I (*3*). PCR was performed using the

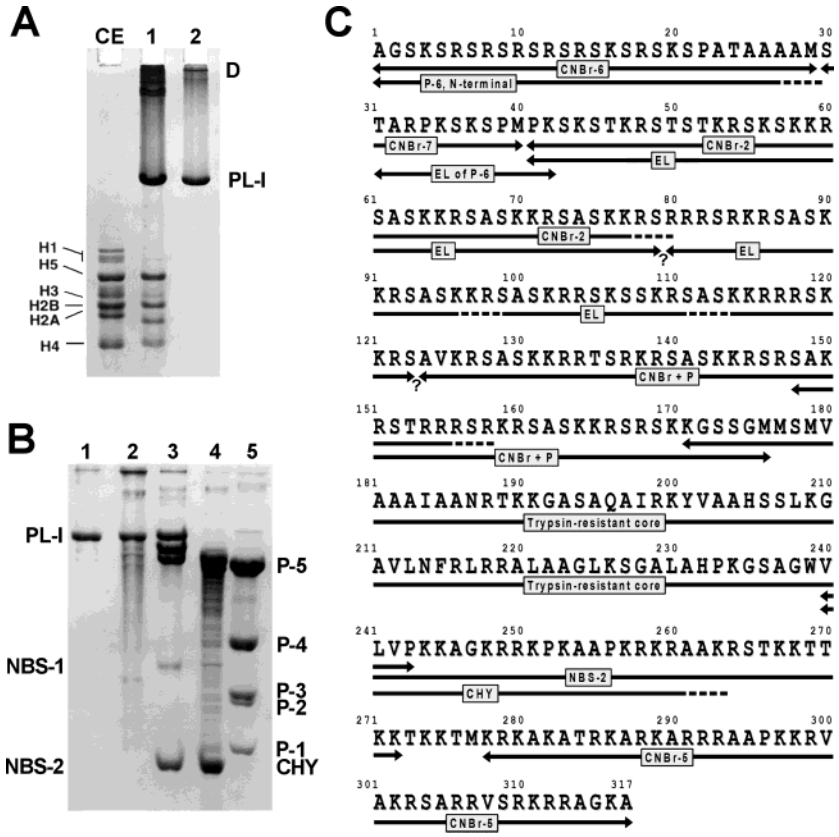

FIGURE 1: (A) Lane 1, acetic acid (5%)/urea (2.5 M) polyacrylamide (15%) gel electrophoresis (AU−PAGE) of crude extract from the sperm of *S. solidissima*; lane 2, RP-HPLC-purified PL-I. CE (chicken erythrocyte) histones were used as a marker. The letter D indicates the intermolecular dimer. (B) AU−PAGE analysis of a few of the proteolytic digestions of PL-I used in the generation of peptides used in the sequencing of the protein. Lane 1, PL-I; Lane 2, digestion of PL-I with 2% formic acid; lane 3, NBS digestion; lane 4, chymotrypsin digestion; lane 5, pepsin digestion. (C) Primary structure of PL-I determined from the overlap of different peptides. The peptides are shown by arrows with the name of the peptide: CHY, chymotrypsin peptide; CNBr, cyanogens bromide peptide; EL, elastase peptides; NBS, *N*-bromosuccinimide peptides; and P, pepsin peptides. The question marks denote the regions for which no overlaps could be determined. Dashed lines refer to areas of lower sequence confidence.

PCRSprint thermal cycler (Interscience) with genomic DNA as the template. A touchdown profile was used for the amplification, with the annealing temperature decreasing from 65 to 45 °C over 20 cycles, followed by 10 cycles at 45 °C.

Genomic Walking. Genomic walking was performed on genomic DNA using adaptors, adaptor primers, and protocols based on ref *26*. DNA was digested overnight with *Spe* I, *Nhe* I, and *Xba* I (New England Biolabs). Adaptors were ligated at 16 °C for 6 h, and PCR reactions were carried out using the adaptor-specific PCR primer PP1 and the gene-specific primers SPISGEN-F1 (5′-GGCTCAGTAGGTTGGGTTCTTGTACC-3′) and SPISGEN-R1 (5′-CATACTTGCGGATAGCTTGGGCTGAAGCACC-3′). A 1:40 dilution was made of the products of the first reaction, and 1 μL of this was added to a second PCR reaction using the nested adaptor-specific PCR primer PP2 and the gene-specific primers SPISGEN-F2 (5′-GGGCAGCAAAGAGGTCCACAAAGAAGACCAC-3′) and SPISGEN-R2 (5′-CAATGGCTGCAGCGACCATGCTCATCAT-3′). Stratagene's Herculase Enhanced DNA polymerase and buffer system were utilized for the PCR reactions. A hot-start and touchdown profile was used for each amplification, exactly as in ref *26*.

Cloning and DNA Sequencing. PCR products were purified using Wizard PCR Preps DNA Purification System (Promega). The purified PCR products were then cloned into pCR 2.1-TOPO vector (Invitrogen) following the instructions of the manufacturer and transformed into TOP10 competent cells (Invitrogen).

RESULTS AND DISCUSSION

Previous Attempts To Determine the Sequence and the Size of PL-I. Having obtained the primary structure of the trypsin-resistant globular core of PL-I of *S. solidissima* (*3*), we attempted to sequence the entire molecule. A significant protein sequencing effort was undertaken, wherein a large amount of PL-I was purified (see Figure 1A) and then digested with a variety of proteases (Figure 1B). The sequencing of over 30 distinct peptides (over 900 amino acids in total) provided an incomplete sequence (see Figure 1C). The complete sequence remained elusive because of the difficulty in establishing proper peptide overlaps in the N-terminal region, which was compounded by the presence of highly repetitive sequences. Furthermore, at this point the molecular weight of the protein and hence the number of its constituent amino acids was not precisely known.

The apparent molecular weight of PL-I of *S. solidissima* had been determined previously to be 33 500 Da using a sedimentation equilibrium (*2*). We utilized MS to obtain a more precise mass of the *S. solidissima* PL-I protein. Analysis of PL proteins by MS is problematic in general because of a high content of charged residues. Our initial results obtained from ESIMS (electrospray ionization mass spectrometry)

instrumentation produced raw data that were particularly complex. Regression probability analysis was performed using MassLynx 4.0 (Waters, Inc.), revealing that the abundant species was 51 454 ± 124 Da (data not shown). Because it appeared at this point that multiple protein species were present, we further refined our molecular weight determination of PL-I by utilizing a more advanced ESIMS/MS (electrospray quadrupole time-of-flight mass spectrometry) instrument. This method resulted in a clear and distinct peak corresponding to a molecular mass of 51 431 ± 28 Da (Figure 1C). Considering our previous experimental estimates, this result was quite unexpected and it underscores the importance of the charge effect on the sedimentation equilibrium analysis of these highly cationic proteins (*27*).

PL-I Gene Encodes a Highly Elongated H1-like SNBP. As described above, determination of the complete PL-I sequence by conventional protein sequencing proved to be quite difficult and time-consuming. We therefore decided to refocus our efforts on the isolation and characterization of the gene(s) encoding this protein. The complete gene sequence for the *S. solidissima* PL-I protein was obtained in a two-step process. Degenerate PCR was used to amplify a 306-bp genomic clone. This information was utilized to create nondegenerate PCR primers that were used for genomic walking in both the 5′ and 3′ directions. Two distinct isoforms of the PL-I coding region were identified in this manner, hereafter referred to as SsPLIa and SsPLIb. Both genomic clones contain a single open-reading frame encoding proteins of 453 and 454 amino acids, respectively (Figure 3), flanked by 442 bp of the upstream sequence and 45 bp of the 3′ UTR. A comparison of the predicted proteins to our previous protein sequencing data is seen in Figure 3B. For the most part, the protein- and DNA-derived sequences are in good agreement. There are a number of single amino acid substitutions over the length of the PL-I protein, though most are conversions between lysine, arginine, serine, and alanine, which together comprise 84% of the entire sequence. These substitutions are not unexpected because of heterogeneity that results from the rapid evolution of the PL proteins. The two isoforms may be separate genomic copies (alleles) of the PL-I gene (see below for a discussion of the copy number), though the MS data suggests that a single protein isoform (SsPLIb) is expressed in the mature sperm. The sequencing data also correctly identified the single cysteine residue that was misidentified during protein microsequencing but has been identified and characterized extensively in previous work (*28*).

The calculated average mass for the entire encoded PL-I protein is 51 139 Da for SsPLIa and 51 437 for SsPLIb. The predicted mass for PL-I SsPLIb was in close agreement with the mass obtained by MS. Because our MS data showed no peak in the region of 51 139 Da, it may be that the gene sequence for SsPLIa is not expressed. Protamines and sperm-specific histone H1s are known to be highly phosphorylated at the outset of spermiogenesis and become rapidly dephosphorylated as the histone replacement proceeds and the chromatin is compacted (*29*, *30*). While we initially surmised that the disparity in molecular mass data may have been due to posttranslational modifications, our subsequent use of more advanced instrumentation allowed us to clearly determine the presence of a single homogeneous species at 51 431 ± 10 Da.

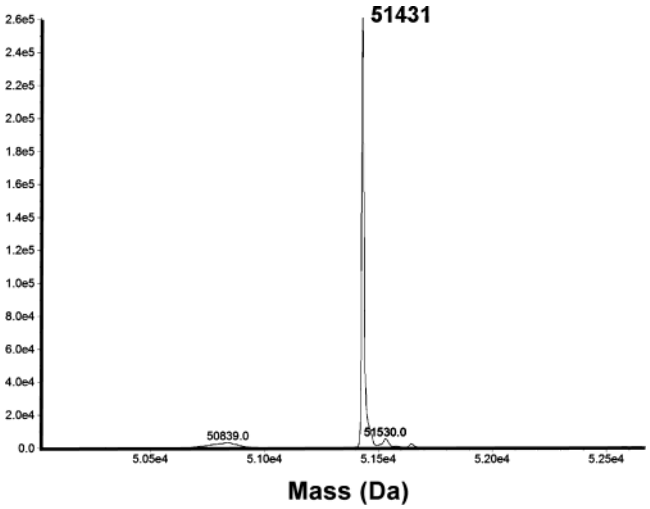

FIGURE 2: (A) ESIMS/MS of HPLC-purified *S. solidissima* PL-I protein. Results obtained represent the uncharged average mass. The mass of PL-I (51 431 ± 28 Da) is highlighted.

PL-I Protein Contains Many Repetitive Motifs. An important potential target for phosphorylation is a domain containing five RS repeats found at the N terminus of the PL-I of *S. solidissima* (green block of Figure 3A). These motifs have been described in a number of the PL proteins, such as the PL-II of *M. californianus* and the EM-1 of *E. minor*, as well as in protamines, such mammalian protamine 1 (*31*) and the alligator Al-I and Al-II protamines (*32*). Protamine 1 is phosphorylated in testis by the SR protein-specific kinase 1 (SRPK1) (*29*), a kinase that also targets the SR motifs of splicing factors (*33*). Because SR proteins are also specifically phosphorylated by topoisomerase I (*34*) and the chromatin remodeling events during spermiogenesis require the activity of topoisomerase I to maintain sufficient DNA relaxation (*35*), it is quite plausible that topoisomerase I may also be involved in the phosphorylation of SNBPs.

The amino-terminal tail of *S. solidissima* PL-I, at 306 amino acids in length, is significantly elongated in comparison to that of other members of the histone H1 family. For example, the chromatin-condensing histone H5 has an N-terminal tail of only 25 amino acids, while the N-terminal tail of the H1-like PL-II protein from the sperm of the bivalve *M. californianus* is 42 amino acids long. The N-terminal tail of *S. solidissima* is comprised mainly of highly basic but simple hexapeptide repeats, at least 39 in total (Figure 2). A total of 19 of these are (K/R)KRSAS (red blocks of Figure 3A), and a further 15 repeats follow this template with a single amino acid substitution (dark orange blocks of Figure 3A). The remaining repeats contain two or three semiconservative amino acid substitutions (dark and light yellow blocks of Figure 3A). These repeating motifs may be involved in the formation of specific structures when interacting with DNA. During spermiogenesis, it is likely that these motifs comprise units whose electrostatic interaction with DNA is modulated by the phosphorylation and dephosphorylation of the serine residues within each unit. The EM-1 PL protein from the sperm of *E. minor* has an N-terminal tail that bears the closest similarity to *S. solidissima* PL-I, with a length of 212 amino acids. It also possesses hexapeptide repeats, as does *E. minor* EM-6, which is an N-terminal cleavage product of a PL precursor (*17*). Comparison of the hexapeptide repeats from both *S. solidissima*

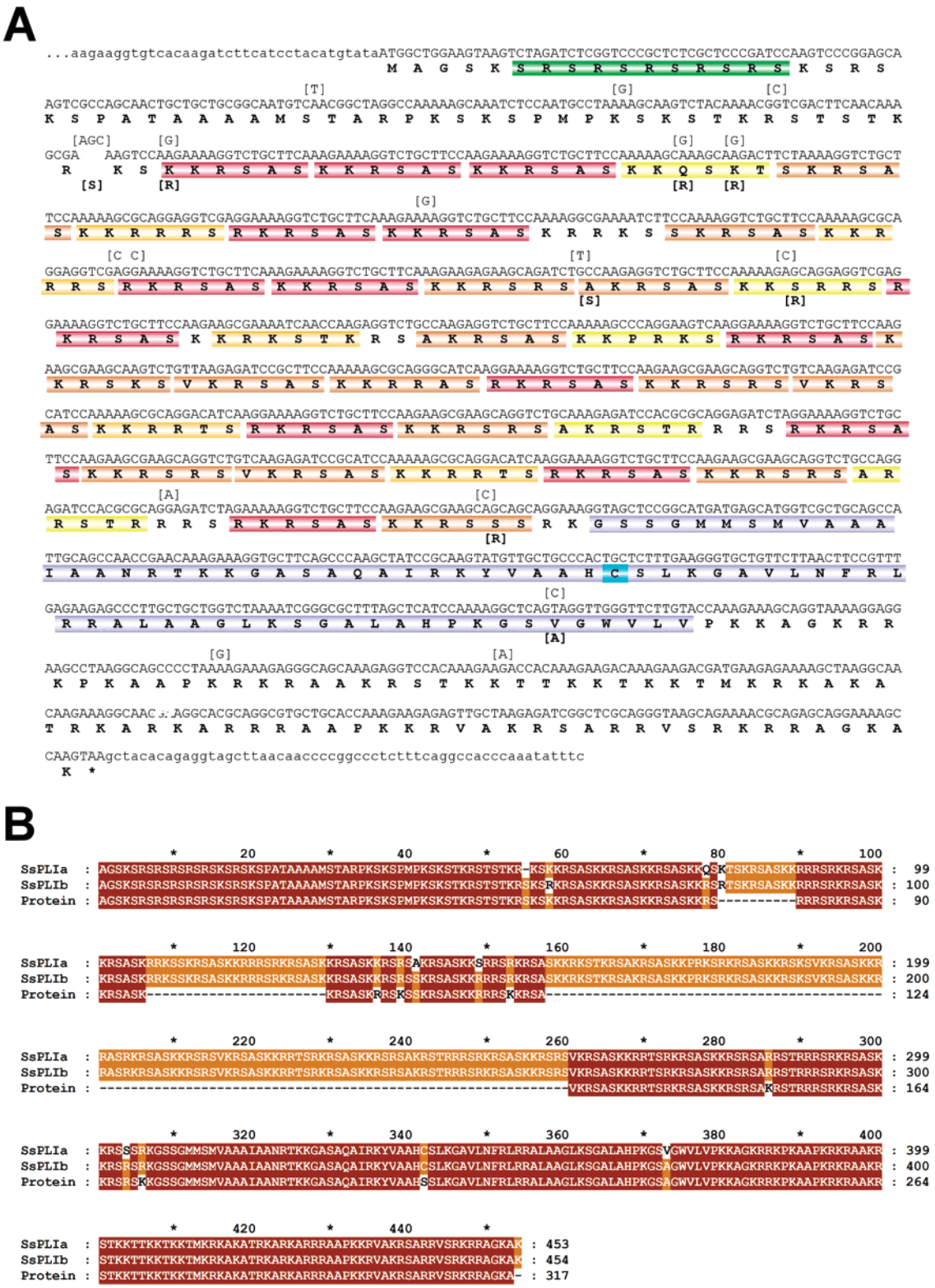

FIGURE 3: (A) Complete gene sequences for two isoforms of the PL-I of *S. solidissima* (AY626224 and AY626225). The coding region is in capital letters, and the flanking regions are in lowercase letters. The SsPLIa isoform is listed, and the SsPLIb DNA sequence and protein sequence differences are displayed in brackets. The green block indicates N-terminal SR repeat motif. Red blocks denote conserved hexapeptide repeats (K/R)KRSAS. Dark orange blocks indicate hexapeptide repeats with a single amino acid substitution. Dark and light yellow blocks indicate hexapeptide repeats with two and three substitutions, respectively. The histone H1-related winged-helix globular region is highlighted with the blue block. The single cysteine residue is indicated by the cyan box. (B) Multiple alignment of *Spisula* PL-I predicted protein sequences from genomic isoforms SsPLIa and SsPLIb with partial PL-I protein sequence determined by amino acid microsequencing. Complete identity denoted with brown blocks and pairwise identity, in orange. The sequences were aligned using the CLUSTAL X multiple sequence alignment program (*54*).

and *E. minor* PL proteins is seen in Figure 4C, which indicates a high conservation of a motif consisting of three basic residues followed by two serines separated by a single amino acid. The carboxy-terminal tail of the *S. solidissima*

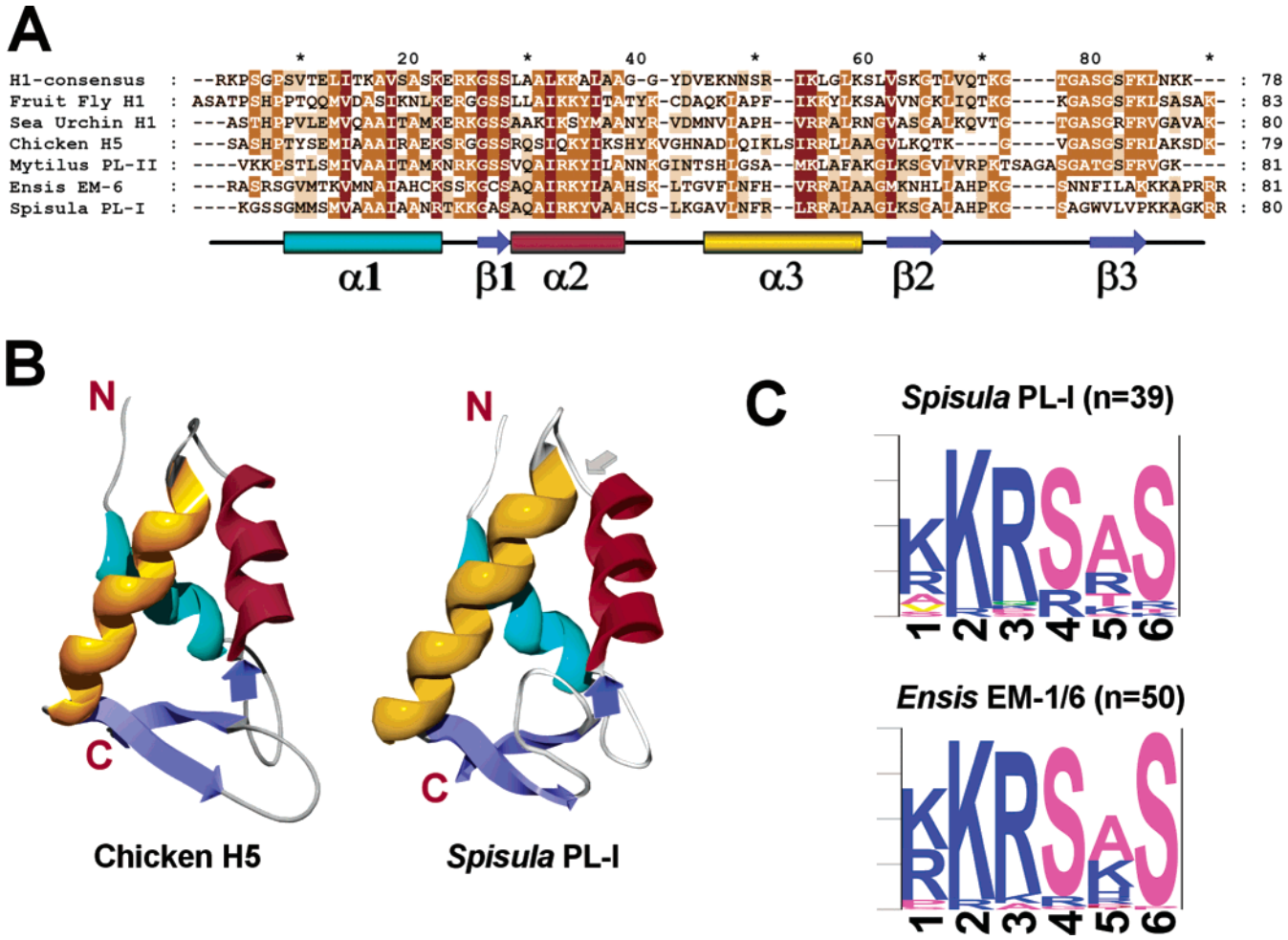

FIGURE 4: (A) Multiple alignment of *S. solidissima* PL-I winged-helix region with the winged helices of other histone H1 and H1-related SNBPs, with the secondary structure highlighted below. The sequences were aligned using the CLUSTAL X multiple sequence alignment program (*54*). The sequence accession numbers, if available, are H1 consensus (*55*), H1 fruit fly (P02255), H1 urchin (P15869), H5 chicken (P02259), PL-II *M. californianus* (A45317), and EM-1 *E. minor* (AAA98076). (B) Left, 3D rendering of the globular core of chicken erythrocyte H5 with coordinates as determined by ref *38*; right, theoretical 3D rendering of the globular core of *S. solidissima* PL-I, generated with the aid of the SWISS-MODEL server (*39*). Secondary structures are color-coded to match the secondary structure in A. The gray arrow indicates the position of cysteine in PL-I. (C) N-terminal hexapeptide repeats of *S. solidissima* PL-I and *E. minor* EM1/6 displayed in a Logos format (*56*). In this representation, the size of the letters is proportional to the frequency with which an amino acid appears at a given position in the sequence, and the overall height of all of the letters in that position is proportional to the conservation of the site. The letters are color-coded according to the physical and chemical structural characteristics of the amino acids that they represent.

PL-I protein does not contain the hexapeptide motifs seen in the N-terminal tail, but it is very lysine- and arginine-rich and also has a significant serine and threonine content.

S. solidissima PL-I Contains a Conserved Winged-Helix Motif. The sequence of the globular core of PL-I puts it unmistakably in the histone H1/H5 family of linker histone proteins (Figure 4A). The structure databases PFAM (protein families database) (*36*) and SMART (simple modular architecture research tool) (*37*), when queried with the PL-I protein, report very significant hits for the linker histone H1 and H5 family of proteins based on the identification of the winged-helix motif. The PL-I globular region exhibits a 41 and 50% similarity to the winged-helix regions of chicken histone H5 and sea urchin (*Strongylocentrotus purpuratus*) histone H1, respectively. The winged-helix region of the closely related bivalve *E. minor* EM-1 is most similar at 69%.

Figure 4B shows the results obtained by using the crystal structure coordinates of the histone H5 globular core (*38*) to extrapolate the three-dimensional structure of PL-I, with tools made available at the SWISS-MODEL server (*39*). Two principle differences between the H5 and PL-I core structures can be seen in this figure. First, the α3 helix of PL-I appears to be extended by roughly a half-turn. In addition, the highly conserved β-sheet structure close to the carboxyl-terminal end of the winged helix appears more twisted and coiled in toward the center of the structure. This is not particularly surprising, considering the presence of a tryptophan residue in the PL-I, which is not present in histone H5. The gray arrow in Figure 4B indicates the position of the cysteine residue in the structure of the winged helix. At this position, it would likely be on the exposed surface and accessible for intermolecular interactions (*28*).

Number of PL-I Gene Genomic Copies. Southern blot analysis was carried out to determine the copy number of the PL-I gene in *S. solidissima* (Figure 5A). The results indicate that the PL-I gene is present in two or a multiple of two copies. The data point strongly toward the existence of four copies, however, because the controls, which contained single copy amounts of PL-I DNA, displayed equivalent intensities to each of those in lanes *Eco*R I and *Xho* I in

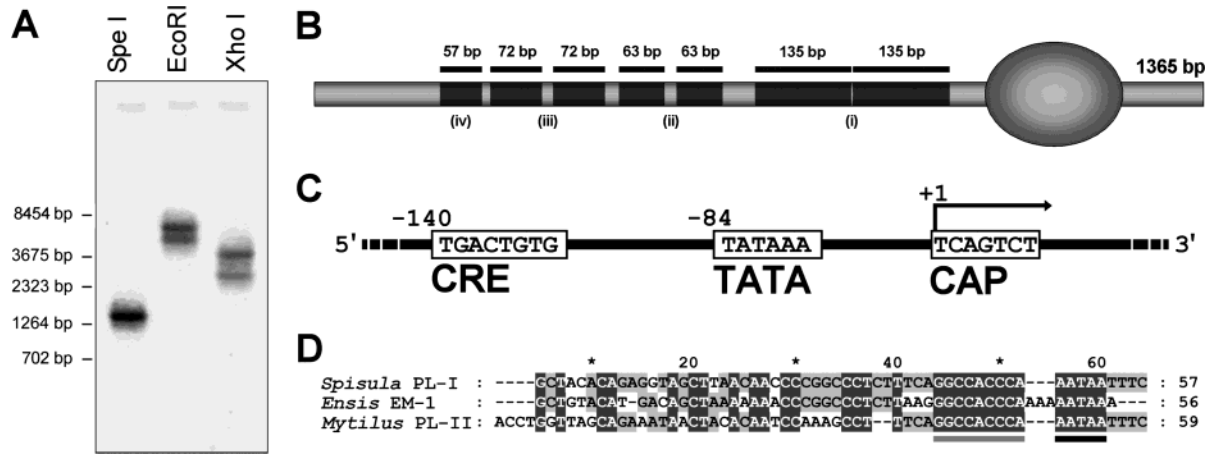

FIGURE 5: (A) Southern blot of *S. solidissima* genomic DNA hybridized with the 306-bp PL-I probe. Each lane contains 10 μg of DNA digested with *Spe* I, *Eco*R I, or *Xho* I as indicated. The λDNA-*Bst*E II marker (New England Biolabs) positions are indicated on the left-hand side. (B) Diagram depicting the extensive sequence duplication within the coding region of the PL-I gene. Identical repeats appear as similar colors. Duplications (i), (ii), and (iii) are direct tandem repeats, while (iv) is a partial repeat of (iii). (C) Putatively identified conserved features of the promoter sequence of *S. solidissima* PL-I. (D) Multiple alignment of 3′ UTR of PLs from the related bivalve molluscs *E. minor* (L41834) and *M. californianus*. The polyadenylation signal is indicated by the black bar. The light gray bar indicates an additional conserved sequence that may be the target of factors responsible for the repression of mRNA translation during spermiogenesis.

Figure 5A (data not shown). Previous studies have shown that there is a wide range of gene copy numbers for SNBPs. The genes for the human protamines P1 and P2 exist as a single linear array of PRM1 (protamine 1), PRM2 (protamine 2), and TNP2 (the gene for transition protein 2) on human chromosome 16 (*31*). The gene for the PL SNBP PL-III of *M. californianus*, on the other hand, has many copies, which are widely dispersed (*40*). The replacement of histones by protamines or PL proteins occurs in the nucleus very rapidly, requiring a large amount of SNBP to be translated in a very short period of time, which in turn relies on the production of significant amounts of mRNA. In some organisms, such as humans, this issue has been solved by building up large amounts of protamine mRNA and subjecting it to translational repression. This is accomplished by factors that specifically bind the 3′ UTR and "silence" the mRNA until it is required for translation during spermiogenesis (*41*). Other organisms, however, seem to use a "brute force" method and rely on high gene copy numbers to produce sufficient amounts of SNBP mRNA in the limited available time.

PL-I Protein Has Elongated through Genomic Duplication. The genomic sequence encoding the amino-terminal tail of *S. solidissima* PL-I provides some insight into the nature and origin of its peptide repeats. There are a number of recurring nucleotide sequences throughout the region coding for the N-terminal tail, arranged into distinct patterns of 18 nucleotides each that correspond with the repeating amino acid hexapeptides. The extremely elongated nature of the N-terminal tail of *S. solidissima* PL-I, however, seems to have been the result of four distinct sequence duplication events (Figure 5B). The largest of these sequence duplications is indicated by (i) in Figure 5B and consists of two tandem identical repeats of 135-bp each, comprising nucleotides 610−744 and 745−849 (numbering from the translational start site). The second duplication (ii) is exactly 63 bp, consisting of the nucleotides 421−484 and 501−564. The last events (iii and iv) depicted in Figure 5B are the result of two distinct overlapping sequence duplications, the first resulting in a 72-bp tandem repeat including the nucleotides 244−315 and 331−402 and the second duplicating a 57-bp region overlapping both of these regions, from nucleotides 283−340 to 174−232. The most obvious result of these duplication events is the extension of the PL-I gene by a minimum of 327 nucleotides, which translates to an elongation of the amino-terminal tail by 109 amino acids.

A similar rapid expansion of repetitive sequences is seen in both the SNBPs of winter flounder (*42*) and *E. minor* (*17*), though the mechanism for such evolutionarily rapid duplication events is not immediately apparent. It is possible that this occurs as a result of unequal crossing over or perhaps slippage during replication. It has been suggested that the presence of repeated protein motifs contained in chromatin-associated proteins (such as histone H1) may actively facilitate an elevated level of recombination in their own genes (*43*). A sperm chromatin-condensing protein in contact with its own coding sequence could theoretically exert significant selective pressure in the propagation of genetic modifications. Evidence for such a mechanism is lacking, however.

Identification of Putative Binding Sites in the UTR of the PL-I Gene. The promoter sequences of the *S. solidissima* PL-I gene were compared with those of histone H1s, histone H5, and protamines. While there is a very distinctive TATA box (Figure 5C), conserved elements such the H4 box of vertebrate and invertebrate H1s (and histone H5) were not detected. A putative CRE element and CAP site were identified, on the basis of the vertebrate consensus sequences (*44*). It will likely be necessary to characterize the genes of a number of other PL proteins so that an adequate comparison of the SNBP promoters can be undertaken. The protamine genes of vertebrates also contain conserved binding sites for the spermiogenesis-specific activating factors PAF-1 and Y-box-binding protein in the first 100 upstream nucleotides (*45*). While it is likely that expression of PL-I is regulated by similar factors, the binding sites have diverged significantly enough so as not to be recognizable.

The 3′ UTR of the *S. solidissima* PL-I gene was compared to those from the closely related bivalves *E. minor* (EM-1/6) and *M. californianus* (PL-II/IV), for which cDNA sequences are available (Figure 5D). All three sequences

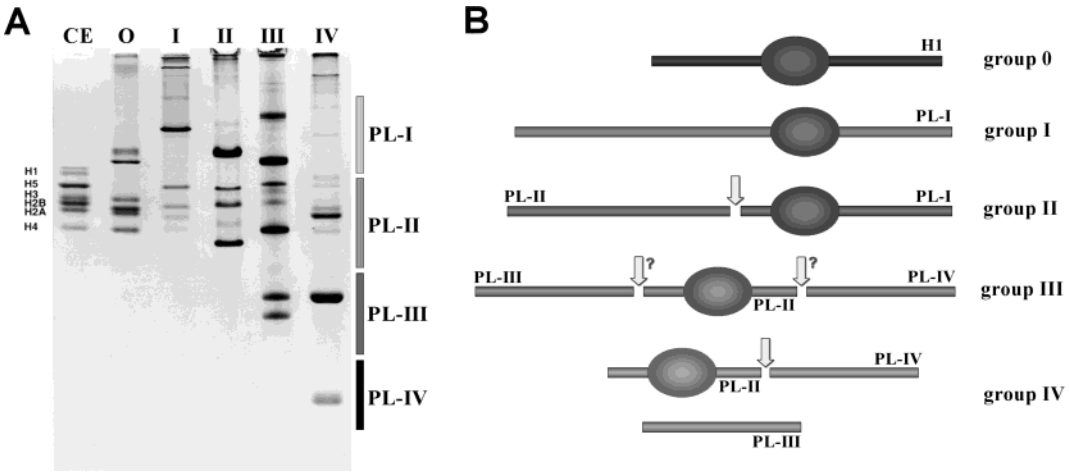

FIGURE 6: (A) AU−PAGE analysis of the SNBP from a representative organism of each of the SNBP groups (0−IV) of bivalve molluscs. 0, *Crassostrea gigas* (pacific oyster); I, *S. solidissima* (surf clam); II, *E. ensis* (razor clam); III, *Macoma nasuta* (bent-nosed clam); and IV, *M. californianus* (California mussel). (B) Schematic representation of the relationship between the different SNBP (H1 and PL) used for the classification of these groups (*5*, *51*).

possess a polyadenylation signal, which is modified from the consensus AATAAA to AATAAT in both *S. solidissima* and *M. californianus*. All protamine genes described to date contain a polyA signal and are polyadenylated. Polyadenylation plays a very important role in the temporal translational regulation of protamine genes (*46*, *47*). Variations on the consensus polyadenylation signal are thought to be a means of additional regulation (*48*). In addition, there is a well-conserved sequence of 9 nucleotides (GGCCACCCA) directly upstream from the polyadenylation signal (light gray bar of Figure 5D). This sequence motif is in the same location as the nucleotide sequences that show a significant extent of similarity among vertebrate protamine genes (*49*), and binding of sequence-specific RNA-binding proteins to these regions has been suggested to play an important translational regulatory role (*47*). Two recently identified proteins in mice, MSY2 and MSY4, bind to a 5′-UCCAUCA-3′ consensus sequence located in the 3′ UTR of the protamine P1 mRNA (*50*). MSY4 in particular has been shown to play an active role in the translational repression of several mRNAs in differentiating spermatids (*41*).

Evolution and Molecular Relationships of PL Proteins. Our findings support the notion that the PL proteins of bivalve molluscs have arisen from a common histone H1 ancestor. Our evidence strongly suggests that the unique protein sequence of *S. solidissima* PL-I is the result of an expansion of the amino-terminal tail of a sperm-specific H1, resulting in an extended and regularly repeating structure consisting primarily of arginine, lysine, alanine, and serine (23.8 + 23.3 + 14.1 + 22.7 = 83.9%). Sequences contained in the PL-I of *S. solidissima* bear close resemblance to both the H1-like and protamine-like SNBPs of *E. minor*, as well as the smaller and more protamine-like proteins PL-III and PL-IV of *M. californianus*.

Several years ago it was shown that bivalve molluscs could be arranged in five major categories (0−IV) based on the types of SNBPs (PL-I, II, III, and IV) expressed in the mature sperm (*5*, *51*) (see Figure 6). Despite their enormous structural variability, bivalve PL proteins exhibit close compositional similarity and can be related to the heterogeneous family of histone H1s (*5*, *51*). At the time, the molecular relationship between the different PL proteins within each group was not clear. Since then, several PLs from representative species belonging to several of these different groups of bivalves in this classification have been characterized: group 0, *O. edulis* (H1-1/H1-2/PL-IV) (*52*); group I, *S. solidissima* (PL-I) (this paper); group II, *E. minor* [PL-I (EM-1)/PL-II (EM-6)] (*17*); and group IV, *M. californianus* (PL-II/PL-IV) (*19*, *23*) and PL-III (*21*). This information makes it possible to establish the molecular relationship between the bivalve mollusc SNBPs (see Figure 6B). The H1-like PL proteins of sperm, which coexist with a fraction of germinal histones, may in fact form novel chromatin structures (*53*). Further genetic characterization of the SNBPs can capitalize on the significant amounts of genetic and protein association data that are being amassed today and should further elucidate the evolutionary origins of these rapidly evolving proteins.

REFERENCES

1. Ausió, J. (1995) in *Advances in Spermatozoal Phylogeny and Taxonomy* (Jamieson, B. G. M., Ausió, J., and Justine, J. L., Eds.) pp 447−462, Memoires du Museum National d'Histoire Naturelle, Paris, France.
2. Ausió, J., and Subirana, J. A. (1982) A high molecular weight nuclear basic protein from the bivalve mollusc *Spisula solidissima*, *J. Biol. Chem. 257*, 2802−2805.
3. Ausió, J., Toumadje, A., McParland, R., Becker, R. R., Johnson, W. C., and van Holde, K. E. (1987) Structural characterization of the trypsin-resistant core in the nuclear sperm-specific protein from *Spisula solidissima*, *Biochemistry 26*, 975−982.
4. Ausió, J. (1988) An unusual cysteine-containing histone H1-like protein and two protamine-like proteins are the major nuclear proteins of the sperm of the bivalve mollusc *Macoma nasuta*, *J. Biol. Chem. 263*, 10141−10150.
5. Ausió, J. (1992) Presence of a highly specific histone H1-like protein in the chromatin of the sperm of the bivalve mollusks, *Mol. Cell. Biochem. 115*, 163−172.
6. Garel, A., Mazen, A., Champagne, M., Sautiere, P., Kmiecik, D., Loy, O., and Biserte, G. (1975) Chicken erythrocyte histone H5; I. Amino terminal sequence (70 residues), *FEBS Lett. 50*, 195−199.
7. Sautiere, P., Kmiecik, D., Loy, O., Briand, G., Biserte, G., Garel, A., and Champagne, M. (1975) Chicken erythrocyte histone H5 II. Amino acid sequence adjacent to the phenylalanine residue, *FEBS Lett. 50*, 200−203.
8. Cirillo, L. A., McPherson, C. E., Bossard, P., Stevens, K., Cherian, S., Shim, E. Y., Clark, K. L., Burley, S. K., and Zaret, K. S. (1998)

Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome, *EMBO J. 17*, 244−254.
9. Schlake, T., Schorpp, M., and Boehm, T. (2000) Formation of regulator/target gene relationships during evolution, *Gene 256*, 29−34.
10. Ausió, J. (1999) Histone H1 and evolution of sperm nuclear basic proteins, *J. Biol. Chem. 274*, 31115−31118.
11. Subirana, J. A., Cozcolluela, C., Palau, J., and Unzeta, M. (1973) Protamines and other basic proteins from spermatozoa of molluscs, *Biochim. Biophys. Acta 317*, 364−379.
12. Kasinsky, H. E., Huang, S. Y., Mann, M., Roca, J., and Subirana, J. A. (1985) On the diversity of sperm histones in the vertebrates: IV. Cytochemical and amino acid analysis in Anura, *J. Exp. Zool. 234*, 33−46.
13. Itoh, T., Ausió, J., and Katagiri, C. (1997) Histone H1 variants as sperm-specific nuclear proteins of *Rana catesbeiana*, and their role in maintaining a unique condensed state of sperm chromatin, *Mol. Reprod. Dev. 47*, 181−190.
14. Saperas, N., Ausió, J., Lloris, D., and Chiva, M. (1994) On the evolution of protamines in bony fish: Alternatives to the "retroviral horizontal transmission" hypothesis, *J. Mol. Evol. 39*, 282−295.
15. Watson, C. E., and Davies, P. L. (1998) The high molecular weight chromatin proteins of winter flounder sperm are related to an extreme histone H1 variant, *J. Biol. Chem. 273*, 6157−6162.
16. Lewis, J. D., Saperas, N., Song, Y., Zamora, M. J., Chiva, M., and Ausió, J. (2004) Histone H1 and the origin of protamines, *Proc. Natl. Acad. Sci. U.S.A.* in press.
17. Bandiera, A., Patel, U. A., Manfioletti, G., Rustighi, A., Giancotti, V., and Crane-Robinson, C. (1995) A precursor-product relationship in molluscan sperm proteins from *Ensis minor*, *Eur. J. Biochem. 233*, 744−749.
18. Watson, C. E., Gauthier, S. Y., and Davies, P. L. (1999) Structure and expression of the highly repetitive histone H1-related sperm chromatin proteins from winter flounder, *Eur. J. Biochem. 262*, 258−267.
19. Carlos, S., Hunt, D. F., Rocchini, C., Arnott, D. P., and Ausió, J. (1993) Post-translational cleavage of a histone H1-like protein in the sperm of *Mytilus*, *J. Biol. Chem. 268*, 195−199.
20. Zalensky, A., and Zalenskaya, I. A. (1980) Basic chromosomal proteins of marine invertebrates. III. The proteins from the sperm of bivalve molluscs, *Comp. Biochem. Physiol., B 66*, 415−419.
21. Rocchini, C., Rice, P., and Ausió, J. (1995) Complete sequence and characterization of the major sperm nuclear basic protein from *Mytilus trossulus*, *FEBS Lett. 363*, 37−40.
22. Subirana, J. A., and Colom, J. (1987) Comparison of protamines from freshwater and marine bivalve molluscs: Evolutionary implications, *FEBS Lett. 220*, 193−196.
23. Carlos, S., Jutglar, L., Borrell, I., Hunt, D. F., and Ausió, J. (1993) Sequence and characterization of a sperm-specific histone H1-like protein of *Mytilus californianus*, *J. Biol. Chem. 268*, 185−194.
24. Jutglar, L., Borrell, J. I., and Ausió, J. (1991) Primary, secondary, and tertiary structure of the core of a histone H1-like protein from the sperm of *Mytilus*, *J. Biol. Chem. 266*, 8184−8191.
25. Sambrook, J., Fritsch, E. F., and Maniatus, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbour, New York.
26. Zhang, Z., and Gurr, S. J. (2000) Walking into the unknown: A "step down" PCR-based technique leading to the direct sequence analysis of flanking genomic DNA, *Gene 253*, 145−150.
27. Ausió, J., and Subirana, J. A. (1982) Conformational study and determination of the molecular weight of highly charged basic proteins by sedimentation equilibrium and gel electrophoresis, *Biochemistry 21*, 5910−5918.
28. Zhang, F., Lewis, J. D., and Ausió, J. (1999) Cysteine-containing histone H1-like (PL-I) proteins of sperm, *Mol. Reprod. Dev. 54*, 402−409.
29. Papoutsopoulou, S., Nikolakaki, E., Chalepakis, G., Kruft, V., Chevaillier, P., and Giannakouros, T. (1999) SR protein-specific kinase 1 is highly expressed in testis and phosphorylates protamine 1, *Nucleic Acids Res. 27*, 2972−2980.
30. Poccia, D. L., and Green, G. R. (1992) Packaging and unpackaging the sea urchin sperm genome, *Trends Biochem. Sci. 17*, 223−227.
31. Domenjoud, L., Nussbaum, G., Adham, I. M., Greeske, G., and Engel, W. (1990) Genomic sequences of human protamines whose genes, PRM1 and PRM2, are clustered, *Genomics 8*, 127−133.
32. Hunt, J. G., Kasinsky, H. E., Elsey, R. M., Wright, C. L., Rice, P., Bell, J. E., Sharp, D. J., Kiss, A. J., Hunt, D. F., Arnott, D. P., Russ, M. M., Shabanowitz, J., and Ausió, J. (1996) Protamines of reptiles, *J. Biol. Chem. 271*, 23547−23557.
33. Gui, J. F., Lane, W. S., and Fu, X. D. (1994) A serine kinase regulates intracellular localization of splicing factors in the cell cycle, *Nature 369*, 678−682.
34. Rossi, F., Labourier, E., Forne, T., Divita, G., Derancourt, J., Riou, J. F., Antoine, E., Cathala, G., Brunel, C., and Tazi, J. (1996) Specific phosphorylation of SR proteins by mammalian DNA topoisomerase I, *Nature 381*, 80−82.
35. Cobb, J., Reddy, R. K., Park, C., and Handel, M. A. (1997) Analysis of expression and function of topoisomerase I and II during meiosis in male mice, *Mol. Reprod. Dev. 46*, 489−498.
36. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database, *Nucleic Acids Res. 30*, 276−280.
37. Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. (1999) SMART: Identification and annotation of domains from signalling and extracellular protein sequences, *Nucleic Acids Res. 27*, 229−232.
38. Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., and Sweet, R. M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding, *Nature 362*, 219−223.
39. Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003) SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res. 31*, 3381−3385.
40. Heath, D. D., and Hilbish, T. J. (1998) *Mytilus* protamine-like sperm-specific protein genes are multicopy, dispersed, and closely associated with hypervariable RFLP regions, *Genome 41*, 587−596.
41. Giorgini, F., Davies, H. G., and Braun, R. E. (2002) Translational repression by MSY4 inhibits spermatid differentiation in mice, *Development 129*, 3669−3679.
42. Watson, C. E., and Davies, P. L. (1999) Recent and rapid amplification of the sperm basic nuclear protein genes in winter flounder, *Biochim. Biophys. Acta 1444*, 337−345.
43. Ohno, S., and Becak, M. L. (1993) Can a protein influence the fate of its own coding sequence?: The amino- and carboxyl-terminal regions of H1 histone, *Proc. Natl. Acad. Sci. U.S.A. 90*, 7341−7345.
44. Oliva, R., and Dixon, G. H. (1990) Vertebrate protamine gene evolution I. Sequence alignments and gene structure, *J. Mol. Evol. 30*, 333−346.
45. Yiu, G. K., and Hecht, N. B. (1997) Novel testis-specific protein-DNA interactions activate transcription of the mouse protamine 2 gene during spermatogenesis, *J. Biol. Chem. 272*, 26926−26933.
46. Hecht, N. B. (1989) *Mammalian Protamines and Their Expression*, CRC Press Inc., Boca Raton, FL.
47. Steger, K. (1999) Transcriptional and translational regulation of gene expression in haploid spermatids, *Anat. Embryol. 199*, 471−487.
48. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., and Gautheret, D. (2000) Patterns of variant polyadenylation signal usage in human genes, *Genome Res. 10*, 1001−1010.
49. Lewis, J. D., Song, Y., De Jong, M. E., Bagha, S. M., and Ausió, J. (2003) A walk through vertebrate and invertebrate protamines, *Chromosoma 111*, 473−482.
50. Giorgini, F., Davies, H. G., and Braun, R. E. (2001) MSY2 and MSY4 bind a conserved sequence in the 3′ untranslated region of protamine 1 mRNA in vitro and in vivo, *Mol. Cell. Biol. 21*, 7010−7019.
51. Ausió, J. (1986) Structural variability and compositional homology of the protamine-like components of the sperm from the bivalve mollusks, *Comp. Biochem. Physiol., B 85*, 439−449.
52. Agelopoulou, B., Cary, P. D., Pataryas, T., Aleporou-Marinou, V., and Crane-Robinson, C. (2004) The sperm-specific proteins of the edible oyster (European flat oyster (*Ostrea edulis*)) are

products of proteolytic processing, *Biochim. Biophys. Acta 1676*, 12−22.
53. Lewis, J. D., and Ausió, J. (2002) Protamine-like proteins: Evidence for a novel chromatin structure, *Biochem. Cell Biol. 80*, 353−361.
54. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res. 25*, 4876−4882.
55. Wells, D., and Brown, D. (1991) Histone and histone gene compilation and alignment update, *Nucleic Acids Res. 19 Suppl.*, 2173-2188.
56. Schneider, T. D., and Stephens, R. M. (1990) Sequence logos: A new way to display consensus sequences, *Nucleic Acids Res. 18*, 6097−6100.

BI0360455