Histone H1 and the evolution of protamines

by

John David MacLean Lewis
B.Sc.Hon, University of Western Ontario at London, 1993

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Biochemistry and Microbiology

We accept this thesis as conforming
to the required standard

Dr. J. Ausió, Supervisor (Department of Biochemistry and Microbiology)

Dr. T.W. Pearson, Department Member (Department of Biochemistry and Microbiology)

Dr. C. Upton, Department Member (Department of Biochemistry and Microbiology)

Dr. E.E. Ishiguro, Department Member (Department of Biochemistry and Microbiology)

Dr. P.C. Wan, Outside Member (Department of Chemistry)

Dr. H.E. Kasinsky, External Examiner (Department of Zoology, UBC)

Supervisor: Dr. Juan Ausió

I. ABSTRACT

It has been proposed that protamines have evolved vertically from an ancestral histone H1. My research has concentrated mainly on the investigation of this proposal by characterizing the sperm nuclear basic proteins (SNBPs) and their genes from a diverse range of organisms which employ histones, protamines, or protamine-like proteins to achieve sperm chromatin compaction. The complete gene sequences were obtained for the large histone H1-related protamine-like PL-I of the bivalve mollusc *Spisula solidissima*, the small protamine-like PL-III protein of related bivalve *Mytilus californianus*, and the protamine of the squid, *Loligo opalescens*, which is the first invertebrate protamine gene to be characterized. In addition, a full-length cDNA from the novel protamine and histone H1-related sperm nuclear protein of the primitive chordate, *Styela montereyensis*, was isolated and characterized. This genetic data, beyond providing valuable information on the regulation and organization of the heterogeneous family of SNBPs, has provided unequivocal support to the hypothesis that the chromatin-condensing protamines of the sperm have evolved from the chromatin-condensing histones of somatic cells. This has in turn allowed a more accurate tracing of the origin of histone H1, protamines and protamine-like proteins in both the protostomes and deuterostomes.

Examiners:

Dr. J. Ausió, Supervisor (Department of Biochemistry and Microbiology)

Dr. T.W. Pearson, Department Member (Department of Biochemistry and Microbiology)

Dr. C. Upton, Department Member (Department of Biochemistry and Microbiology)

Dr. E.E. Ishiguro, Department Member (Department of Biochemistry and Microbiology)

Dr. P.C. Wan, Outside Member (Department of Chemistry)

Dr. H.E. Kasinsky, External Examiner (Department of Zoology, UBC)

II. TABLE OF CONTENTS

SECTION B : PROTAMINE-LIKE PROTEINS

IV. LIST OF TABLES

IV. LIST OF FIGURES

Chapter 7

Chapter 8

V. ABBREVIATIONS

A - adenine

APS - ammonium persulfate

bp - base pair

BSA - bovine serum albumin

C - cytosine

cDNA - complementary deoxyribonucleic acid

Da - Dalton

DEPC - diethyl pyrocarbonate

DNA - deoxyribonucleic acid

DNase - deoxyribonuclease

dNTP - deoxynucleoside triphosphate

dT - deoxythymidine

DTT - dithiothreitol

EDTA - ethylenediaminetetraacetic acid

ESI-MS - electrospray ionization mass spectrometry

FPLC - fast performance liquid chromatography

G - guanine

HCl - hydrochloric acid

HPLC - high performance liquid chromatography

IPTG - isopropylthio-β-D-galactoside

kDa - kiloDalton

LB - Luria-Bertani

LTR - long terminal repeat

mRNA - messenger ribonucleic acid

$MgCl_2$ - magnesium chloride

MOPS - 3-(N-morpholino)propane sulfonic acid

NaCl - sodium chloride

NDSB - nondenaturing sample buffer

OD - optical density

PAGE - polyacrylamide gel electrophoresis

PCA - perchloric acid

PCR - polymerase chain reaction

PL - protamine-like

RNA - ribonucleic acid

RNase - ribonuclease

rNTP - ribonucleoside triphosphate

SDS - sodium dodecyl sulfate

SNBP - sperm nuclear basic protein

T - thymine

T_m - melting temperature

TE - tris-EDTA

TEMED - N,N,N',N'-tetramethylethlenediamine

TLCK - tosyllysine chloromethyl ketone

X-gal - 5-bromo-4-chloro-3-indoyl-β-D-galactoside

VI. ACKNOWLEDGEMENTS

After all of these years as a student again, there are a lot of people who are in no small way involved in the realization of this thesis, and for which I would like to express thanks:

Juan, for all of his enthusiasm, his inspiration, his understanding, and his friendship which was always there when I needed it.

The members current and past of the Ausió lab, collaborators, co-conspirators, co-dependents, and even co-habitators!

Harold for his enthusiasm for the connectedness of everything, and some fantastic discussions about histone evolution.

Aaron, Kim, Glen, Rodney, Ellen, Liz, Dustin and everyone else in the Department who got things done for me in the nick of time or let me use their stuff in the middle of the night when I was trying to get things done myself in the nick of time.

God for giving *Mytilus* so many pseudogenes.

Everyone at Asilomar, ASCB and Friday Harbour who at least pretended to be excited about what the Sperm Guy had to say.

Mom and Dad for all of their love and unending support (and support!) during my long years of being a starving student all over again.

My new Mom for all of her love and support and encouragement.

Mike, Ian, Peter, Andrew, and Mike M, for being the most wicked bunch of guys that could possibly be.

Most of all, Nat, the love of my life, who makes my life feel more meaningful every single day I spend with her. It was the pursuit of this thesis that brought us together and for that I am forever grateful.

INTRODUCTION

In the nuclei of all eukaryotic cells, DNA is highly folded and organized by histones and non-histone proteins into chromatin (van Holde 1989). At the structural level, the most important function of this assembly is to compact the lengthy DNA molecule inside the limited available nuclear space. In somatic cells, chromatin is a dynamic structure as DNA must be accessible for replication, repair and transcription. The major protein component of chromatin is histones, and these can be structurally grouped in two major categories: the "core" and "linker" histones. Distinct levels of chromatin organization are dependent on the dynamic higher order structure of nucleosomes, which represent the basic repeating unit of chromatin (Figure 1). Each nucleosome core particle consists of 146 bp of DNA wrapped around a histone octamer core in approximately two left handed superhelical turns, the protein constituent consisting of a $(H3\text{-}H4)_2$ tetramer associated with 2 adjacent H2A-H2B dimers (Eickbush and Moudrianakis 1978). The core histones (histones H2A, H2B, H3, and H4) are relatively small proteins (11,000 to 16,000 Da), and have an arginine and lysine content of over 20% (Wolffe 1992). The structure of core histones consists of a well-characterized globular "histone" motif (Luger et al. 1997), flanked by less structured amino- and carboxy-terminal domains commonly referred to as "tails". Core histones are amongst the most highly evolutionarily conserved proteins (Isenberg 1978). Adjacent nucleosomes are connected by a variable stretch of linker DNA, which is often associated with histone H1, which as a result is commonly referred to as the "linker histone". Histone H1 is larger (>20,000 Da) (Wolffe 1992) and more lysine-rich than the core histones (Johns 1971; Isenberg 1978). Linker histones contain a trypsin-resistant globular

core with charged amino- and carboxyl-terminal tails. The crystallographic structure of the globular core has revealed that it adopts a conformation known as the "winged helix" motif (Ramakrishnan et al. 1993). This region of histone H1 interacts with the

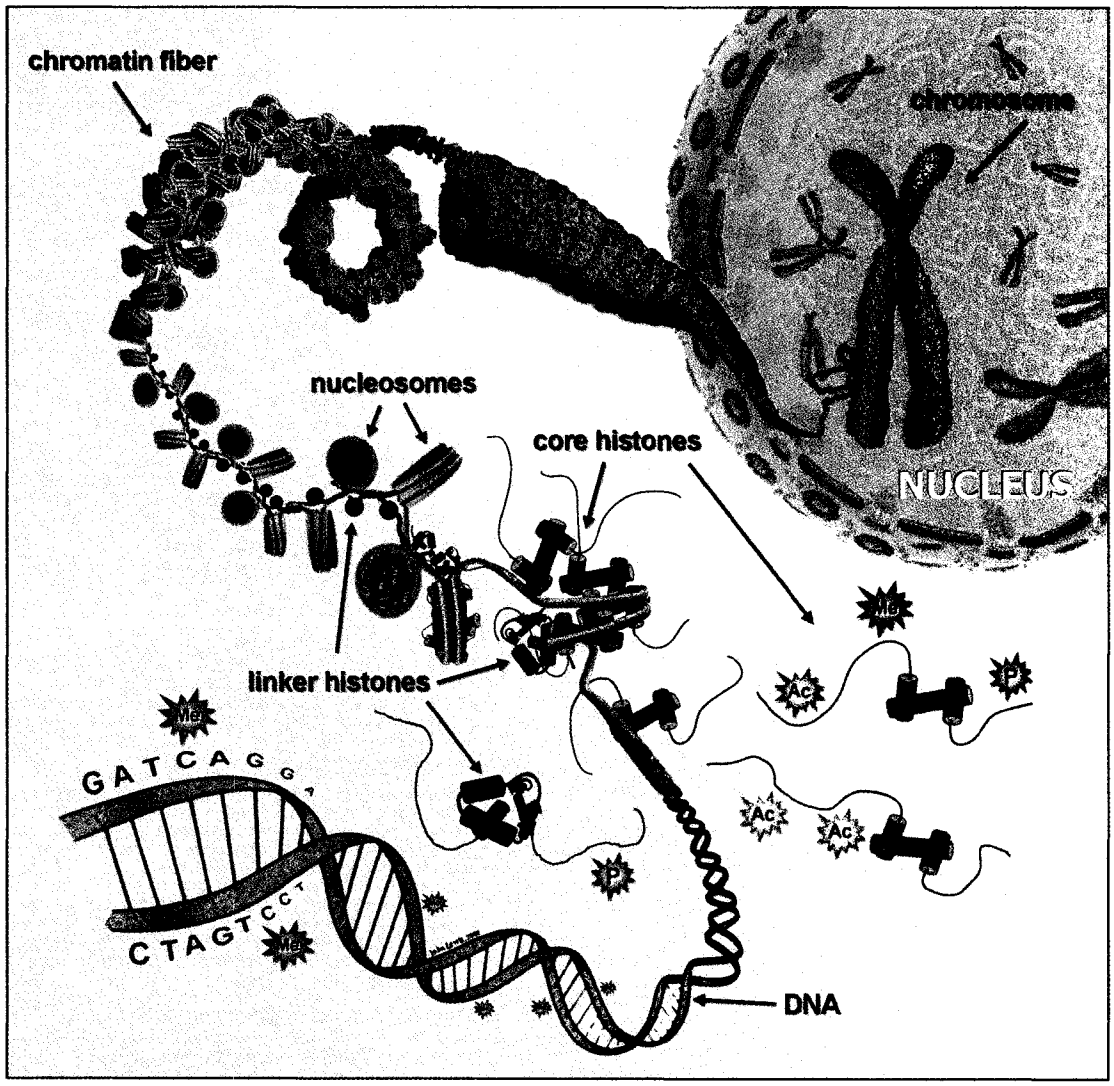


Figure 1. A schematic representation of successive levels of chromatin folding, from free DNA to its packaging within the nucleosome to the formation of higher order structures and finally a condensed metaphase chromosome. Original artwork by John Lewis.

nucleosome at a region close to the entry and exit points of the DNA strand (Zhou et al. 1998). In contrast to core histones, linker histones are much less conserved evolutionarily (Isenberg 1978; Cole 1984). When histone H1 becomes associated with the nucleosome,

a total of 168 bp of DNA is protected, and this structure is referred to as the chromatosome (van Holde 1989). Upon binding of histone H1 to the linker DNA, the polynucleosomal fiber will fold into a chromatin fiber of 30 nm in diameter (van Holde 1989), contributing significantly to the formation of a compact chromatin structure.

SPERMATOGENESIS

All sexually reproducing organisms have a specialized developmental pathway for gametogenesis, in which diploid cells undergo meiosis to produce haploid germ cells. Spermatogenesis is the biological process whereby a gradual transformation of germ cells into spermatozoa occurs over an extended period of time. This process involves cellular proliferation by repeated mitotic divisions, duplication of chromosomes, genetic recombination through crossing-over, reduction-division by meiosis to produce haploid spermatids, and finally terminal differentiation of spermatids into spermatozoa (Figure 2).

Spermatogonia, which comprise the first phase, are the most immature cells and are located along the base of the seminiferous epithelium. They proliferate by mitotic division and multiply repeatedly to continually replenish the germinal epithelium. Spermatogonia divide mitotically into both stem cells that remain along the base (type A spermatogonia) as well as committed cells, the B spermatogonia, that will progress to become spermatozoa. In most species, these B spermatogonia are the last to divide by mitosis. Their division produces the first cell of the second phase, the preleptotene spermatocyte, which migrates upwards away from the base of the seminiferous tubule and crosses through the Sertoli-Sertoli junction.

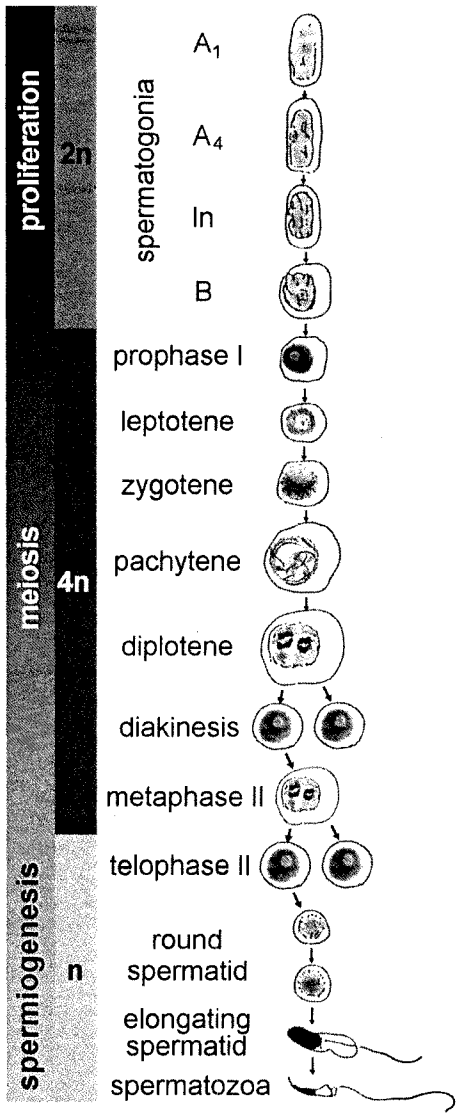


Figure 2. Diagram depicting the stages of mammalian spermatogenesis and meiosis, showing cell morphology at each relevant stage of spermatogenesis. Adapted with permission from (Lewis et al. 2003a).

Reduction-division is a biological mechanism by which a single germ cell doubles its DNA content, then divides twice to produce four individual haploid germ cells. Initially, a round of DNA synthesis occurs to produce the preleptotene spermatocytes (4N). Prophase of the first meiotic division may last for nearly three weeks, during which time the chromosomes first unravel as thin impaired filaments in leptotene. Homologous chromosomes become paired in the zygotene cell, and the synaptonemal complex is formed. Pachytene spermatocytes enlarge greatly as the chromosomes become shorter and thicken. During diplotene the synaptonemal complex dissociates and the chromosomes spread apart in the nucleus, followed by diakinesis, where the nuclear envelope disappears and chromosomes condense. The subsequent meiotic divisions occur rapidly, producing first small secondary spermatocytes (2N) after meiosis I and then very small round spermatids (1N) after meiosis II.

The haploid germ cells undergo a prolonged phase of terminal differentiation

during spermiogenesis. Dramatic species-specific changes occur, including the following major modifications:

(i) The nucleus elongates and the chromatin is condensed into a very dark-staining structure.

(ii) the Golgi apparatus produces a lysosomal-like granule that elaborates over the nucleus to form the future acrosome.

(iii) the cell forms a long tail lined with mitochondria in the proximal region as excess cytoplasm is discarded.

The final mature spermatozoan cell consists of four parts: the head, acrosome, midpiece and tail. Progression through spermatogenesis is associated with significant transformations in chromosome condensation and organization. The structure of chromatin, however, is changed most dramatically during the final stages of spermiogenesis as the genome is condensed and inactivated by the binding of the sperm nuclear basic proteins (SNBPs).

SPERM NUCLEAR BASIC PROTEINS

Classification and composition

Early studies of chromatin showed that while the major nucleoprotein complexes in somatic cells were histones, the protein composition of chromatin in sperm cells consisted of either histones (i.e. carp (Kossel 1928)) or protamines (i.e. salmon (Miescher 1874)). The continued chemical characterization of the SNBPs revealed that unlike the somatic histones, these sperm proteins exhibited a large degree of compositional variability and structural heterogeneity (Felix 1960; Ando et al. 1973; Subirana et al.

1973).

An early attempt to classify the SNBPs was carried out by David Bloch in 1969 (Bloch 1969), who distinguished among the following types:

(i) *Salmo* type: or "monoprotamines", arginine-rich protamines from fish such as salmine from the salmon (Miescher 1874).

(ii) *Mammalian* type: or "stable protamines", with a high arginine content but also containing sulfhydryl groups such as the protamine P2 from human (Domenjoud et al. 1990).

(iii) *Mytilus* type: or "di/triprotamines", containing high levels of two or three of the basic amino acids lysine, arginine or histidine. This type was the most heterogeneous of the groups and included those proteins whose composition was intermediate to that of histones and protamines, such as those from the surf clam (Ausió 1986).

(iv) *Rana* type: sperm-specific and/or somatic-type histones similar to those found in somatic cells, such as those from grass carp (Kadura et al. 1983).

(v) *Crab* type: containing no basic proteins in the mature sperm, resulting in a large uncondensed nucleus (Vaughn et al. 1969).

As more information has become available, the relationships underlying SNBP variability have become somewhat clearer. In recent years, studies have gathered a wealth of information regarding SNBPs from a range of both distant and closely related organisms. Consequently, the classification has been simplified and organized based on both protein structure and composition to comprise three main groups, the Histone type (H), the Protamine type (P), and the Protamine-like type (PL) (Ausió 1986). This is the

classification in general use at present and will be used for the remainder of this thesis.

Histones (H Type)

The H type corresponds to Bloch's *Rana* type. These proteins consist of sperm-specific and/or somatic-type histones that are similar in structure to those found in somatic cells. While they resemble very closely their somatic counterparts, there are often sperm-specific variants of H1, H2B and H2A. Examples include the spH1 and spH2B from the sperm of echinoderms (Zalenskaya et al. 1980; Poccia and Green 1992), the sperm-specific variants of H1, H2B and H2A from grass carp (Kadura et al. 1983), and the H1 variants found in bivalve molluscs such as the giant Pacific oyster. They are presumably involved in mediating the highly compacted state of sperm chromatin in these organisms (Poccia and Green 1992).

Protamines (P Type)

The P type SNBPs are relatively small (generally $4000 \leq M_r \leq 12000$), arginine-rich (Arg ≥ 30%) proteins that correspond to Bloch's *Salmo* and *Mammalian* types. During spermiogenesis, these proteins replace the majority of the histone complement, either directly or subsequent to the appearance of transition proteins and/or protamine precursors. This group includes the protamines of mammals, marsupials, birds, fish and reptiles (reviewed by (Oliva and Dixon 1991)), and those that have been identified more recently in the invertebrates (Wouters Tyrou et al. 1995; Lewis et al. 2003b) (Fig. 3, lane SL). Please refer to Chapter 3 for an in-depth review of protamines.

Protamine-like proteins (PL type)

It is the third group, the PL type SNBPs, that are structurally quite heterogeneous, while maintaining a very consistent chemical composition; one intermediate to that of protamines and histones. While initially described in the bivalve molluscs, they are pervasive across the animal kingdom, having been identified in such phylogenetically diverse organisms as Cnidaria (Rocchini et al. 1995b; Rocchini et al. 1996), chordates (Saperas et al. 1992), and vertebrates (Saperas et al. 1994). Despite the common function of PL proteins, there is a remarkable variability in the size and number of expressed PL proteins in the sperm of even closely related organisms. Like protamines, PL proteins are highly basic, with an arginine + lysine content of at least 35-50 mol%, and some also contain cysteine (Zhang et al. 1999). They can vary in molecular

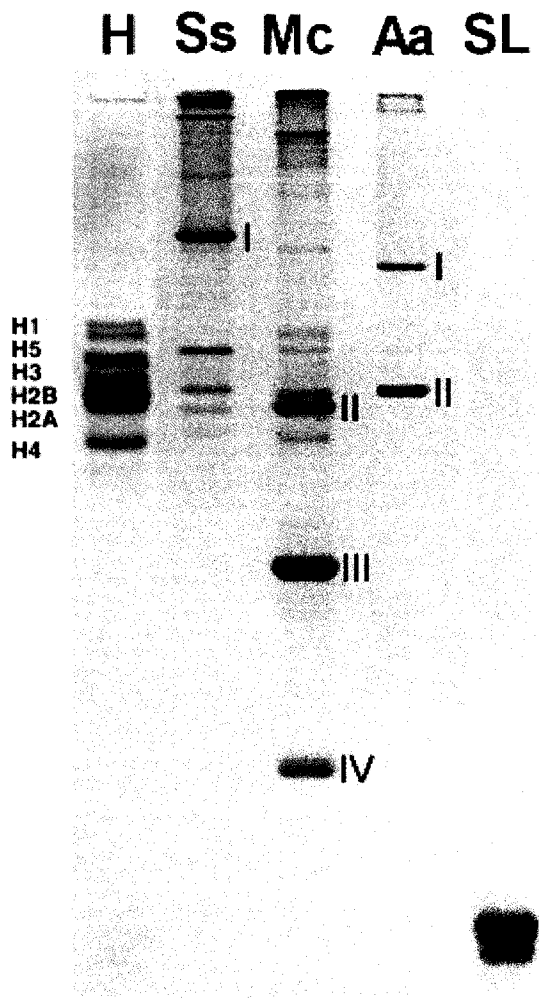


Figure 3. Urea (2.5 M)-acetic acid (5%) polyacrylamide gel electrophoresis analysis of the SNBPs from several representative invertebrate and vertebrate organisms. *3, Aurelia aurita* (moon jellyfish, class Scyphozoa, phylum Cnidaria); *5, S. solidissima* (surf clam, phylum Mollusca, class Bivalvia); *6, Mytilus californianus* (California mussel, phylum Mollusca, class Bivalvia); Chicken erythrocyte histones *(H)* and salmine *(SL,* salmon protamine) were used as markers. The Roman numerals I, *II, III,* and *IV* designate the PL-I, PL-II, PL-III, and PL-IV components.

mass from 6500 Da up to 200000 Da for the SNBPs of winter flounder (Watson and Davies 1998).

Due to their heterogeneity, PL proteins are generally sub-classified into four basic categories based on their relative electrophoretic mobilities; PL-I, PL-II, PL-III, and PL-IV (Ausió 1986) (see Fig. 3, lanes Ss, Mc & Aa). In addition, since many PL proteins have been identified in the bivalve molluscs, the bivalve molluscs themselves have been classified according to the number and size of PLs present in their mature sperm (Ausió 1986): Pectinidae (group O), Veneridae (group I), Cardiidae (group II), Tellinidae (group III) and Mytilidae (group IV).

Pectinidae (group O)

The SNBPs of this group are histones that are similar to those found in somatic cells (Ausió 1992), but containing a sperm-specific H1 with a lower electrophoretic mobility than the somatic H1 and also displaying microheterogeneity. This observed microheterogeneity may be the result of post-translational cleavage of a PL precursor, a situation that is found in protamines of both vertebrates and invertebrates (Lewis et al. 2003b), and also in other PL proteins (Carlos et al. 1993a; Bandiera et al. 1995). An example of a member of this group is the bivalve mollusc, *Swiftopecten swifti* (Zalenskaya et al. 1982).

Veneridae (group I)

The organisms in this group have a single PL protein of very low electrophoretic mobility (Ausió 1992) (Fig. 3, lane Ss). The sperm PL of the surf clam, *Spisula*

solidissima is quite large, containing significant amounts of lysine and arginine, 24.8 mol% and 23.1 mol%, respectively (Ausió and Subirana 1982b). Like histone H1, the PL-I proteins have an internal trypsin resistant globular core (Ausió et al. 1987). Two other members of this group, *Agriodesma saxicola* and *Mytilimeria nuttalli*, have PL-I proteins with the highest arginine content found within the PL classification (Ausió 1992).

Cardiidae (group II)

Sperm from the organisms in this group express two PL proteins, a PL-I and a PL-II. While the PL-I has a low electrophoretic mobility, the PL-II proteins of this group have a similar mobility to histone H4 in urea-acetic acid PAGE (Ausió 1992) (see Fig. 3, lane Aa). The sperm of the razor clam, *Ensis minor*, contains proteins designated EM6 and EM1, which correspond to the PL proteins PL-I and PL-II respectively (Giancotti et al. 1983). Like the PL-I of *Spisula* (group I), these proteins contain significant amounts of lysine and arginine, while only EM6 (PL-I) possesses a trypsin-resistant globular core (Bandiera et al. 1995). Isolation of the cDNA of these SNBPs has revealed that EM6 and EM1 are products of post-translational cleavage of a PL precursor (Bandiera et al. 1995).

Tellinidae (group III)

In the sperm of this group of bivalve molluscs, there are three PL proteins: a PL-I, PL-II, and PL-III. PL-III exhibits some electrophoretic microheterogeneity and has a higher electrophoretic mobility than PL-I, PL-II, and all of the somatic histones (Ausió 1992). The sperm of the bent-nose clam, *Macoma nasuta* contains a PL-I that, like the PL-I of *Spisula*, is rich in lysine and arginine and has a trypsin-resistant globular core

(Ausió 1988). The PL-II and PL-III of this organism do not contain a trypsin-resistant core, and are very similar to each other in amino acid composition. The PL-II and PL-III of *Macoma nasuta* contain 138 and 68 amino acids, respectively (Ausió 1988).

Mytilidae (group IV)

The sperm of Mytilidae, like the Tellinidae, contain three PL proteins. These SNBPs, however, are of higher electrophoretic mobility, consisting of PL-II, PL-III and PL-IV (Fig. 3, lane Mc). PL-IV has the highest electrophoretic mobility seen of all PL proteins (Ausió 1992). Work within this group has concentrated on the SNBPs of the closely related *Mytilus californianus* (Ausió and McParland 1989; Jutglar et al. 1991; Carlos et al. 1993b), *Mytilus trossulus* (Mogensen et al. 1991; Rocchini et al. 1995a), and *Mytilus edulis* (Subirana et al. 1973; Ausió and Subirana 1982c). The PL-II of *Mytilus sp.* possesses a trypsin-resistant globular core, while PL-III and PL-IV do not. Similar to the proteins in *Ensis minor*, cDNA data of the PL-II has revealed that PL-II and PL-IV are products of post-translational cleavage of a PL precursor (Carlos et al. 1993a).

Rationale for the study of bivalve molluscs

Much of the study of SNBPs has been carried out on the bivalve molluscs, for three principal reasons. First, many species of bivalve molluscs, especially mussels, are easy to collect around Vancouver Island. Second, since molluscs achieve fertilization in the open water, they amass a very large amount of sperm in their gonads, which can account for up to 80% of their weight when they are "ripe". They are, therefore, an extremely abundant source of SNBPs, and very large preparative amounts can be

obtained with relative ease. Finally, examples of all three classifications (H, PL, P) of SNBPs can be found in the sperm of different species of bivalve molluscs. For example, the giant Pacific oyster, *Crassostrea gigas*, and the bay scallop, *Aequipecten irradians*, have SNBPs of the H type (Ausió 1986). The sperm of the surf clam, *Spisula solidissima*, the razor clam, *Ensis minor*, and the California mussel, *Mytilus californianus*, contain the PL type of SNBPs (Ausió 1986), while the octopus, *Eledone cirrhosa*, and the snail, *Gibbula divaricata*, possess P type SNBPs (Subirana et al. 1973; Gimenez-Bonafe et al. 2002).

SPERMIOGENESIS AND HISTONE-SNBP REPLACEMENT

There is a dramatic remodeling of local and global chromatin structure during the final stages of spermiogenesis, as somatic-type histones are replaced by the sperm nuclear basic proteins. A number of organisms replace the somatic-type histones with germinal sperm-specific histone variants that are ultimately responsible for condensation of the sperm chromatin (H type). In the majority of organisms, however, the germinal histones are replaced during spermiogenesis by the even more specialized PL or P type SNBPs.

In those organisms that contain protamines in the mature sperm, the protamine mRNA is transcribed much earlier than its expression, usually in the post-meiotic spermatid stage. Newly synthesized protamine mRNAs are stored for up to 7 days before translational activation (Giorgini et al. 2002). In many mammals, germinal histones are first displaced by the highly basic transition proteins (TP1 and TP2) before protamines are deposited. It is unclear exactly what the function of the transition proteins is, but temporal expression studies have shown that during rat spermiogenesis, TP2 is expressed

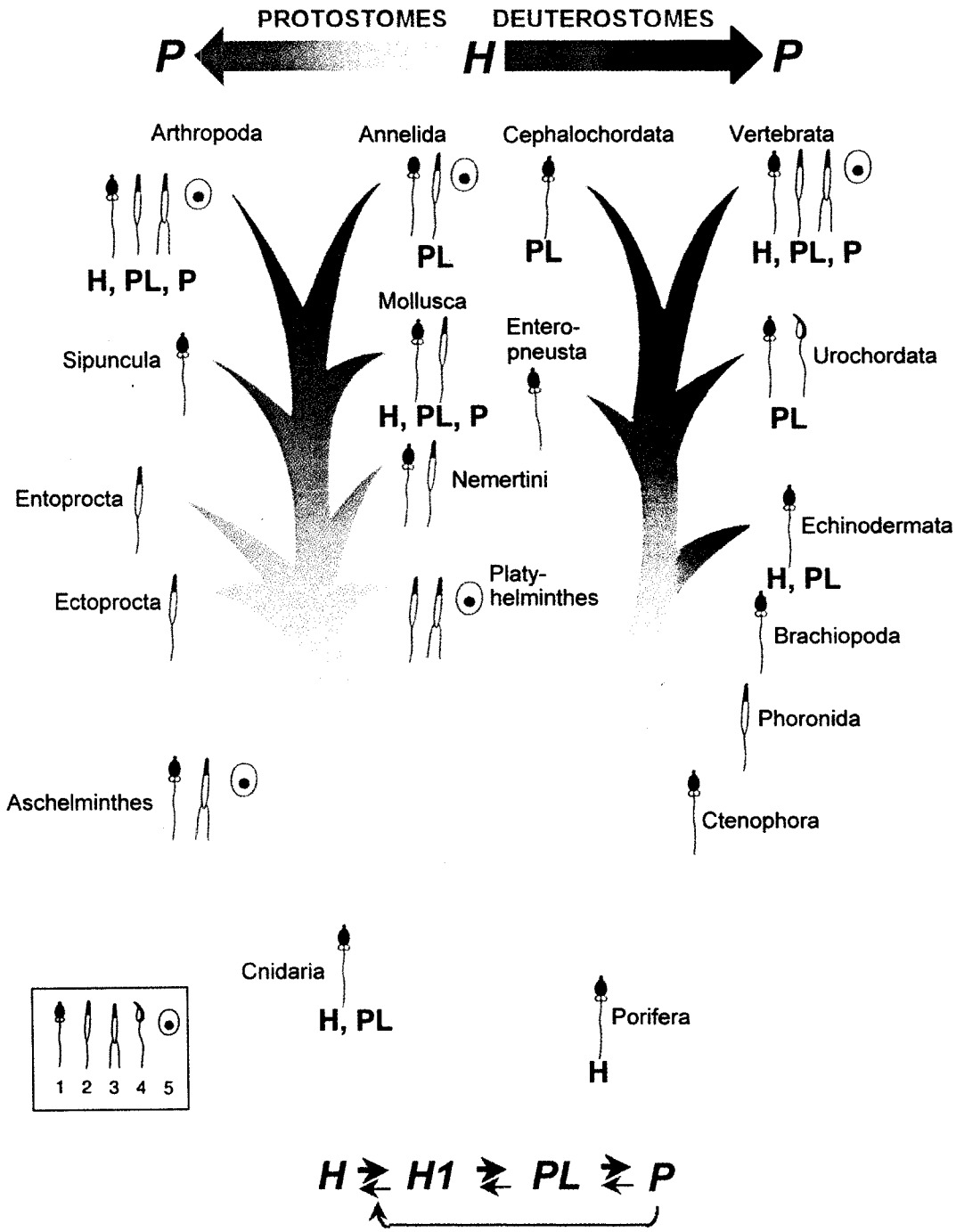


Figure 4. Schematic representation of the evolution of the major SNBP types. The basic pattern of evolution among the different SNBP types is shown at the base of the tree with black arrows. H, primitive histone protein precursor; H1, primitive sperm histone H1 precursor; P, arginine-rich protamine. The red arrows indicate the existence of reversions among the different major SNBP types (Ausió 1999). This pattern appears to have occurred on repeated occasions during evolution. The SNBPs present in different taxonomic groups along the phylogenetic tree are shown in black, as in Fig. 1. The pink- and blue-colored arrows at the top indicate the direction of the evolutionary trend from primitive histone protein to arginine-rich protamine in the protostome and deuterostome branches.

first and may be involved in the initial disassembly of the ordered nucleosome structure. The expression of TP1 begins after the appearance of TP2 and may facilitate the deposition of protamines (Kistler et al. 1996), although it has been suggested that replacement of TPs by protamines could occur simply due to electrostatic competition for the DNA (Oliva and Dixon 1991).

The degree that histones are replaced by the SNBPs varies in a species-specific manner. In humans, typically 85% of the nucleosomal structure is replaced by a nucleoprotamine complex. The remaining 15% retain nucleosomes containing germinal histones. Fluorescence *in situ* hybridization and confocal microscopy studies with sperm nuclei have described an organized and well-defined higher order compartmentalization of chromatin (Zalensky et al. 1995). The nuclear architecture in the human sperm is characterized by the clustering of the 23 centromeres into a compact chromocenter positioned well inside the nucleus. The ends of the chromosomes are exposed to the nuclear periphery where the telomere sequences of the chromosome arms are joined into dimers, looping the chromosomes into a hairpin-like configuration (Zalensky et al. 1995). Studies in which the sperm chromatin structure is specifically probed with DNase I have revealed that the regions that remain packaged in nucleosomes include the telomeres and also the promoters and relevant nuclear matrix attachment regions (MARS) of genes active during chromatin condensation (specifically PRM1, PRM2) (Choudhary et al. 1995; Wykes and Krawetz 2003), while the members of the β-globin gene family, for instance, were tightly packaged with protamines (Gardiner Garden et al. 1998). Genes important for early embryonic development may also be located in the nucleohistone fraction (Gatewood et al. 1987). The nucleosomal fraction of mammalian sperm

chromatin has also been shown to be enriched in histone variants such as H2A.X and H2A.Z (Gatewood et al. 1990).

In those organisms that express PL proteins, the mature sperm retain a higher proportion of germinal histones, from 30-40% of the total SNBPs. While there are no transition proteins, the expression of SNBP precursors and their subsequent post-translational cleavage may provide added levels of control. As in mammalian sperm, the chromatin in PL-containing sperm may consist of two distinct fractions of chromatin organization; a nucleosomal fraction containing somatic-type histones and a fraction highly saturated with protamine-like proteins. Due to the similarity of many of the PL proteins to linker histones, other novel chromatin structures are possible (Lewis and Ausió 2002) (see Chapter 6).

EVOLUTION OF SPERM NUCLEAR BASIC PROTEINS

All three main types of sperm nuclear basic proteins are widespread through the phylogenetic groups in the animal kingdom (Saperas et al. 1997). Organisms that replace their histones with protamines in the mature sperm are always found at the furthermost tips of the evolutionary branches (Ausió 1999), while the histone type of SNBPs are found in the sperm of more primitive organisms such as the sponge *Neofibularia* (Ausió et al. 1997).

Regardless of the variability in size and number of the sperm nuclear basic proteins, they are all significantly enriched in the basic amino acids arginine and lysine. Early theories proposed to account for the relationship between SNBPs involved the partial gene duplication of a pentapeptide core of Ala-Arg-Arg-Arg-Arg (Black and

Dixon 1967), with subsequent insertions and deletions evolving to the modern day protamines. The idea that protamines had evolved from a histone precursor was introduced in 1973, when Subirana proposed a novel mechanism of vertical evolution. He suggested that an ancient histone H1 had evolved from a somatic-type histone precursor, then proceeded through a number of PL type intermediates until it finally became a

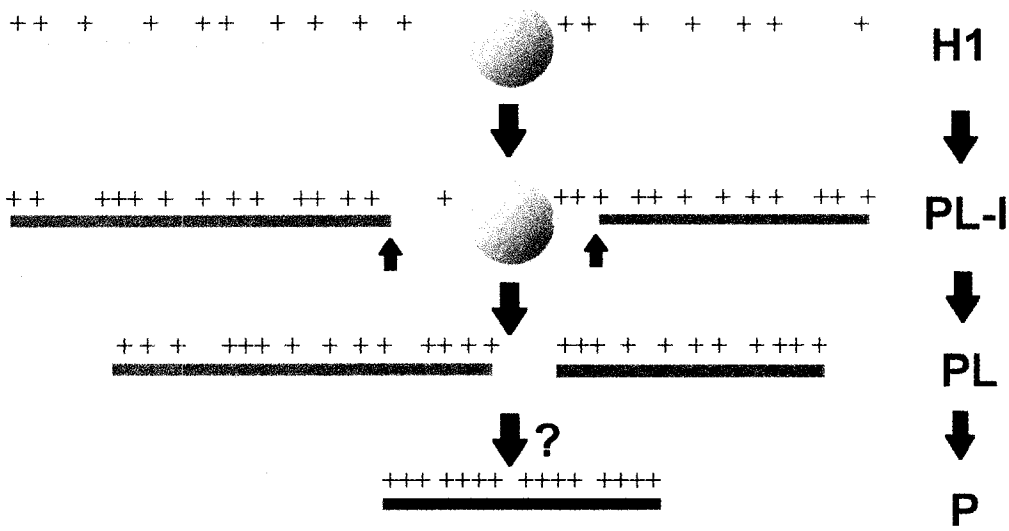


Figure 5. Proposed evolution of the sperm nuclear basic proteins. Histone H1 precursors expand and become more arginine-rich like the PL-I of bivalve molluscs. Proteins segregate first by post-translational cleavage, then genetic segregation to lose the H1 winged helix. These smaller PL proteins increase in arginine content until they are indistinguishable from protamines.

protamine (Subirana et al. 1973). This theory was later refined, as it became apparent that histone H1 and the core histones had separate origins (see Chapter 2 for an extensive discussion of linker histones). Ausió proposed that all of the sperm nuclear basic proteins arose from a primitive histone H1 precursor (Ausió 1986). Since H1s are more lysine-rich than arginine-rich, over the course of evolution the arginine content would rise slowly and the protein would become more "protamine-like", or similar to the H1-related PL-I proteins (Ausió 1999). This PL-I would expand and then begin to fragment (Fig. 5), first

at the post-translational level through cleavage and then at the genetic level, such as seemed to be the case in *Mytilus*. Continued evolution and eventual down-regulation of the H1-related portion of the proteins would favour the expression of the smaller arginine-rich PL proteins such as PL-IV from *Mytilus*. Because H1-related proteins are thought to coordinate with the nucleosome in some way, as the H1-related proteins were lost, the amount of nucleosomes required in the mature sperm would decrease. This would set the stage for the final evolutionary step to protamines.

It has also been postulated that protamines arose not from an ancient eukaryotic protein, but instead have a retroviral origin (Jankowski et al. 1986). In his characterization of the fish protamine genes, Dixon's group found a high incidence of viral long terminal repeat sequences near protamine genes (Oliva and Dixon 1991). To account for the seemingly random distribution of protamines in fish species, he proposed that horizontal evolution of protamines had instead taken place via the uptake of virally encoded repeating sequences. While the tendency for protamine genes to evolve rapidly is well known (Wyckoff et al. 2000), the evolutionary lineage of the protamines in fish was later elucidated (Saperas et al. 1994). The critical issue with the horizontal theory of protamine evolution, however, is not the apparent randomness of the distribution of protamines throughout the animal kingdom. It is the apparent instantaneous conversion of a PL protein with 25% arginine and 25% lysine to a protamine with 60% arginine and little or no lysine.

THESIS OBJECTIVES

The main objective of my thesis was to investigate the opposing theories of protamine evolution by characterizing the SNBPs and their genes from a diverse range of organisms which employ either histones, protamines, or protamine-like proteins to achieve sperm chromatin compaction. The main questions regarding the plausibility of the vertical evolution of SNBPs revolve around a few simple questions:

1. Many organisms are closely related phylogenetically yet contain very different SNBP numbers and sizes. By what mechanism does this rapid change occur?

2. What are the differences at the genetic level of the different types of sperm nuclear basic proteins? Extensive characterization of the vertebrate protamine genes has provided good insight into the regulation of these proteins, but with limited scope.

3. All of these SNBPs fulfill the common function of sperm chromatin condensation. How can PL proteins of such variability in size and number adequately perform this function in such a structurally indistinguishable way?

4. How have somatic linker histones that are by definition lysine-rich, evolved so rapidly into the highly arginine-rich protamines of the sperm?

Thesis organization

Each chapter that follows is a separate manuscript representing work that has been either published in a refereed journal or has been submitted, and each addresses the answers to one or more of the above questions.

This thesis is organized in the following way:

Section A contains two chapters and is an overview of protamines and histone H1. Each is an inclusive overview of the subject matter.

> **Chapter 2** is a comprehensive review of ALL of the histone H1 and H1-like proteins examined up until 2001, combining all of the sequence and compositional data to trace both the origin of the lysine-rich DNA-binding component of histone H1, and the inclusion of the conserved globular winged helix that is characteristic of metazoan H1s.
>
> **Chapter 3** is a review of protamines from vertebrates and invertebrates, with extensive discussion of protamine composition, regulation, expression, modifications and evolution.

Section B covers much of my experimental work with the PL type SNBPs.

Chapter 4 is the genetic characterization of the large, H1-like PL-I protein of the surf clam, *Spisula solidissima*. A mechanism for the rapid expansion of a sperm-specific histone H1 is described, as well as an initial characterization of the promoter and UTRs of the gene encoding this SNBP.

Chapter 5 is the characterization of the genes from *Mytilus californianus* that encode the PL-II, PL-III and PL-IV sperm nuclear basic proteins. These sequences are compared to those of *Spisula*'s PL-I SNBP, with the main conclusion that the arginine-rich PL-III gene has segregated from the histone H1-like PL-II and PL-IV genes.

Chapter 6 is a review and hypothesis concerning the question of SNBP variability and their ability to condense sperm DNA with comparable efficiency. A novel chromatin structure is proposed based on information compiled from a range of experimental sources.

Section C covers my experimental work with protamines in invertebrates.

Chapter 7 is the isolation and characterization of the protamine gene from squid, representing the first characterization of a complete protamine gene. The

insights provided into the regulation and evolutionary origin of protamines are discussed.

Chapter 8 includes the isolation and genetic characterization of a novel H1 with a highly arginine-rich protamine tail in the sperm of the primitive chordate, *Styela montereyensis*. The real breakthrough came when we compared our DNA sequence with that of the closely related tunicate, *Ciona intestinalis*, which has a sperm-specific H1 with a lysine-rich tail. Examination revealed that the wholesale conversion of lysine to arginine had occurred as the result of a frameshift mutation and extreme codon bias. This finding provides the first solid evidence for a direct evolutionary relationship between the lysine-rich histone H1 of somatic cells and the arginine-rich protamines of sperm.

Origin of H1 Linker Histones*

Harold E. Kasinsky^§†, John D. Lewis^§, Joel B. Dacks‡ and Juan Ausió§¶

§ Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6 and †Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, B.C., Canada, V6T 1Z4

‡ Program in Evolutionary Biology, Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, N.S., Canada, B3H 4H7

¶ To whom all correspondence should be addressed. Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6. Phone: 250-721 8863, fax: 250-721 8855, e-mail: jausio@uvic.ca

^ These authors have contributed equally to this work.

* This article is dedicated to Professor R. David Cole.

Kasinsky, Lewis et al. (2001) FASEB J. **15**:34-42.

ABSTRACT

In which taxa did H1 linker histones appear in the course of evolution?

Detailed comparative analysis of the histone H1 and histone H1-related sequences available to date suggests that the origin of histone H1 can be traced to bacteria. The data also reveal that the sequence corresponding to the "winged helix" motif of the globular structural domain, a domain that is characteristic of all metazoan histone H1 molecules, is evolutionarily conserved and appears separately in several divergent lines of protists. Some protists, however, appear to have only a lysine-rich basic protein which has compositional similarity to some of the histone H1-like proteins from eubacteria and to the carboxy-terminal domain of the H1 linker histones from animals and plants. No lysine-rich basic proteins have been described in archaebacteria. The data presented in this review provide the surprising conclusion that while DNA-condensing H1-related histones may have arisen early in evolution in eubacteria, the appearance of the sequence motif corresponding to the globular domain of metazoan H1s occurred much later in the protists, after and independently of the appearance of the chromosomal core histones in archaebacteria.

Key Words: Histone H1, evolution, protists, bacteria

Kasinsky, Lewis et al. (2001) FASEB J. **15**:34-42.

INTRODUCTION

Recent crystallographic analysis of histones has provided a detailed structural characterization of the histone fold of the core histones (Arents et al. 1991) and the globular winged helix domain of the linker histones (Ramakrishnan et al. 1993). While the latter is ubiquitous among animals, plants and fungi, it is absent in some protist taxa. Both the pattern of distribution of the H1 winged helix and an examination of the remaining C-terminal region should provide insights into the evolution of this protein family.

The characterization of the histone fold (Arents et al. 1991) has shed an important insight on the evolution of core histones, whose origin can be traced to archaebacteria (Arents and Moudrianakis 1995). However, the origin of the linker histones has not been established. In what has already become a classic work (van Holde 1989) for researchers in the chromatin field, van Holde declared:

> *"The relationship of H1 to other histones is obscure. So far as we can tell the H1 sequences seem unrelated to either other histone sequences or those of prokaryotic proteins. This may of course, simply be a consequence of the rapid evolution of this protein, which has obscured its origins: alternatively, H1 may have evolved from an entirely different protein".*

In this review, we examine several important questions of H1 linker histone evolution: Can the origin of this family of linker histones be traced back to prokaryotes? If so, have H1 linker histones evolved from the same or entirely different genes than the core histones?

For this purpose, we survey the recent literature on histone H1 and H1-like protein and gene sequences in protists and bacteria and analyze their similarity to that of the sequence of their animal, plant and fungal counterparts.

Kasinsky, Lewis et al. (2001) FASEB J. **15**:34-42.

CORE HISTONES, LINKER HISTONES AND CHROMATIN.

In the eukaryotic cell, DNA exists as a nucleoprotein complex known as chromatin (van Holde 1989). Histones are the major protein component of chromatin and can be structurally grouped in two major categories: "core" and "linker" histones. Core histones (histones H2A, H2B, H3 and H4) are arranged as a globular octameric core in

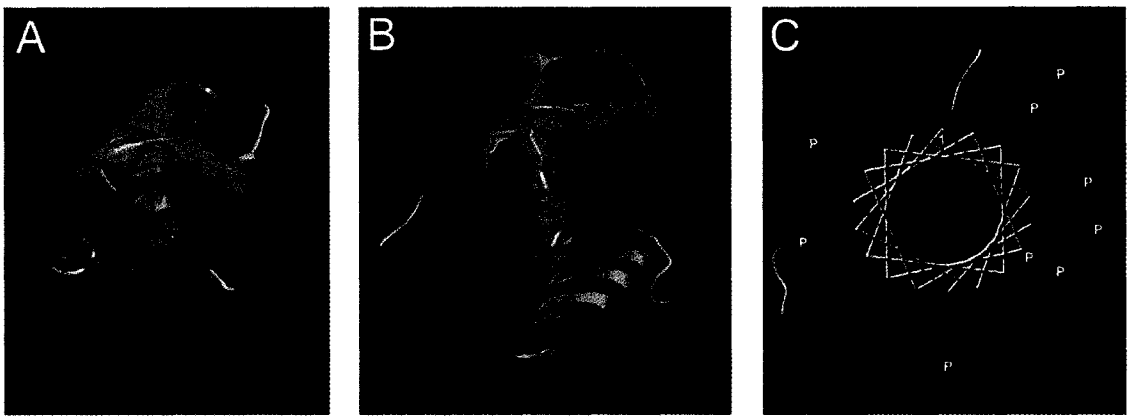


Figure 1. Comparison of the structure of the winged helix motif of histone H1 and the conservative domain of core histones. **A:** Histone fold for core histones H2A, H2B, H3 and H4 (van Holde 1989) **B:** Linker H1 histone winged helix motif (Ramakrishnan et al. 1993). **C:** Helical wheel representation of the putative helical requirements of the C-terminal domain of histone H1 from the sea urchin *Strongylocentrotus purpuratus*. Note the sequential distribution of proline (P) residues which would introduce kinks along the helix. N=amino terminus; C=carboxy terminus. The α-helices are denoted in cyan; β-sheet in purple.

which an H3-H4 tetramer serves as scaffold to two adjacent H2A-H2B dimers (Eickbush and Moudrianakis 1978). Between 146-180 bp of DNA are wrapped around this protein core in approximately two left-handed superhelical turns. The nucleosome structures resulting from such association (Luger et al. 1997) are connected by a variable stretch of linker DNA.

Each of the core histones has a histone fold domain (Arents et al. 1991) (see Fig. 1A) which extends into less structured amino and carboxy terminal domains commonly

referred to as "tails". The N-terminal tail of core histones has a highly basic amino acid composition and together with the linker histones play an important role in chromatin folding. Core histones are amongst the most highly evolutionarily conserved proteins (Isenberg 1978) and are present in all eukaryotic cells. They are thought to have evolved from a DNA-binding protein such as Hmf found in the thermophilic archaeon *Methanofermus fervidus* (Baxevanis et al. 1995). Such DNA binding proteins consist of the histone fold but lack the C- and N-terminal tails found in eukaryotic organisms. They are present in the euryarchae, a major kingdom of archaebacteria, but are absent from the one crenarchaeal genome sequenced thus far (Faguy and Doolittle 1999; Kawarabayasi et al. 1999).

Histones of the H1 family interact extensively with linker DNA and hence are known as linker histones. Upon binding of histone H1 to the linker DNA, the polynucleosomal fiber folds into a 30 nm chromatin fiber (van Holde 1989). The linker histones of multicelled eukaryotes exhibit a tripartite structural organization in which a globular domain is flanked by two less structured basic amino and carboxy terminal domains. The crystallographic structure of the globular domain has been determined and shown to consist of a winged helix motif (Ramakrishnan et al. 1993) (see Fig. 1 B). This domain interacts with the nucleosome at a region close to the pseudodyad axis of symmetry (Zhou et al. 1998). In contrast to core histones, linker histones are less evolutionarily conserved (Isenberg 1978; Cole 1984). While the sequence of the winged helix motif is relatively well conserved through evolution in animals, plants and fungi (see Fig. 2), the N- and C-terminal domains are extremely heterogeneous, both in their length and in amino acid composition. The histone H1 family in metazoans and other

multicelled eukaryotes is a heterogeneous family of developmentally regulated histones (Cole 1984) that includes highly tissue-specific proteins such as histone H5 from the nucleated erythrocytes of birds (Neelin et al. 1964) and sperm PL-I proteins (Ausió 1999). Henceforth, "H1" will represent the entire histone H1 family.

THE LYSINE-RICH C-TERMINAL DOMAIN OF H1: A CRITICAL STRUCTURE FOR LINKER HISTONE FUNCTION

The first eukaryotic linker histones that were purified and characterized all had a

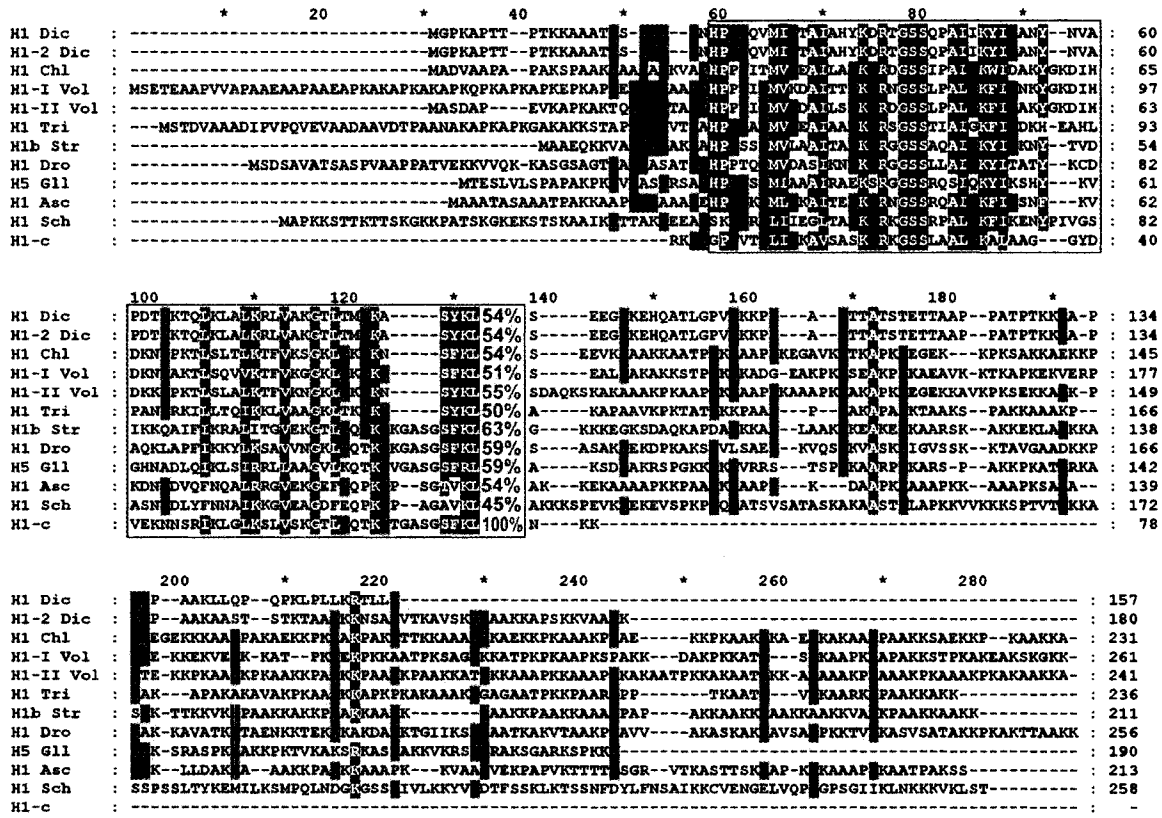


Figure 2. Sequence alignment of encoded H1 linker histones in protists, animals, a plant and a fungus, generated with Clustal X (Thompson et al. 1997). Shading indicates the range from completely identical amino acid residues in the same position in all sequences (purple), to similar residues at a particular position (light purple, more similar; blue, less similar). The sequence of the winged helix motif is demarcated by a red box. Percentiles indicate the extent of similarity to H1-c, the histone H1 core consensus sequence (Wells and Brown 1991). See the legend of Table 1 for the species nomenclature.

high lysine composition compared to core histones and hence were called lysine-rich histones (Johns 1971; Cole 1984). Early sequence analysis showed that the lysine-rich nature of the linker histone was mainly due to the frequent occurrence of this amino acid in the C-terminal domain of these proteins (Cole 1984). The alternating occurrence of lysine (K) and alanine (A) residues (two highly helicogenic amino acids) in this region and the resulting charge distribution (Subirana) have been postulated to result in a proline-kinked AK α-helix organization (Churchill and Travers 1991) that we will refer to as the AKP helix (see Figure 1C). In many instances, these putative α-helical domains exhibit a clear amphipathic nature (Subirana) that may play a role in linker histone-linker histone interactions in the chromatin fiber or in the inter-chromatin fiber association mediated by these histones. It is this particular distribution of AKP in the C-terminus that confers to histone H1 the unique ability to bind to the linker DNA (Subirana), and its presence is essential for the processes of chromatin folding and condensation. As has already been mentioned, the major function of histone H1 is to condense the linker DNA to induce folding of the polynucleosome fiber into chromatin structures of approximately 30 nm in diameter (van Holde 1989). These can eventually condense into larger superstructures (chromosomes) during mitosis. Although a polynucleosome fiber lacking linker histones is able to fold to a certain extent (Hansen and Ausió 1992), additional folding into the 30 nm fiber, under physiological conditions, can only occur upon binding of histone H1 to the linker DNA. Chromatin reconstitution experiments carried out with histone H1 fragments consisting of the globular and C-terminal domain have shown that these fragments are able to fold the chromatin fiber as effectively as the intact native H1 molecule (Allan et al. 1986). In contrast, the globular histone H1 domain alone is unable

to condense the chromatin fiber to any similar extent (Allan et al. 1980). Thus, of the three structural domains of the linker histones, the C-terminal domain appears to be critical for chromatin folding (Allan et al. 1986).

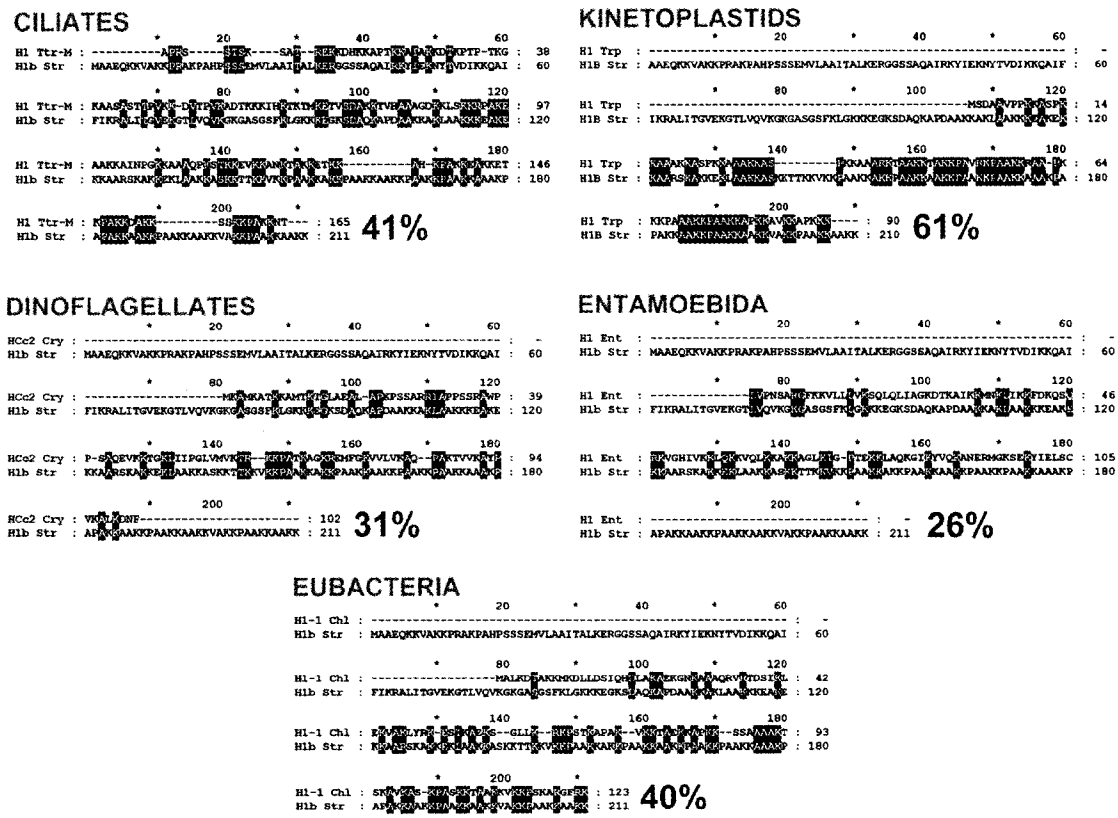


Figure 3. Pairwise comparison of encoded H1 histone and H1 histone-like sequences lacking the winged helix motif in selected protists and bacteria with linker histone H1b of *S. purpuratus*. Purple shading indicates identity. Ciliates: *Tetrahymena thermophila* macronuclear histone H1; Kinetoplastids: *Trypanosoma brucei* histone H1 (M1 genomic DNA clone); Dinoflagellate: *Crypthecodinium cohnii* HCc2. Entamoebidae: *Entamoeba histolytica* histone H1. Bacteria: *Chlamydia pneumonia* histone H1-1. Alignments created using Clustal X (Thompson et al. 1997).

H1 LINKER HISTONES IN SOME PROTISTS LACK THE WINGED HELIX MOTIF

While a great deal of work has been done on histone H1 in animals, other eukaryotic taxa have been largely ignored relative to the question of H1 origin.

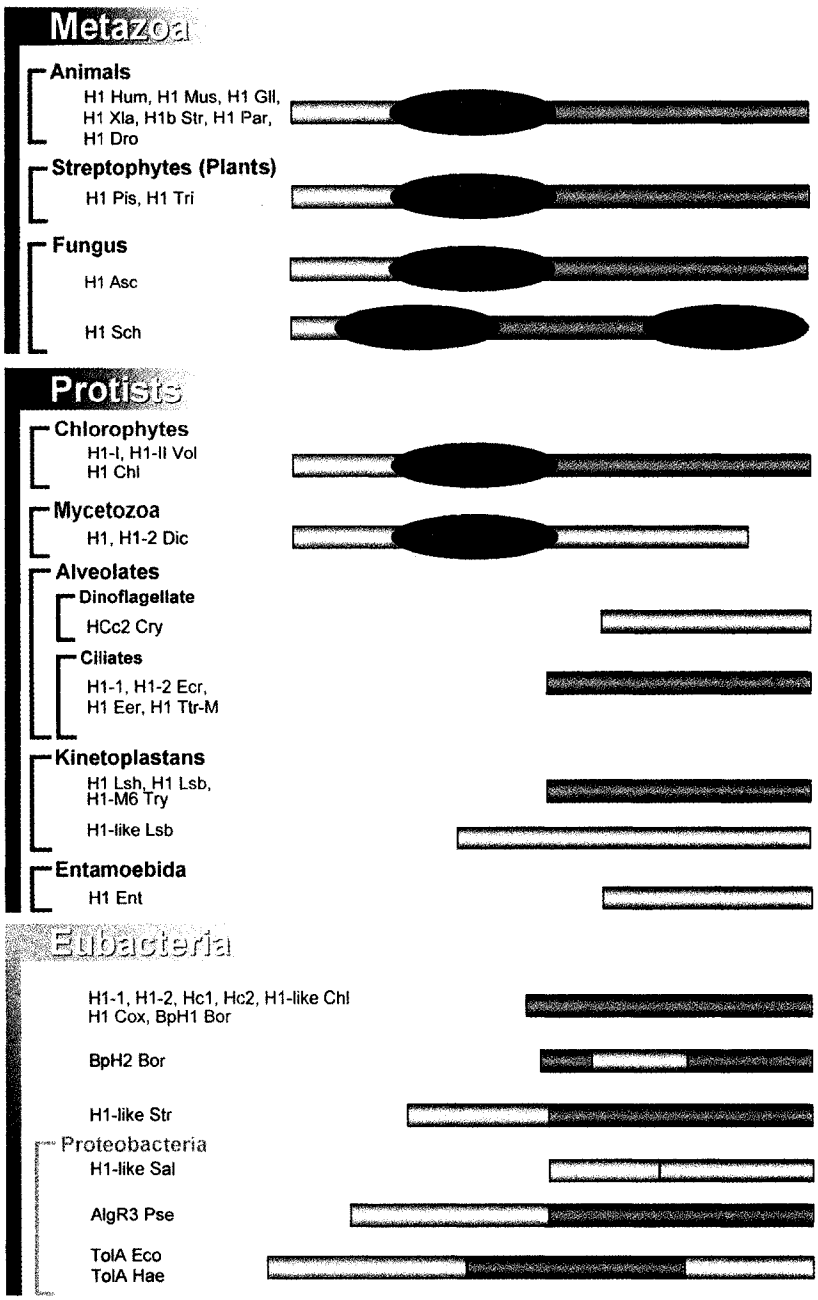


Figure 4. Schematic diagram of the evolution of the winged helix motif in H1 linker histones. The green oval denotes the winged helix motif and the dark purple rods the lysine-rich carboxyl-terminus of linker histones similar to histone H1b in the sea urchin *Strongylocentrotus purpuratus*. Lighter shades of purple indicate sequences with decreasing similarity to the carboxy-terminus tail of *S. purpuratus* histone H1b. Yellow stands for the amino-termini as well as other sequences that are not similar to either the carboxy-terminus or globular core of *S. purpuratus* histone H1b. See the legend of Table 1 for a description of the species nomenclature.

Nonetheless, H1 homologues have been characterized from a surprisingly varied taxon diversity, including plants, animals, fungi and a wide variety of protozoans.

Euglenozoan protists, such as the kinetoplastids *Trypanosoma cruzi* (Toro and Galanti 1988) *and T. brucei* (Burri et al. 1993), possess linker histones that lack the winged helix motif. These are small proteins that are compositionally

and structurally very similar to the C-termini of histone H1 in animals, plants, chlorophytes and mycetozoans (see Table I and Fig. 3) and bind to the linker DNA of the nucleosomally organized chromatin of these organisms (Burri et al. 1993). In addition to trypanosomes, a gene encoding a protein with a similar amino acid composition is present in another kinetoplastid, *Leishmania major* (see Fig. 4 and Table I). A similar protein has been purified from *Euglena gracilis* (Jardine and Leaver 1978), also from the phylum Euglenozoa. (see Table I and Fig. 4). However, not all kinetoplastid H1 proteins match the consensus C-terminal sequence so well. A protein has been isolated by perchloric acid extraction (a method initially devised by Johns (Johns 1971) to selectively fractionate histone H1 from core histones) and the gene identified for a H1 homologue in the insect trypanosomatid *Crithidia fasciculata.* Although related to histone H1 (Duschak and Cazzulo 1990), the protein has an amino acid composition that significantly departs from the consensus amino acid composition of the histone H1 C-terminus and bears very low similarity to the linker histone consensus sequence of the winged helix.

Similarly, proteins related to the histone H1 C-terminus both in amino acid composition (Table I) and in sequence (Fig. 3) can be found (see Fig. 4, 5) in the protist phylum Alveolata (Hausmann and Hülsmann 1996). Examples of this are the encoded histone H1 gene of the oligohymenophoran ciliate *Tetrahymena thermophila* (Hayashi et al. 1987), the histones of the hypotrich ciliate *Oxytricha* sp. (Caplan 1975) and the encoded histone H1-1 gene from the hypotrich ciliate *Euplotes eurystomus* (see Fig. 4 and Table I). The *Tetrahymena* gene is expressed in macronuclei, where the H1 linker histone has been characterized by gel electrophoresis (Wu et al. 1994). Within the alveolates, a lysine-rich basic protein, HCc2, from the dinoflagellate *Crypthecodinium*

TABLE I. Composition (mol%) of abundant amino acid residues in H1 linker histones^

	Carboxyl-terminal region					Whole protein				Total
	K%	A%	P%	aa's	Length	K%	A%	P%	Accession	length
Animals, plants, and fungi										
H1 Hum	37.5	19.2	10.6	111-214	104	26.6	19.6	8.9	X57130	214
H1 Mus	39.8	20.4	11.7	110-212	103	27.4	19.8	8.5	S43949	212
H1 Gll	41.3	32.1	11.9	110-218	109	28.9	26.2	9.2	P09987	218
H1 Xla	38.8	30.6	11.2	114-228	98	28.1	25.9	9.7	S69089	228
H1 Str	43.8	33.9	7.4	90-210	121	32.4	25.7	5.7	P15869	210
H1 Par	42.9	37.8	5.1	109-206	98	25.7	25.7	7.3	S09388	206
H1 Dro	34.5	23.0	5.0	117-255	139	26.7	19.6	5.5	P02255	255
H1 Pis	34.6	23.3	10.5	32-264	133	26.5	17.1	9.9	P08283	264
H1 Tri	34.0	37.0	16.0	124-223	100	23.3	28.3	11.7	P27806	223
H1 Sch	32.2	18.6	8.5	114-172	59	22.5	9.7	5.8	P53551	258
H1 Asc	30.8	34.2	12.5	94-213	120	23.5	26.8	10.3	AAF16011	213
H1 Asp	29.7	20.7	10.8	90-200	111	23.0	17.0	9.5	CAB72936.1	200
H1 Ncr	32.4	31.6	9.6	101-236	136	23.3	26.3	8.9		236
Average	36.3	27.9	10.1		110.1	26.0	22.1	8.5		225.9
Algae and protists										
H1-I Vol	41.5	20.0	13.1	131-260	130	31.2	18.9	12.7	Q08864	260
H1-II Vol	40.3	37.5	13.9	98-241	144	32.4	26.6	11.6	Q08865	241
H1 Chd	43.0	28.2	11.1	97-231	135	33.3	22.9	9.5	S59589	231
H1 Dic	22.6	18.9	17.0	105-157	53	17.8	14.7	11.5	AAA93483	157
H1-2 Dic	29.0	29.0	10.5	105-180	76	21.1	19.4	9.4	P54671	180
H1 Phy	17.6	18.1	16.2			17.6	18.1	16.2		
H1 Chr	26.7	17.8	8.7			26.7	17.8	8.7		
Average	34.6	26.6	11.1		109.4	25.9	21.3	9.5		222.6
Protists										
HCc2 Cry	19.6	15.7	9.8	1-102	102	19.6	15.7	9.8	B56581	102
H1-1 Ecr	32.9	19.1	6.6	1-152	152	32.9	19.1	6.6	AAD32600	152
H1-2 Ecr	29.8	12.3	5.3	1-171	171	29.8	12.3	5.3	AAD32601	171
H1 Eer	25.9	28.2	5.2	1-135	135	25.9	28.2	5.2	S34952	135
H1 Ttr-M	33.5	15.9	7.3	1-164	164	33.5	15.9	7.3	A26490	164
H1-like Lsb	15.1	12.5	6.3	1-192	192	15.1	12.5	6.3	AAD26571	192
H1 Lsb	31.3	36.6	8.9	1-112	112	31.3	36.6	8.9	AAD26570	112
H1 Lsh	35.2	20.0	4.8	1-105	105	35.2	20.0	4.8	CAA11592	105
H1-M6 Try	37.8	33.3	13.3	1-90	90	37.8	33.3	13.3	P40274	90
H1 Myc	33.0	35.0	7.8	112-214	103	19.2	24.8	6.5	P95109	214
H1 Ent	26.7	6.7	1.9	1-105	105	26.7	6.7	1.9	BAA21981	105
H1 Cri	17.2	14.1	7.0	1-128	128	17.2	14.1	7.0	2206467C	128
H1 Oxy	31.6	29.2	5.1			31.6	29.2	5.1		
H1 Oli	19.5	16.3	5.4			19.5	16.3	5.4		
H1 Eug	35.0	22.6	10.1			35.0	22.6	10.1		
Bacteria										
Hc1 Chd	28.8	18.4	2.4	1-125	125	28.8	18.4	2.4	A39396	125
H1-1 Chd	28.5	18.7	4.1	1-123	123	28.5	18.7	4.1	AAD19024	123
H1-like Chd	25.6	19.7	4.3	1-117	117	25.6	19.7	4.3	JH0658	117
Hc2 Chd	27.4	24.9	4.0	1-201	201	25.1	23.3	4.0	A36884	223
H1-2 Chd	30.8	19.2	4.2	1-120	120	25.0	17.4	2.9	AAD18528	172
H1 Cox	24.8	18.8	2.6	1-117	117	24.8	18.8	2.6	AAB36614	117
BpH1 Bor	37.3	37.3	7.6	1-158	158	32.4	34.6	8.8	S61926	182
BpH2 Bor	25.6	25.6	12.2	1-41,105-145	82	20.0	20.0	8.3	JC6029	145
H1-like Str	31.5	29.1	5.5	92-218	127	21.1	22.9	4.6	CAA20004	218
H1-like Sal	15.5	15.5	3.5	80-137	58	8.8	12.4	2.2	AAB61148	137
AlgR3 Pse	18.6	48.2	15.5	121-340	220	17.1	35.9	10.6	A35630	340
TolA Eco	23.0	50.3	0.0	104-294	191	15.7	30.9	2.4	P19934	421
TolA Hae	22.5	43.4	0.8	121-249	129	13.4	20.2	2.9	AAC44596	382

^Hum: human, Mus: *Mus musculus* (mouse), Gll: *Gallus gallus* (chicken), Xla: *Xenopus laevis* (frog), Str: *Strongylocentrotus purpuratus* (urchin), Par: *Parechinus angulosus* (urchin), Dro: *Drosophila melanogaster* (fruit fly), Pis: *Pisum savitum* (pea), Tri: *Triticum aestivum* (wheat), Sch: *Saccharomyces cerevisiae* (yeast), Asc: *Ascobolus immersus* (fungi), Asp: *Aspergillus nidulans* (fungus), Ncr: *Neurospora crassa* (fungus), Vol: *Volvox carteri*, Chd: *Chlamydomonas reinhardtii*, Dic: *Dictyostelium discoidium*, Phy: *Physarum polycephalum*, Chr: *Chlorella ellipsoidea*, Cry: *Crypthecodinium cohnii*, Ecr: *Euplotes crassus*, Eer: *Euplotes eurostomas*, Ttr-M: *Tetrahymena thermophila* (macronuclear), Lsb: *Leishmania brasiliensis*, Lsh: *Leishmania major*, Try: *Trypanosoma brucei*, Myc: *Mycobacterium tuberculosis*, Ent: *Entamoeba histolytica*, Cri: *Crithidia fasciculata*, Oxy: *Oxytricha sp.*, Oli: *Olisthodiscus luteus*, Eug: *Euglena gracilis*, Chd: *Chlamydia trachomatis*, Cox: *Coxiella burnetii*, Bor: *Bordetella pertussis*, Str: *Streptomyces coelicolor*, Sal: *Salmonella typhimurium*, Pse: *Pseudomonas aeruginosa*, Eco: *Escherichia coli*, Hae: *Haemophilus influenza*. K = lysine; A = alanine; P = proline. aa's = amino acid residues. References for protists lacking accession numbers are as follows: H1 Phy (Mende et al. 1983), H1 Oxy (Caplan 1975), H1 Oli (Rizzo et al. 1985), H1 Eug (Jardine and Leaver 1978), H1 Chr (Iwai 1964).

cohnii, has also been identified (Vernet et al. 1990; Sala Rovira et al. 1991). However, in this instance, the extent of sequence similarity with the C-terminus of histone H1 is lower than in ciliates (see Fig. 3). This is not surprising, as dinoflagellates are known for their unusual nuclear organization (van Holde 1989).

In the subphylum Heterokonta (Hausmann and Hülsmann 1996), a histone H1-related protein has also been purified from the unicellular golden alga *Olisthodiscus luteus* (Rizzo et al. 1985). Although the sequence of this protein is unknown, the similarity of its amino acid composition to that of *Crypthecodinium* (see Table I) suggests that their C-terminal domains may also be similar.

Finally, the encoded H1 histone for *Entamoeba histolytica* (see Fig. 4 and Table I) lacks the winged helix motif entirely but shows considerable similarity (Fig. 3) to the C-terminal tail of histone H1. Molecular sequence data indicates a deeply diverging position for *Entamoeba* (Fig. 5), although this has been brought into question (Stiller and Hall 1999). It thus may represent one of the most primitive protists for which a histone H1-related protein has been characterized.

EVOLUTIONARY APPEARANCE OF THE WINGED HELIX MOTIF IN PROTISTS

Proteins that contain the winged helix motif comprise a family of DNA-binding proteins (mainly transcription factors) whose DNA recognition helices are related in structure and function to the helix-turn-helix motif, despite a relative absence of significant primary sequence identity.

Figure 4 summarizes the results from the comparative analysis of the sequences from histone H1 and histone H1-related proteins using the alignment analysis shown in Figures 2 and 3, as well as the information from Table I. As illustrated in Figure 4, genes coding for H1 linker histones with an evolutionarily conserved winged helix motif can be found in diverse protist groups: chlorophytes and mycetazoans (Fig. 5). Figure 2 indicates that the putative protein products of encoded genes H1-I and H1-II from the multicellular green alga *Volvox carteri* and an H1 gene from the unicellular green alga *Chlamydomonas reinhardtii* both display a sequence alignment typical of the winged helix motif of linker histone H1 from plants and animals (see also Table I). Of these, expression of the linker histone genes has been demonstrated only in *C. reinhardtii* (Morris et al.). H1 histone has been purified by chromatography from the unicellular green alga *Chlorella ellipsoidea* (Iwai 1964). The composition in mol %, of its three most abundant amino acyl residues Lys, Ala and Pro (KAP) is similar to that of the putative protein of the H1-I gene for *V. carteri* (see Table I). In the yeast *Saccharomyces cerevisiae,* there is an H1 gene of 258 amino acyl residues encoded in the genome (Landsman 1996) that is expressed as a poly-A^+ RNA and whose gene product is localized to the nucleus (Ushinsky et al. 1997). Although there is biochemical *in vitro* evidence to suggest that this protein indeed behaves as a canonical linker histone (Patterton et al. 1998) evidence for this role *in vivo* is still lacking. The sequence data suggests that it may also contain a second winged-helix motif at its C-terminal end. On the contrary, the histone H1s from the multicellular fungi *Neurospora crassa* (Folco et al. 2003), *Ascolobus immersus* (Barra et al. 2000) and *Aspergillus nidulans* (Ramon et al. 2000) contain the tripartite structural organization that is found in all multicelled

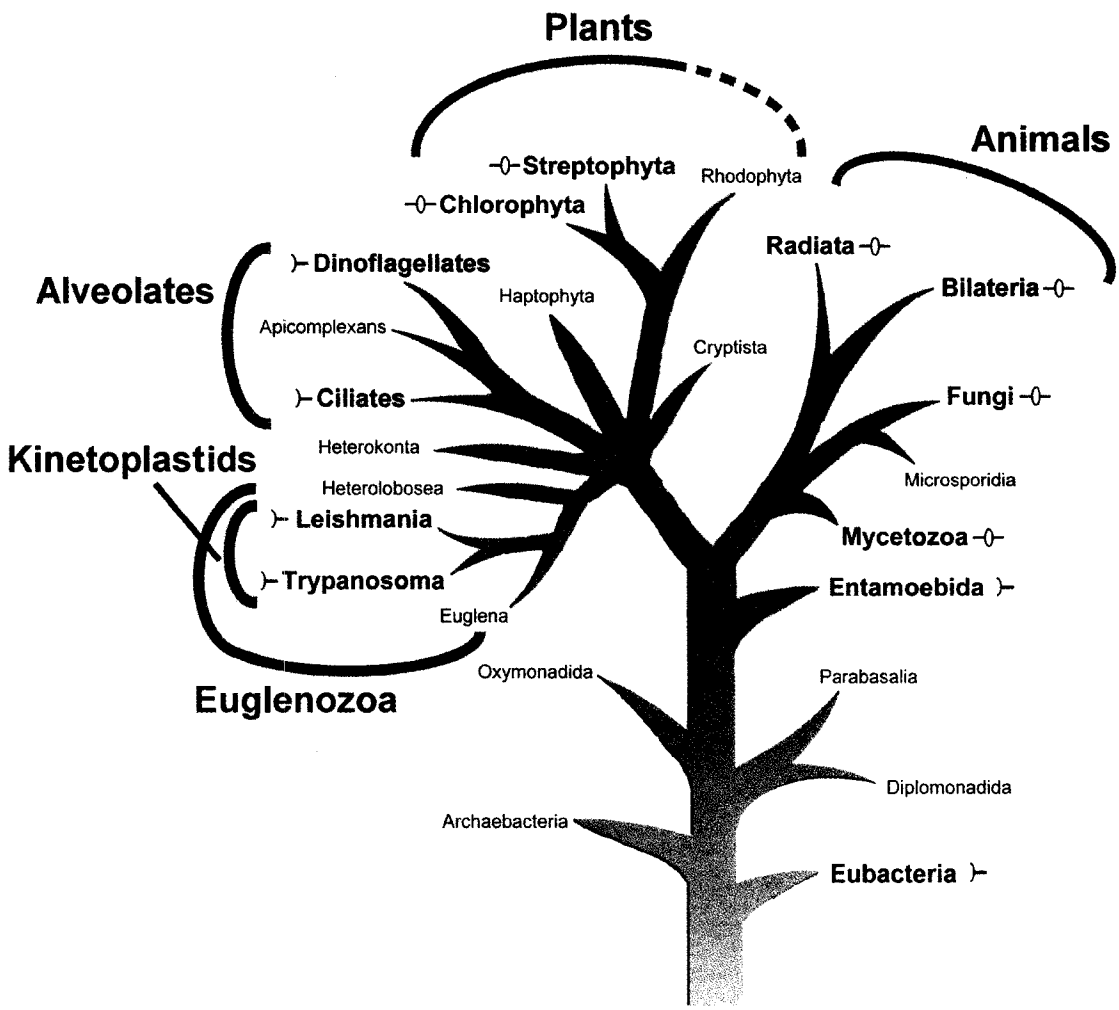


Figure 5. Distribution of encoded H1 linker histones in protists superimposed on a phylogenetic tree of eukaryotes and prokaryotes, adapted from Dacks and Kasinsky (Dacks and Kasinsky 1999; Dacks and Roger 1999). –o– = linker histone H1 with winged helix motif;)– = C-terminus of linker histone H1.

eukaryotes (see Fig. 2, Fig. 4 and Table I).

Genes encoding linker histones H1 and H1-2 are also present in a mycetozoan, the cellular slime mold *Dictyostelium discoideum* (Fig. 2), where the proteins have also been purified by chromatography (Charlesworth and Parish 1977), but not sequenced. In addition, a histone H1 was chromatographically purified from plasmodia of the acellular slime mold *Physarum polycephalum* (Mende et al. 1983), another mycetozoan. These proteins have the lowest carboxy-terminal (KAP) amino acid composition of all the winged helix containing H1s (see Table I). Yet, they contain a sequence with extensive

similarity to the consensus sequence of the winged helix motif of H1 histones in multicelled eukaryotes (see Fig. 2). These findings indicate, therefore, that H1 linker histones with a winged helix motif appeared separately in at least two disparate lines of eukaryotes (Fig. 5), possibly as the result of two separate fusion events between the C-terminal domain of H1 and the proto-winged helix domain protein..

HISTONE H1-RELATED PROTEINS IN EUBACTERIA AND THE C-TERMINI OF METAZOAN H1 HISTONES

By looking across eukaryotic diversity we can see that the ancestral eukaryote most likely possessed a lysine-rich proto-linker histone. But where did this gene arise? If core histones evolved from archaeal histones, did the H1 linker histones do so as well? Not likely. To date there is no evidence for the presence of H1 histone related genes in any of the archaea, including those whose genomes have been completely sequenced (Bult et al. 1996; Klenk et al. 1997; Smith et al. 1997; Kawarabayasi et al. 1998):R. Heilig, unpublished results).

Interestingly, the encoded genes for basic proteins in several eubacteria show more similarity to the C-terminus of a typical metazoan histone H1 than do the encoded H1 histone genes of some of the alveolates and *Entamoeba* described in the preceding section (Fig. 3, Fig. 4 and Table I). The amino acid compositional similarity of the histone H1-related proteins in *Chlamydia pneumonia* and *C. trachomatis* (Hackstadt et al. 1991) to that of the C-terminus of histone H1 is striking (see Table I). Furthermore, their sequence similarity is in some instances even higher than in their protist counterparts (see

Fig. 3). Other examples of this kind include *Coxiella burnetii*, *Streptomyces coelicolor* and *Bordetella pertussis* (see Fig. 4 and Table 1).

In the proteobacteria (a subgroup of eubacteria), similar results can be seen (Fig. 4) for *Salmonella typhimurium, Pseudomonas aeruginosa, Escherichia coli* and *Haemophilus influenzae*. However, the lysine content of these proteins is lower than expected and, in general, they have a higher alanine content (see Table I).

It is therefore quite likely that the lysine-rich DNA binding proteins found in eubacteria are evolutionarily related to the histone H1 truncated versions also found in protists (see above) and to the C-terminus of histone H1 in metazoans and other multicelled eukaryotes. The occurrence of histone H1-related proteins in eubacteria stands in contrast to archaebacteria, which have core histones but not H1-like histones in any of the five genomes that have been sequenced from the euryarchaeal kingdom (Bult et al. 1996; Klenk et al. 1997; Smith et al. 1997; Kawarabayasi et al. 1998):R. Heilig, unpublished results).

OVERVIEW

The main function of eukaryotic histone H1 is in the condensation of chromatin and/or block access to nucleosomal DNA. This is mainly achieved by screening of the negative charges on the linker DNA connecting adjacent nucleosomes. Mechanistically, such charge neutralization could possibly occur in many different ways and thus it should not be surprising that lysine-rich "linker histones" of the H1 family are less evolutionarily conserved and much more variable in size and sequence than the "core histones". However, there are some compositional constraints to this variability. Although arginine-

rich proteins such as protamines from sperm chromatin are also very efficient at condensing DNA, they bind very tightly and decondensation (during fertilization) requires the assistance of highly specific proteins from the egg, such as nucleoplasmin. In contrast, the lysine-rich nature of linker histones allows them to be associated dynamically with the DNA (chromatin) substrate to provide the compaction required by the changing physiological needs of the cell during the different stages of the cell cycle. Therefore it comes as no surprise that the first DNA condensing proteins to be found in eubacteria, such as in *Chlamydia*, are lysine-rich.

In the linker histones of multicelled eukaryotes, the lysine-rich nature of these proteins is the result of the high lysine content of their C-terminal domains. As mentioned earlier, the C-termini of these molecules play a critical role in their function of chromatin condensation. Because of these very particular yet rather simple compositional constraints of such domains, it should in principle be possible to identify structurally and functionally related proteins of this family by simply examining their amino acid composition (see Table I). In our comparative analysis we have used histone H1 from the sea urchin *S. purpuratus* as a reference because it exhibits a high degree of sequence similarity to the consensus sequence of the globular domain, while containing a good representation of the KAP repeats that are characteristic of the C-termini of H1s in multicelled eukaryotes (Subirana 1990; Churchill and Travers 1991).

From the analysis obtained with these tools (Table I, Figs. 2-4), it is possible to trace the evolutionary origin of histone H1 to the DNA-condensing, lysine-rich proteins of eubacteria (see Fig. 5). In fact, the extent of sequence similarity of *Chlamydia* H1-related proteins to the C-terminal domain of metazoan H1 histones is striking,

considering the enormous overall variability of this histone family.

One interesting question to address is the following: if core histones arose in archaea and lysine-rich basic proteins arose in bacteria, then how did they come together in eukaryotes? This might have occurred as a result of a lateral gene transfer (LGT) event in the early proto-eukaryote (thought to resemble an archaeal ancestor), as LGT has recently come to light as a major evolutionary force in both bacteria and archaea (Doolittle 1999). The conjunction of the ancestral histone proteins in eukaryotes could also have arisen from the large-scale transfer of genes that accompanied the first endosymbiosis of the alpha proteobacteria that gave rise to the mitochondria.

The ancestor of eukaryotes might already have had archaeal histones precursors able to incorporate the C- and N- terminal domains present in the core histones of all eukaryotic organisms. These proteins may have led to the appearance of the nucleosome organization and hence representing the origin of eukaryotic chromatin. In a broad sense, the core histones provide the structural substrate upon which eukaryotic gene regulation takes place. The incorporation of the bacterial H1-like histone precursors might either have been a by-product of, or the step allowing for, the large scale expansion that generally characterizes eukaryotic genomes. DNA already bound by these proteins may have gained further protection from the incorporation of the lysine-rich bacterial H1 precursors, providing an additional condensation of the "naked" linker DNA regions connecting the nucleosome structures that resulted from the core histone-DNA interactions. It has been shown that the small lysine-rich H1-related protein from *Trypanosoma brucei* contributes both to the spacing of nucleosomes on the DNA and the extensive condensation of the chromatin fiber (Burri et al. 1993), demonstrating that this

function for histone H1 is not restricted to animal and plant taxa.

The acquisition of the globular winged helix occurred later in eukaryotic evolution (see Figs. 4 and 5), possibly to provide specificity for the targeting of the H1 molecules to the linker DNA regions by providing structural recognition of the four-way junction DNA-like structures that are present at the sites of DNA entry and exit in the nucleosome (Zlatanova and van Holde 1998). Furthermore, such an incorporation most likely contributed to enhance chromatin folding into the 30nm fiber (Ramakrishnan 1997), which is present in all multicelled eukaryotes but is absent in *Trypanosoma* (Toro and Galanti 1988). It is also important to notice that we have not found any H1 molecules in multicelled eukaryotes that did not have an N-terminal domain. Thus, it appears that the N-terminal and globular domains have evolved together, although the rate of variation has been higher in the N-terminal region. Despite the fact that the N-terminal domain of linker histones does not seem to be critical for chromatin folding (Allan et al. 1986) and the functional and structural role of this part of the molecule are still obscure, this region may play an important role in the modulation of the binding affinity of the whole histone H1 molecule to chromatin and/or in the head-to-tail interactions of the linker histones that occur in the chromatin fiber (Lennard and Thomas 1985).

If the histone fold of the core histones (see Fig. 1A) provides a structural signature for these proteins such as that provided by the winged helix and C-terminal domains of the linker histones (see Fig. 1B-C), then no genes or proteins have been identified to date in eubacteria that resemble or have any similarity to the core histones. It thus appears that "core" and "linker" histones, despite their common *histone* nomenclature, have evolved quite independently from entirely unrelated genes in archaebacteria and eubacteria,

respectively. In contrast with the evolutionarily conserved core histones (Isenberg 1978), the variability of linker histones and their increasing evolutionary complexity mirrors the developmental variability and complexity of living organisms, starting from eubacteria. Thus, the increase in complexity of the linker histones most likely occurred in response to (and reflects) the increasing functional complexity of the different chromatin domains within the eukaryotic cell and has been developmentally driven. Despite this variability, and the limited sequence information available so far, the information compiled in this review indicates that linker histone evolution had its origin in eubacteria.

ACKNOWLEDGMENTS

We are grateful to Michael Frietag of the Institute of Molecular Biology at the University of Oregon, as well as Eric Selker, D. Folco, and A. Rosa, for allowing us access to the unpublished sequence data for H1 of *Neurospora crassa*. We also thank R. Heilig for permission to cite his unpublished results. We would like to thank W.F. Doolittle for his critical reading of the manuscript. J.B. Dacks would like to thank W.F. Doolittle for financial support and for allowing him the latitude to pursue this project. Funding was provided by NSERC, Canada, to J. Ausió and H.E. Kasinsky. The latter wishes to thank the faculty and staff of Biochemistry and Microbiology at the University of Victoria for their hospitality during his sabbatical year.

A Walk through Vertebrate and Invertebrate Protamines

John D. Lewis, Yue Song, Miriam E. de Jong, Sabira M. Bagha and Juan Ausió*

Department of Biochemistry and Microbiology, University of Victoria,
Victoria, B.C., V8W 3P6, Canada.

*Corresponding author: Department of Biochemistry and Microbiology,
University of Victoria,
P.O. Box 3055, Petch Building Room 220,
Victoria, B.C.
V8W 3P6, Canada

Phone: (250) 721-8863; Fax: (250) 721-8855
email: jausio@uvic.ca

Lewis et al. (2003) Chromosoma **111**:473-82.

" Das Nuclein ist nun aber im Samen nicht frei enthalten, sondern in einer unlöslichen , salzartigen Verbindung mit einer organischen Base, dem ***Protamin.*** *"*

" However, nuclein in the semen is not in free form, but it is an insoluble salt-like complex with an organic base, the ***Protamine.*** *"*

Miescher, F. (1874) "Das Protamin, eine neue organische Base aus den Samenfäden des Rheinlachses" Ber 7, 376-379.

ABSTRACT

An updated comparative analysis of protamines and their corresponding genes is presented, including representative organisms from each one of the vertebrate classes and one invertebrate (squid, *Loligo opalescens*). Special emphasis is placed on the implications for sperm chromatin organization and the evolutionary significance. The review is based on some of the most recent publications in this field and builds upon previously published reviews on this topic.

INTRODUCTION

The term "protamin" (protamine) was first coined by Friederich Miescher (Miescher 1874) almost 130 years ago to refer to an "organic base" that was found associated with "nuclein" in the sperm nuclei of Rhine salmon. However, it was not until about 20 years later that the true protein nature of protamines was established by Albrecht Kossel (Kossel 1896a; Kossel 1896b). In this early work it was already recognized that nuclei from somatic cells and spermatic cells had a different chromosomal protein composition. Whereas somatic cells consisted of histones, sperm cells consisted of protamines. An

exhaustive catalogue of sperm nuclear basic proteins (SNBP) carried out by David Bloch (Bloch 1969) showed that the protein composition of sperm chromatin was very heterogeneous and that protamines were one of several SNBP types which included histones [(Bloch 1969), see also Kasinsky (Kasinsky 1989) for a more recent update].

In 1960, the first comprehensive review on the chemical characterization of protamines was published by Kurt Felix (Felix 1960), which was followed in 1973 by that of Toshio Ando and collaborators (Ando et al. 1973). In both reviews, a detailed description of the fractionation, purification and characterization of these proteins was provided. However, most of this work revolved around protamines that had been obtained from vertebrate organisms. For a more recent overview on vertebrate protamines the reader is referred to the thorough review by Oliva and Dixon (Oliva and Dixon 1991).

In this minireview we will briefly summarize the main structural characteristics of vertebrate and invertebrate protamines and the information available about their genes in a comparative fashion, with a primary focus on the most recent findings. The implications of this information for sperm chromatin and the evolution of this group of proteins will also be discussed.

THE PROTAMINE FAMILY OF PROTEINS

Figure 1A shows a synoptic comparison of vertebrate and invertebrate protamines from representative organisms in each group. A high arginine content is the primary common feature of all these proteins, which occurs in clusters. A chemical definition of protamines

can be drawn from the sequence information available: His+Lys+Arg = 45-80 mol %, Arg ≥ 30 mol %, Ser+Thr+Gly = 10-25 mol% (Ausió 1999). Despite these compositional constrains, these proteins exhibit very little conservation (see Fig. 1A), and a high extent of heterogeneity both at the protein and gene levels, a characteristic that reflects their

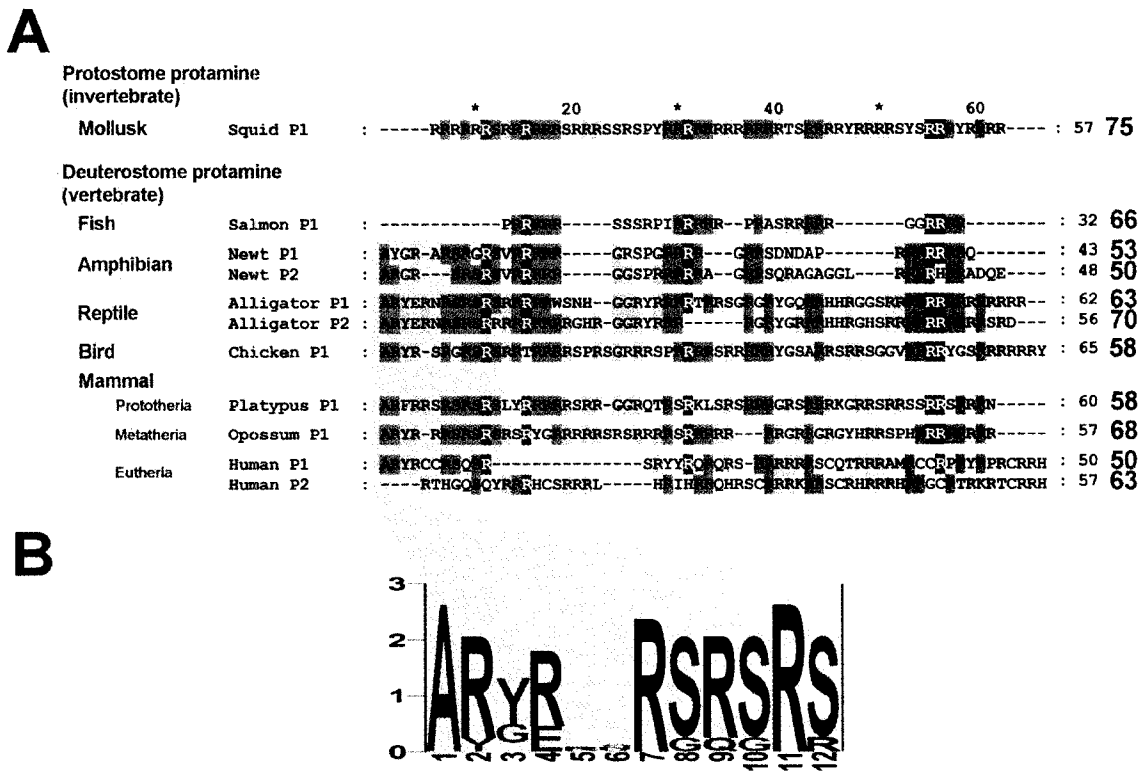


Figure 1. A. Primary structure of several invertebrate and vertebrate protamines. The sequences were obtained from the following sources. Squid P1 from *Loligo opalescens* (Ausió et al. unpublished results [see Chapter 6]); Salmon P1 (salmine SLI) from *Oncorhynchus keta* (chum salmon) (Hoffmann et al. 1990); Newt P1/P2 from *Cynops pyrrhogaster* (japanese newt) (Yoshinobu et al. 1997); Alligator P1/P2 from *Alligator mississipiensis* ((Hunt et al. 1996) and unpublished results); Chicken P1 (galline) from *Gallus domesticus* (fowl) (Nakano et al. 1975); Platypus P1 from *Ornythorhynchus anatinus* (Retief et al. 1993a); Opossum P1 from *Didelfis marsupialis* (Winkfein et al. 1993); Human P1/P2 from *Homo sapiens* (Domenjoud et al. 1990). The sequences were aligned using the CLUSTAL X multiple sequence alignment program (Thompson et al. 1997) and the BLITZ server at EMBL, which uses the best local similarity algorithm (Smith and Waterman 1981). The bold numbers (in red) at the right hand side of the sequences indicate the percentile of basic amino acids. **B.** The N-terminal conserved region of vertebrate protamine P1 displayed in a Logos format (Schneider and Stephens 1990). 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 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 they represent.

rapid rate of evolution.

Most of the information on invertebrate protamines comes from molluscs (Subirana et al. 1973), which is one of the groups where they have been more extensively characterized. This includes the protamines from polyplacophors, gastropods and cephalopods (Chiva et al. 1991; Wouters Tyrou et al. 1998). Some compositional information is available in a few insects (Bloch 1969; Kasinsky 1989) [including a cDNA sequence (Trewitt et al. 1990)] and in algae (Reynolds and Wolfe 1984). As with their vertebrate counterparts, the main characteristic of these proteins is the highly abundant presence of Arginine clusters (see Fig. 1A, Squid P1) within a relatively short amino acid sequence. To date, the largest

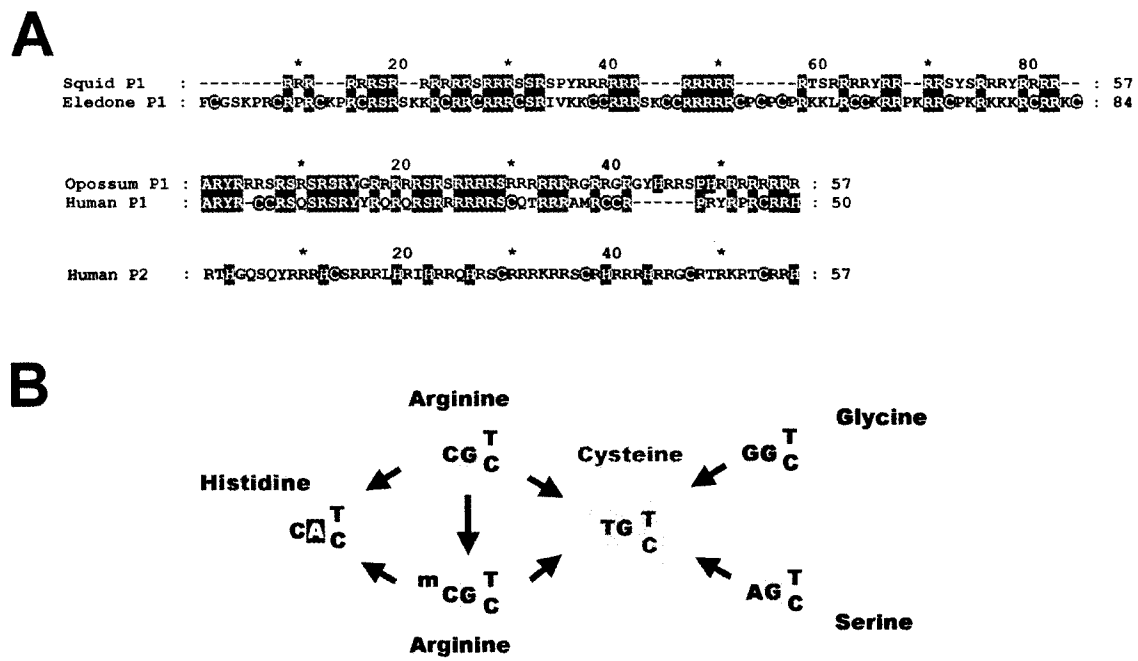


Figure 2. A. Occurrence of Cysteine in invertebrate and vertebrate protamines. The sequences of cysteine-containing protamines of an octopus protamine Eledone P1 (from *Eledone cirrhosa*) (Gimenez-Bonafe et al. 2002) and human P1 (from *H. sapiens*) (Domenjoud et al. 1990) are shown in comparison to the protamine P1 sequences of phylogenetically related organisms, squid (squid P1) and opossum (Opossum P1) (see legend to Figure 1A). **B.** Possible molecular mechanisms to explain the amino acid transitions undergone by protamines in the course of evolution, including the recent acquisition of Cysteine (Oliva and Dixon 1990).

protamine to be described is from molluscs (Daban et al. 1995).

By comparison, the vertebrate protamines have been extensively characterized (Kasinsky 1989; Oliva and Dixon 1991). However, information has become available only recently regarding amphibian and reptilian protamines (Takamune et al. 1991; Hunt et al. 1996; Yoshinobu et al. 1997; Wouters Tyrou et al. 1998) and their genes. Thus, it is now possible for the first time to perform a complete comparative analysis of these proteins such as that shown in Fig. 1A. Several general observations can easily be made with regards to this Figure. To begin with, during the course of vertebrate evolution there has been a gradual increase in the length of the protamines, a trend that appears to have begun already in fish (as it will be discussed later). Among vertebrates, fish contain the shortest protamine molecules. A sequence analysis of the N-terminal region of the tetrapod protamines (see Fig. 1B) reveals the appearance of a signature ARYR domain which, while incipient in amphibians, appears to be clearly established in amniotes and hence could be labelled as the amniote motif (Hunt et al. 1996). This sequence is followed by a stretch of alternating SR residues that are a phosphorylation target (Papoutsopoulou et al. 1999), which is also common in invertebrate SNBPs (Ausió 1999). As it will be discussed below, this SR domain may be important for protamine deposition. Finally in the transition from Metatheria to Eutheria, Cysteine, which is otherwise uncommon in other chromosomal proteins (van Holde 1989), appears to form an important part of the protamine sequence (see Fig. 1A). Cysteine has also been shown to be present in cephalopods (Gimenez-Bonafe et al. 2002) (see Fig. 2A).

Some eutherian mammals contain a second distinct protamine (P2) in addition to the major P1 component. The sequence of this protein differs substantially from that of P1 (see Fig. 1A). In addition to containing Cysteine, it is also rich in Histidine, a feature shared with alligator and ostrich P1 (paleognathus bird) protamines that also contain this amino acid (Fig. 1A) (Ausió et al. 1999). The possible nucleotide sequence transitions involved in the appearance of Histidine and Cysteine in protamines is shown in Figure 2B.

The role of this protamine P2 variant is very intriguing. While all eutherians (placental mammals) contain the P2 gene, only several of them (like mouse, hamster, human and stallion) (Hecht 1989) actually express P2. For instance, in boar and bull the absence of this protein has been shown to be due to mutations within the P2 gene (Maier et al. 1990), and its absence in rat has been shown to be the result of suppression at both the transcriptional and translational level (Hecht 1989). The sequence variability of protamines, whether of invertebrate or vertebrate origin, has not only important evolutionary implications, but is of important functional significance for the inter- and intra-species fecundation ability of sperm. Indeed, partial replacement of mouse P1 by galline in transgenic mouse did not only affect chromatin stability (Rhim et al. 1995), but it also led to impairment of fertilizing activity (Maleszewski et al. 1998). Similarly, haploinsufficiency of either protamine 1 or 2 in mice has been shown to cause infertility (Cho et al. 2001), as has lack of expression of P2 in humans (Balhorn et al. 1988; de Yebra et al. 1993). However, at a higher developmental level, the presence or absence of protamines does not appear to impose a barrier in the development of the zygote. It has been shown that normal development can occur in mammals when round (histone-

containing) spermatids are inserted in oocytes (Kimura and Yanagimachi 1995). This observation could account for the broad heterogeneity of SNBPs and developmental indifference toward the SNBP type (Ausió 1999) that is observed in sexually reproducing organisms (Bloch 1969; Kasinsky 1989).

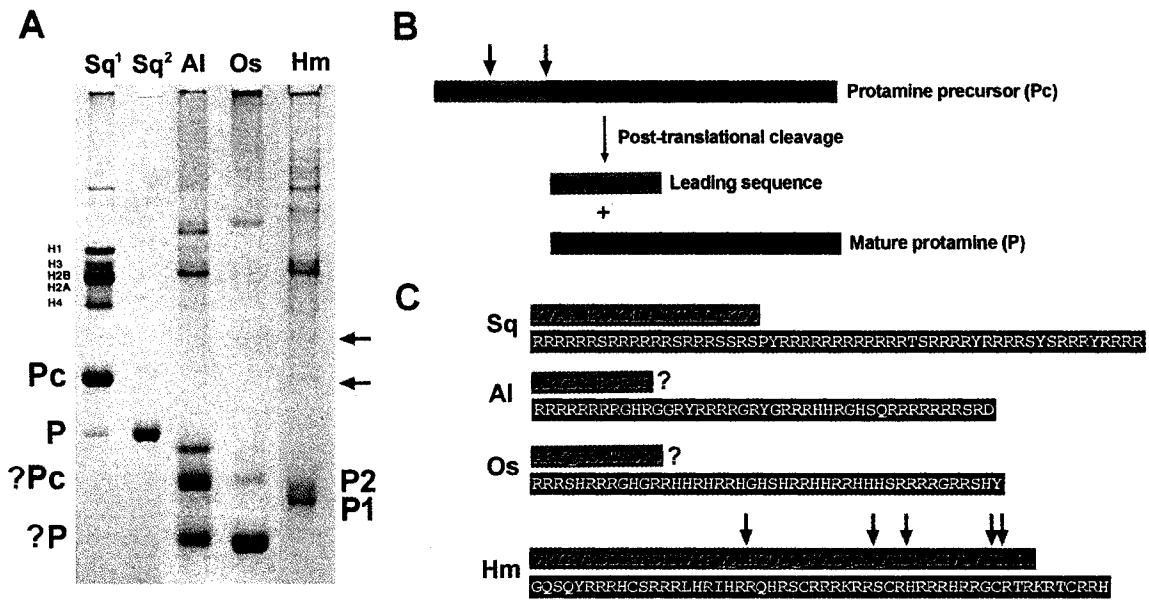


Figure 3. A. Acetic acid (5%)-urea (2.5 M) polyacrylamide gel electrophoretic analysis of the SNBPs from: Sq^1, immature squid (*L. opalescens*) testes; Sq^2, squid (*L. opalescens*) sperm; Al, mature alligator (*A. mississipiensis*) testes; Os, ostrich (*Struthio camelus australis*) sperm (Ausió et al. 1999); and Hm, human sperm. The arrows point to some of the human P2 precursors. Pc = precursor protamine, P = processed protamine. **B.** Post-translational (cleavage) processing of protamines is frequently observed in both invertebrate protamines. The portion of the sequence highlighted in dark blue corresponds to the mature processed molecule and the sequence removed is shown in light blue.

PROTAMINE PROCESSING AND MICROHETEROGENEITY

The functional significance of protamines in male gametogenesis (spermiogenesis) must also be understood in light of protein processing and microheterogeneity. It is still not clearly understood why some organisms process their protamines while others do not, or why certain organisms contain multiple protamine genes while some contain only a

single copy.

Post-translational processing of SNBP precursors is a fairly common process in both vertebrate and invertebrate organisms (Chiva et al. 1995) (see Fig. 3). The process usually involves the removal of a leading peptide. This can take place in one step as it happens with the protamines of some cephalopods (Wouters Tyrou et al. 1998) (Fig. 3C Sq), or it could be the result of a sequential cleavage, as occurs in the case of mammalian P2 (Hecht 1989) (Fig. 3C Hm) or in neogastropod molluscs (Caceres et al. 1999). In molluscs, the processing is accompanied by important changes in chromatin condensation (Caceres et al. 1999) and thus it appears that the leading sequences of the protamine precursors may have an important role in the processes of histone replacement and proper protamine deposition during spermiogenesis. In mammals, the functional relevance of P2 processing has been related to the protein half life (Hecht 1989), but it is possible that it also plays a role in proper sperm chromatin condensation. Recent experimental evidence from our laboratory (Ausió et al. unpublished) suggests that reptile and ostrich protamine P1 may also undergo a post-translational cleavage that involves the removal of the N-terminal domain amniote motif (see Fig. 3C Al-Os).

It has long remained a mystery why there are only a very restricted number of protamine genes in some organisms, whereas in others there are a significantly larger number that tend to display protein microheterogeneity. This microheterogeneity is indeed one of the striking features of protamine chemical and structural composition (Subirana 1983). In vertebrate organisms, microheterogeneity has been observed in salmonid fish (Oliva and Dixon 1991) and in some reptiles (like alligators and turtles). However, other fish and

reptiles contain one or two protamine genes at most (Oliva and Dixon 1991). A comparative analysis carried out in reptiles (Hunt et al. 1996) suggested that such microheterogeneity takes place at the onset of the gene duplication processes involved in the evolution of protamine genes (Oliva and Dixon 1991). The functional relevance of such microheterogeneity, if any, still remains to be elucidated.

PROTAMINES AND CHROMATIN STRUCTURE

While protamines exhibit a random coil conformation in solution, the possibility exists that they adopt a certain extent of secondary structure upon neutralization of the Arginine positive charge as a result of their electrostatic interaction with the phosphate backbone of DNA. Although the binding location of these highly charged proteins to either the major or minor groove of DNA has long been debated [see (Raukas and Mikelsaar 1999) for a recent review], the argument appears to be finally settled in favour of the major groove (Subirana 1991). The charge neutralization resulting from this interaction is responsible for the bending of DNA (Kosikov et al. 2002) that ultimately results in the highly compact toroidal nucleoprotamine structures (Brewer et al. 1999) observed in mammalian sperm (Ward and Coffey 1991) or in the globular and lamellar or structures observed in the sperm of other vertebrates (Gusse and Chevaillier 1978) and invertebrate organisms (Suzuki and Wakabayashi 1988; Caceres et al. 1999; Gimenez-Bonafe et al. 2002). The occurrence of Cysteine provides an efficient mechanism to form inter chromatin fiber associations [see (Balhorn et al. 1992; Gimenez-Bonafe et al. 2002)]. The single molecule approach technology recently developed to study protamine-DNA interactions (Brewer et al. 1999) may prove extremely valuable in ascertaining the

detailed molecular mechanisms involved in these processes.

At the onset of spermiogenesis, during the early replacement of histones by protamines, phosphorylation of Ser/Thr residues plays a very important role in the initial deposition of protamines and ensures the proper formation of the nucleoprotamine complex [see (Raukas and Mikelsaar 1999) for a review]. In this regard, the SR repeats that are present at the N-terminal region of many invertebrate (Ausió 1999) and vertebrate protamines (see Fig 1A) may play a very important role in this process. Several human protamine kinases including a "SR" protein-specific kinase that phosphorylates human P1 protamine have been recently identified (Papoutsopoulou et al. 1999; Wu and Grunstein 2000). It is also possible that some of the post-translational cleavage describe in the preceding section may also play an important role in mediating the protamine deposition process (Hecht 1989).

Regardless of the chromatin organization of the nucleoprotamine complexes, chromosomes exhibit in many instances an ordered arrangement within the sperm nucleus (Ward and Zalensky 1996).

THE PROTAMINE GENES

Most protamine genes are closely linked within the same chromosome and at least in the case of mammalian chromosomes they are located in autosomes (chromosome 16 in both mouse and human) rather than in sex chromosomes as it had initially been hypothesized (Bloch 1969).

At the nucleotide sequence level, the regions corresponding to the coding domains also exhibit a high extent of sequence variability that mirrors that of the corresponding protein sequences encoded. In contrast to what is observed in eukaryotic proteins, the majority of the genes encoding DNA-binding chromosomal proteins (histones and protamines) are intronless. The gene information currently available on protamines indicates that neither the squid protamine genes (Ausió et al. unpublished data [see Chapter 6]) nor those of

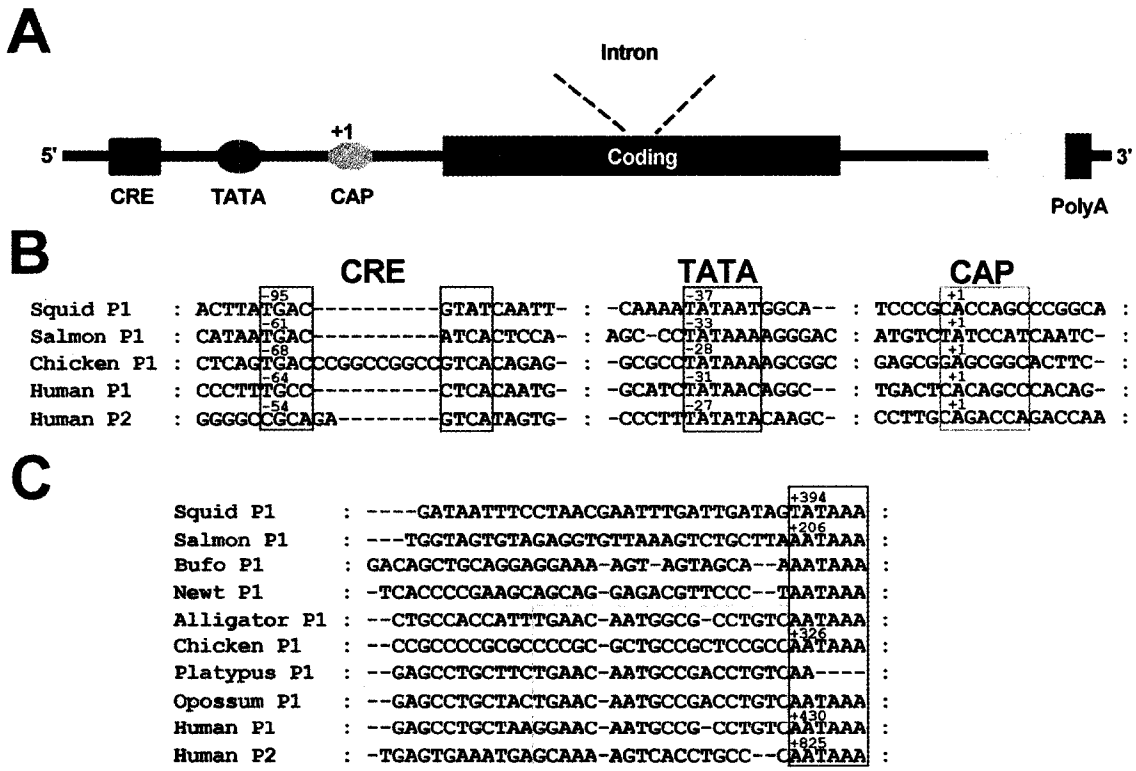


Figure 4. A Schematic representation of a protamine gene (Oliva and Dixon 1990). **B.** Alignment of the nucleotide sequences of the CRE, TATA box and CAP site at the promoter region of several representative invertebrate and vertebrate protamine genes. The nucleotide positions with respect to the first nucleotide of the mRNA CAP site are indicated (Oliva 1995). **C.** Sequence alignment of a portion of the 3' UTR of several protamine genes. The polyadenylation signal (indicated by the box) and the preceding 31 nucleotides in the 5' direction are shown. The nomenclature of the genes is the same as that in Figure 1. Bufo P1 corresponds to the gene for the toad protamine P1 from (*Bufo japonicus*) (Takamune et al. 1991). The GenBank sequence accession numbers are: Squid P1 (AY269798); Salmon P1 (X07511); Bufo P1 (X56529); Newt P1 (D85426); Alligator P1 (Ausió et al. unpublished); Chicken P1 (M28100); Platypus P1 (Z26849); Opossum P1 (X74044); Human P1/P2 (Z46940). See page 54 for explanation of dashed orange box.

fish, amphibians, reptiles or birds contain any intron with the exception of mammalian P1 and P2 protamines (Oliva 1995). The biological significance of this recent intron acquisition is still unclear. It is worth noting that the exons of these latter protamines exhibit a higher variability than that of their corresponding introns (Retief and Dixon 1993; Retief et al. 1993b). The high variability of the protamine gene sequences has historically complicated the analysis of these genes (Oliva and Dixon 1991).

Despite this, comparison of the 5' end flanking non-coding regions of protamine genes (see Fig. 4B, C), including that of an invertebrate (squid), has allowed us to identify, in all instances, the major transcriptional regulatory elements initially ascribed to vertebrate protamine genes (Fig. 4A) (Oliva and Dixon 1990; Oliva and Dixon 1991).

In vertebrates, protamine mRNA is synthesized in the haploid post-meiotic genome and the transcripts are stored for several days before protein synthesis starts in elongated spermatids (Hecht 1989; Oliva and Dixon 1991; Steger 1999). It is the 3' flanking region of the protamine genes (Fig. 4C) that contain elements involved in translational regulation. Without exception, 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 (Hecht 1989; Steger 1999). All *cis*-acting elements involved in the translational regulation of mammalian protamine P1 (as is likely in all protamine genes) are found within the 3'-UTR (Steger 1999). The 20-22 nucleotide sequence preceding the 5'AAUAAA 3' polyadenylation sequence (see Fig. 4C) shows a significant extent of similarity among the amniote protamine genes and

binding of sequence-specific RNA-binding proteins to these regions has been suggested to play an important translational regulatory role (Steger 1999). Two such proteins, MSY2 and MSY4, have recently been identified in mouse (Giorgini et al. 2001) that bind to a 5'-UCCAUCA-3' consensus sequence located in the 3' UTR of the protamine P1 mRNA. They belong to an evolutionarily ancient family of Y-box nucleic acid binding proteins (Wolffe 1994). Putative Y-box binding sequences have been identified in other mammalian protamines including P2 (Giorgini et al. 2001).

THE EVOLUTION OF PROTAMINES

As shown in Fig. 5A, protamines are amongst the proteins with the highest rate of evolution (Oliva and Dixon 1991; Oliva 1995), a trait they share with other reproductive proteins (Swanson and Vacquier 2002). This characteristic allows protamines to be used as very reliable evolutionary markers to distinguish between closely related species (see Fig. 5B).

The reason for the protein sequence divergence that results from this high rate of evolution is still controversial. As can be seen in Fig. 1A, protamines exhibit a large variability in the positions occupied by Arginine residues and in the overall length of the protein, which is also reflected at the nucleotide level. Yet, despite such sequence variability in protamines from different taxonomic groups, the relative proportion of Arginine remains fairly constant (50-70 %) (Rooney and Zhang 1999) (see bold numbers in Fig. 1A).

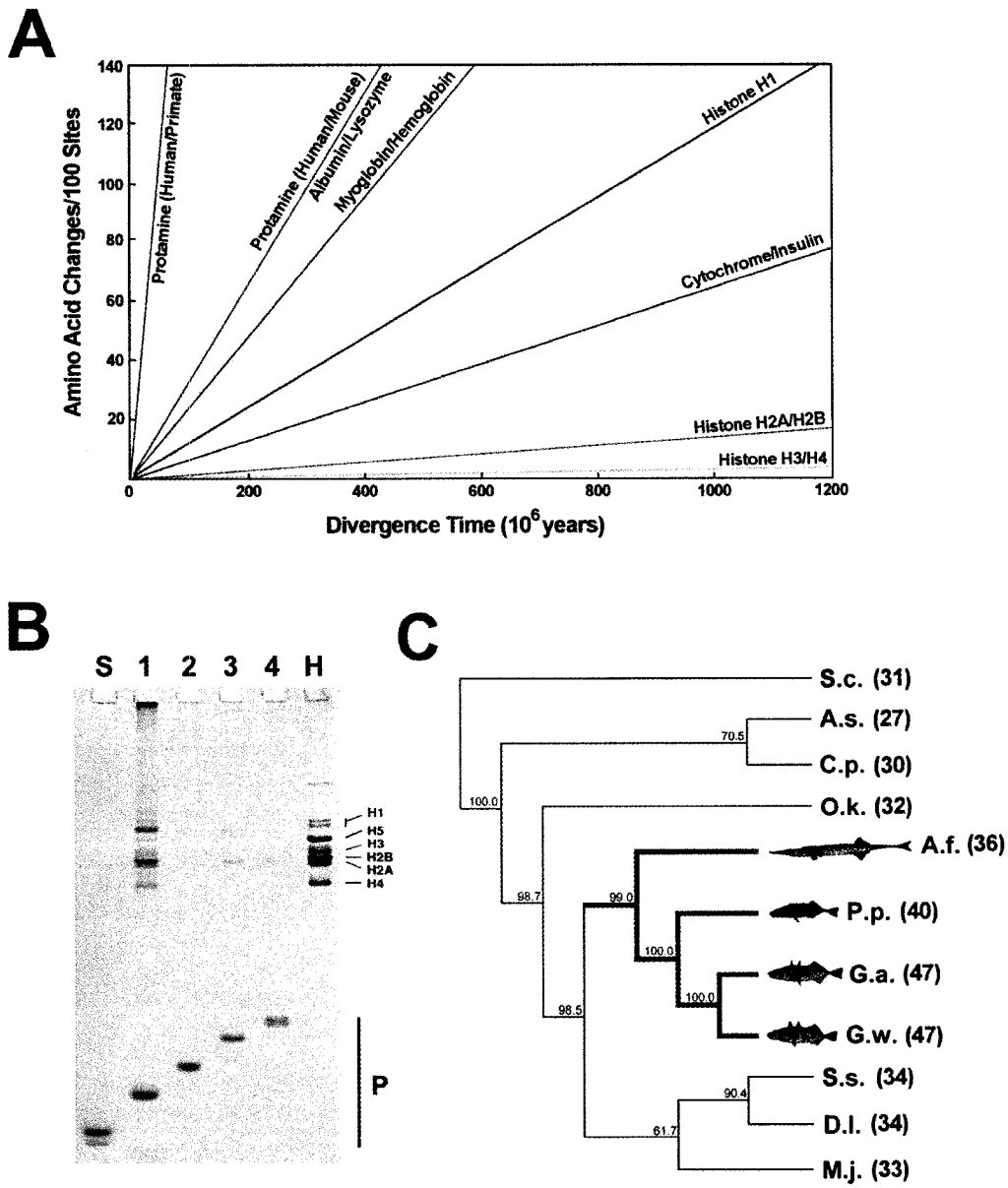


Figure 5. A. Evolutionary rates (amino acid change/100 sites) of mammalian P1 protamines (Retief et al. 1993b) compared to histones (Isenberg 1978) and other proteins. Protamine (human/mouse) and (human primate) indicate the rate of protamine P1 change between human and mouse and humans and other primates respectively. **B.** Acetic acid (5%)-urea (2.5 M) polyacrylamide gel electrophoretic analysis of the SNBPs of different stickleback species. (Gimenez-Bonafe et al. 2000) 1. *Aulorhynchus flavidus*; 2. *Pungitius pungitius*; 3. *Gasterosteus wheatlandi*; 4. *Gasterosteus aculeatus*. S and H are salmine (salmon protamine) and chicken erythrocyte histones respectively, used as markers. **C.** Bootstrap protein parsimony tree generated from amino acid sequences of protamines from different fish (Gimenez-Bonafe et al. 2000) : S.c., *Scyliorhynus canicula* (lesser-spotted dogfish); A.s., *Acipenser sturio* (sturgeon) ; C.p., *Clupea pallasii* (herring) YII. O.k., *Onchorhynchus keta*; A.f., *A. flavidus*; P.p., *P. pungitius*; G.a., *G. aculeatus*; G.w., *G. wheatlandi*; S.s., *Scomber scombrus* (mackerel); D.l., *Dicentrarchus labrax* (sea bass); and M.j., *Mugil japonicus* (mullet). The numbers at the nodes indicate the number of times that the group consisting of the species to the right of the fork occurred among the trees generated out of 100 trees. The bold (green) numbers in brackets indicate the number of amino acids of the corresponding protamines. [Figure modified from (Gimenez-Bonafe et al. 2000). Reproduced with permission from Wiley-Liss Inc.]

Although still controversial (Clark and Civetta 2000), this rapid evolution has been ascribed to positive Darwinian selection (Rooney and Zhang 1999; Wyckoff et al. 2000; Swanson and Vacquier 2002). In the case of primate P1/P2, evidence in support of protamine evolution being the result of adaptive evolution came from comparison of the coding nucleotide sequences. It was found that the ratio of non-synonymous substitutions per site to the number of synonymous substitutions per site was considerably larger than 1 (Wyckoff et al. 2000), a fact that is taken as a good indication of Darwinian selection (Rooney and Zhang 1999; Rooney et al. 2000; Wyckoff et al. 2000; Swanson and Vacquier 2002). The term "non-synonymous change" refers to nucleotide changes that alter the amino acid sequence, whereas "synonymous changes" are silent changes that do not alter the amino acid sequence.

The arginine-rich selection involved in this process has been attributed to an unusual form of purifying selection that maintains a high proportion of Arginine as opposed to conserving its position (Rooney et al. 2000). Because Arginine is encoded by six codons, all of which contain G in the second position, selection pressure operates, according to these authors, in a way that maintains the occurrence of G at a high frequency in the second position of any given protamine codon (see Fig. 6). As can be seen in Fig. 6, using the analysis described by Rooney et al. (Rooney et al. 2000) the G content in the second codon position is highly elevated for all the protamine genes analyzed including the squid (Fig. 6 squid), and the other only invertebrate protamine of an insect (Boll weevil, *Anthonomous grandis*) for which gene information is available (Rooney et al. 2000).

The driving force behind the selection process is probably determined by the function of protamines in tightly condensing sperm chromatin. The infertility resulting from protamine P1/P2 haploinsufficiency in mice (Cho et al. 2001), as well as the impaired fertilizing activity of sperm expressing heterologous protamines in vertebrates (Maleszewski et al. 1998), suggest that sexual selection may also play a role (Wyckoff et al. 2000).

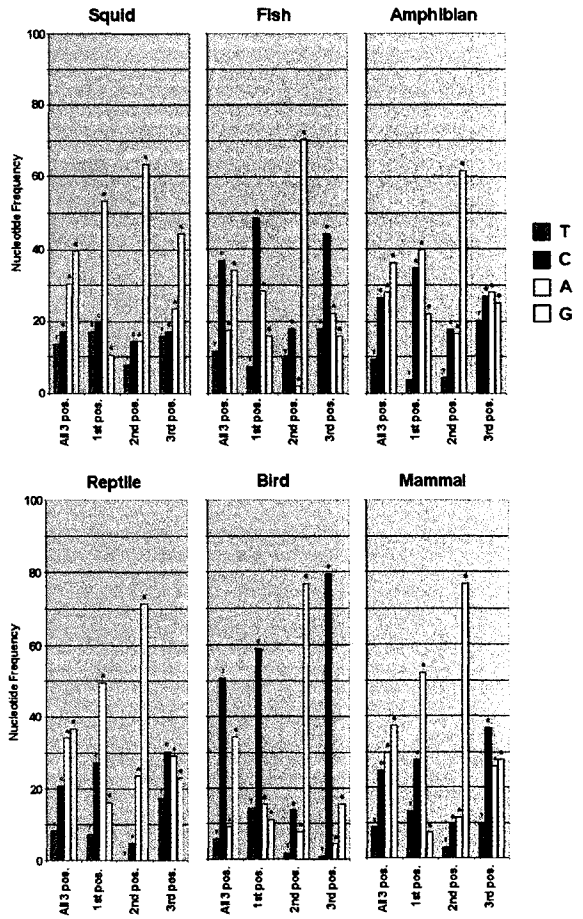


Figure 6. Nucleotide composition of the protamine P1 genes from selected invertebrate and vertebrate organisms. The species used in each group are: **Squid**: *L. opalescens* (AY269798); **Fish:** Salmon P1 from *O. keta* ; Trout P1 from *Oncorhynchus mykiss* (X01599); Dogfish protamine from *Scyliorhinus canicula* (X04517); **Amphibians:** Newt P1 and P2 (D85427) from *C. pyrrhogaster* and Toad P1 (X56529) and P2 (X56530) from *B. japonicus*; **Reptiles:** *A. mississipiensis* (PI and PII) (Ausió et al. unpublished results); **Birds:** Chicken P1 (galline) from *G. domesticus* and Quail P1 *Coturnix coturnix* (M30275); and **Mammals:** Platypus P1 from *O. anatinus*; Opossum P1 from *D. marsupialis*; Human P1 from *H. sapiens*; Killer whale P1 from *Orcinus orca* (Z11496). The nucleotide sequences were analyzed as described in (Rooney et al. 2000) using the computer program MEGA, version 1.03 (Kumar et al. 2001). The accession numbers for species other than those described in the legend to Fig. 4 are given in parenthesis.

Another important question regarding protamine evolution has to do with the origin of the protamine genes and their relation to other SNBP types (Ausió 1999). A process of horizontal evolution in which protamines were proposed to have a retroviral origin (Jankowski et al. 1986; Oliva and Dixon 1991) was put forward in order

to account for the apparent random distribution of protamines in fish. However, an exhaustive analysis carried out over a large number of species from different groups of fish later revealed that the sporadic distribution of SNBP observed in this group was not random and could be traced phylogenetically (Saperas et al. 1994). An alternative vertical process in which the three major SNBP types (H= histone; PL= protamine-like; and P= protamine) are related by H→ PL→ P (starting from a primitive histone precursor) has been proposed in which the random occurrence of SNBP types in organisms belonging to both the protostome and deuterostome branches can be explained by repeated reversions on this process (Ausió 1999). The apparent similarity of protamine primary structures (including the occurrence of Cysteine and post-translational cleavage), as well as that of the flanking regions of their genes (Figs. 2, 3), strongly suggests that the evolution of these proteins provides one of the best examples of both structural and functional convergence (Doolittle 1994).

SUMMARY, CONCLUSION, AND REMAINING QUESTIONS

More than a century has passed since the discovery of protamines by Miescher (Miescher 1874), and much knowledge has been amassed since the first identification of their protein nature (Kossel 1896a; Kossel 1896b). The primary structures and gene sequences of protamines from several representative organisms of each of the vertebrate classes are available (see Figs. 1 and 4). Similar information, although to a much lesser extent, is also available for protamines of several invertebrate groups. Therefore, for the first time now it is possible to have a comprehensive view of this important group of SNBP. The occurrence of protamines in organisms from protostomes and deuterostomes is

additionally remarkable in that some protamines from both groups exhibit similar post-translational cleavage (Fig. 3) and Arginine to Cysteine transitions (Fig. 2), making a strong case for evolutionary convergence.

Despite the amount of compositional information gathered on protamines and their genes, there are still some unresolved issues at this level. Variation in the extent of protamine microheterogeneity among different organisms and the functional role of this microheterogeneity remains unexplained. Similarly, the significance of the post-translational cleavage undergone by many vertebrate and invertebrate protamines during spermiogenesis and the possible involvement or significance of this processing in the formation of the nucleoprotamine complexes found in mature sperm are not well understood.

The structure of nucleoprotamine chromatin complexes has been characterized by a broad spectrum of biophysical techniques (Raukas and Mikelsaar 1999), more recently using state of the art single molecule analysis methods (Brewer et al. 1999). Yet, some of the intricate molecular details involved are still unknown. The secondary and tertiary structural organization, if any, of protamines upon their interaction with DNA is still an unsettled topic (Raukas and Mikelsaar 1999).

At the gene level, several putative *cis*-acting regulatory regions have been identified both at the 5' and 3' region flanking the coding domain (Fig. 4). However, additional work is required to identify the *trans*-acting factors involved.

Lewis et al. (2003) Chromosoma **111**:473-82.

Recently, the study of protamines has been particularly focused on their evolution. Due to the rapid evolutionary rates of these proteins, they are highly valuable as phylogenetic markers (see Fig. 5). In the case of primate protamines, a strong case has been made for the role of natural selection in the evolution of these proteins (Wyckoff et al. 2000), and the involvement of a purifying process of selection (Rooney et al. 2000) (see also Fig. 6). In this regard, it will be extremely important to shift the focus of this research toward a broader spectrum of invertebrate and vertebrate protamines from groups other than mammals, in order to vindicate the general validity of the current evolutionary theories with regards to the overall protamine family of chromosomal proteins.

ACKNOWLEDGEMENTS

This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant OGP 0046399.

The PL-I gene of *Spisula solidissima* encodes a novel and highly elongated sperm-specific histone H1.

John D. Lewis and Juan Ausió§

Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6

§ to whom all correspondence should be addressed

Department of Biochemistry and Microbiology
University of Victoria
Petch Building, Room 220
Victoria, BC
Canada V8N 5Y2
Tel: 250-721-8863
e-mail: jausio@uvic.ca

Lewis and Ausió (2003) JBC (submitted).

ABSTRACT

In the mature sperm of the surf clam, *Spisula solidissima*, the DNA is complexed with a small subset of germinal histones and a single histone H1-related PL-I protein of low electrophoretic mobility. We describe here the complete gene sequence of the PL-I of *Spisula solidissima*. The gene encodes a protein of a surprising 453 amino acids, making it the largest H1-like PL to be described in the bivalve molluscs. The predicted mass (51039 Da) was confirmed using ESI mass spectrometry, which indicated a tetra-phosphorylated state for the protein in the mature sperm. The amino-terminal tail of the *Spisula* PL-I is greatly elongated due to the presence of 39 tandem hexapeptide repeats; 14 of the motif (K/R)KRSAS and the remaining having single, double and a few triple semi-conservative amino acid substitutions. These repeats are very closely mirrored by their encoding DNA sequence, which indicates that an expansion due to sequence duplication most likely occurred. Our analysis of the promoter and 3' UTR has revealed a number of conserved binding sites, one of which appears be the target of factors responsible for the repression of mRNA translation during spermiogenesis.

Lewis and Ausió (2003) JBC (submitted).

INTRODUCTION

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). The SNBPs can be grouped into three major types: Histone (H type), protamine-like (PL type) and protamines (P type) (Ausió 1995). Protamine-like proteins are a structurally heterogeneous group of chromosomal proteins with a rather homogeneous amino acid composition intermediate to that of histones and protamines (Ausió 1995). They are referred to as protamine-like because, like protamines, they replace most 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) (Ausió and Subirana 1982b). 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 (Ausió et al. 1987). 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 (Ausió 1988; Ausió 1992). The winged-helix motif is found in a variety of proteins, including histone H5 (Garel et al. 1975; Sautiere et al. 1975) and a number of transcription factors such as HNF3 (Cirillo et al. 1998) and the FOXn1 family of oncogenes (Schlake et al. 2000). The highest sequence similarity was found with the globular core of the chromatin-condensing protein histone H5, which condenses the chromatin of terminally differentiated chicken erythrocytes.

Lewis and Ausió (2003) JBC (submitted).

Spisula PL-I binds to DNA in an extremely cooperative fashion, with cooperativity parameters and strength of binding larger than those observed for fish protamines and histone H1 (Libertini et al. 1988). The initial characterization of *Spisula* PL-I revealed an aggregation behaviour that could be visualized as a band that was present on top of the wells on highly cross-linked polyacrylamide urea-acetic acid gels (Ausió and Subirana 1982b). This was a particularly puzzling behaviour due to the high composition of arginine and lysine. Such behavior was therefore attributed to hydrophobic interactions (Ausió and Subirana 1982b), but it was later discovered that it was the result of the presence of a single cysteine residue in the globular region that was overlooked during conventional amino acid microsequencing (Zhang et al. 1999).

The evolution of the SNBPs in bivalve molluscs is hypothesized to have originated with an ancient histone H1 (Ausió 1999). Through a process of extension of the N-terminal tail and enrichment in arginine, followed by segregation of the proteins (by post-translational cleavage) and eventually the genes encoding them, the SNBPs of bivalve molluscs appear to be evolving towards "true" protamines. Histone H1-related PL-I proteins similar to the one found in *Spisula* have also been found in the sperm of the razor clam, *Ensis minor* (Bandiera et al. 1995) and *Mytilus californianus* (Carlos et al. 1993b). The *Spisula* PL-I is distinct in the regard that it does not undergo post-translational cleavage to produce smaller SNBPs, as do the PL precursors from both *Ensis* and *Mytilus* (Carlos et al. 1993a; Bandiera et al. 1995). *Mytilus* has a third SNBP (PL-III), that while resembling the other *Mytilus* sperm proteins (PL-II and PL-IV), appears to have undergone gene segregation

(Lewis and Ausió, unpublished [see Chapter 5]). The PL-I of *Spisula*, therefore, seems to be the most primitive of the bivalve mollusc SNBPs previously mentioned.

We set out to characterize the gene of the PL-I SNBP of *Spisula solidissima* with the hopes of gaining insight into the mechanism of histone H1-like PL expansion and evolution, and to describe for the first time the regulatory regions of an SNBP gene from bivalve molluscs. Analysis of the coding region, as well as promoter and 3' UTR regions are discussed herein with these ideas in mind.

MATERIALS AND METHODS

Living Organisms

Specimens of *Spisula solidissima* were obtained from the Department of Marine Resources at the Marine Biological Laboratory (Woods Hole, MA).

Preparation and isolation of the sperm nuclear basic proteins

Sperm nuclear basic proteins were routinely extracted with 0.4N HCl following the procedures described previously (Subirana and Colom 1987). Buffers used during the isolation of proteins contained Complete protease inhibitor cocktail tablets (Boehringer). The dried pellets were stored at -80 °C.

Protein Fractionation

Reverse phase HPLC was performed on a 5mm Vydac C18 column (25 x 3 x 0.46 cm) with 0.1% trifluoroacetic acid as eluant with varying acetonitrile gradients (Ausió 1988). 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.

Protein Gel Electrophoresis

Acetic acid (5%)-urea (2.5 M) polyacrylamide gels were prepared as described in (Jutglar et al. 1991). Like protamines, PL-proteins exhibit a very low or no solubility in SDS and therefore SDS-PAGE cannot be used for the analysis of the proteins.

DNA Extraction

DNA was extracted from gonadal (0.1g) tissue according to the protocol described by Sambrook with some modifications (Sambrook et al. 1989). The tissue was weighed, frozen in liquid nitrogen and ground to a powder using a mortar and pestle. The powder was suspended in 50 mL of 10 mM Tris-HCl (pH 8.0), 0.1 M EDTA (pH 8.0), 20 µg/mL pancreatic RNase I (Boehringer Mannheim), 0.5 % SDS and incubated for 1 hour at 37 °C. Proteinase K (Boehringer Mannheim) was then added to a final concentration of 100 µg/mL, followed by incubation for 3 hours at 50 °C with gentle swirling. This solution was cooled to room temperature and extracted twice in phenol:chloroform. A final extraction was performed with chloroform:isoamyl alcohol (49:1). The aqueous phase was then mixed with 0.2 volumes of 10 M ammonium acetate and the DNA was precipitated with 2 volumes of 95 % ethanol. A glass pipette hook was swirled in the solution to collect the high molecular weight precipitated DNA. The DNA was then immersed in 1 mL of 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0) (TE buffer) and tumbled at 4 °C until the DNA was fully resuspended (~48 hours).

Hybridization probe preparation

Due to 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 towards 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'-CATCGTCTTCTTTGTCTTCTTTGTGGTC - 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 either SpeI, EcoRI or XhoI. Marker was λDNA-*Bst*E II marker (New England Biolabs). The gel was blotted using the VacuGene® XL Vacuum Blotting System (Pharmacia Biotech) following the manufacturer's instructions. The blotting membrane used was Zeta-Probe® GT (BioRad). The gel was first depurinated for 20 minutes in 0.2 N HCl, denatured for 20 minutes in 0.5 M NaOH, 1.5 M NaCl, neutralized for 20 minutes in 1 M Tris-HCl, pH 7.5, 1.5 M NaCl, and then transferred for 1 hour to a membrane with 20 x SSC. The blot was then washed for 5 minutes in 20 x SSC to remove any agarose, air dried for 30 minutes, and vacuum-dried for 30 minutes at 80 °C. The double-stranded 306 bp insert was labelled by nick translation according to (Sambrook et al. 1989). The labelled probe was purified from the free label using a microcon® 10 (Amicon) following the manufacturer's protocol. 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), 7 % SDS at 65 °C for 30 minutes. The probe was heat-denatured at 95 °C for 5 minutes and added to the hybridization solution, 5 mL of 0.5 M Na_2HPO_4 (pH 7.2), 7 % SDS. Hybridization was carried out for 18 hours at 65 °C.

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

Degenerate PCR

Degenerate primers for PCR were created based on the determined amino acid sequence of the globular core of PL-I (Ausió et al. 1987). PCR was performed using the PCRSprint thermal cycler (Interscience) with genomic DNA as template. A touchdown profile was used for the amplification, with the annealing temperature decreasing from 65C to 45C over 20 cycles, followed by 10 cycles at 45C.

Genomic Walking

Genomic walking was performed on genomic DNA using adaptors, adaptor primers, and protocols based on (Zhang and Gurr 2000). DNA was digested overnight with *Spe*I, *Nhe*I and *Xba*I (New England Biolabs). Adaptors were ligated at 16C for 6 hours, 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 1μl 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 (Zhang and Gurr 2000).

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 manufacturer's instructions, and transformed into TOP10 competent cells (Invitrogen). DNA was sequenced by the dideoxynucleotide method (Sanger et al. 1977) using a Sequenase 2.0 kit (USB Corp).

RESULTS AND DISCUSSION

Isolation and mass determination of PL-I

A preparation of the nuclear basic proteins of mature *Spisula solidissima* sperm shows that PL-I is present as a highly concentrated homogeneous band of low electrophoretic mobility, and that approximately 30-40% of the germinal histones persist (Fig. 1, lane 1). Reverse phase HPLC fractionation of these proteins allowed resolution of PL-I from the core histones, and while H1 co-elutes slightly with PL-I, the earlier fractions of the PL-I

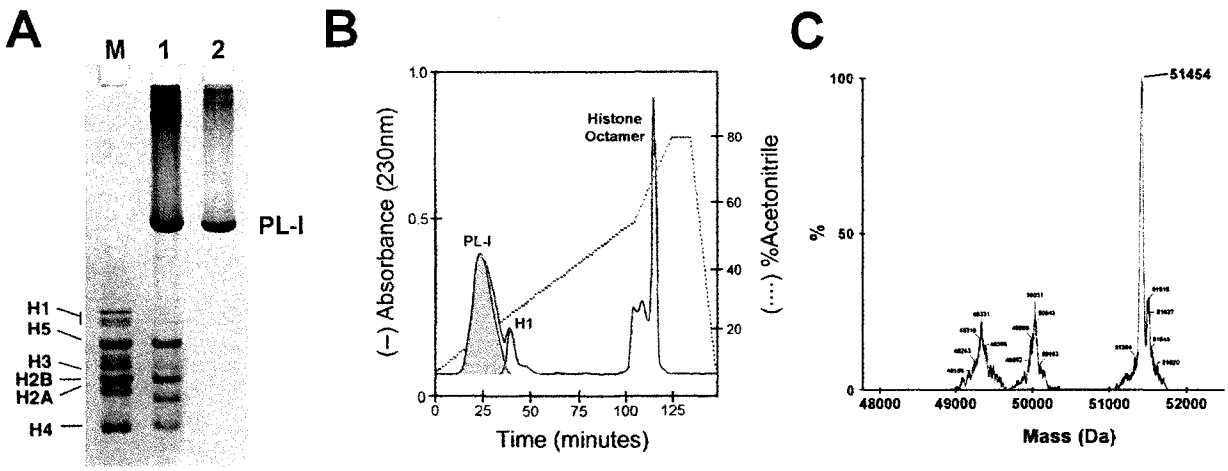


Figure 1. A. Urea (2.5M) - acetic acid (5%), polyacrylamide (15%) gel electrophoresis of 1: 0.4N HCl extract from the sperm of *S. solidissima*, and 2: HPLC fraction taken from leading edge of peak 1. M: chicken erythrocyte histones were used as a marker. **B.** Reverse-phase HPLC fractionation of HCl-extracted sperm nuclear proteins. PL-I fraction is shaded grey. **C.** Regression analysis of results obtained from electrospray ionization mass spectrometry of PL-I (lane 2 from A).

peak contain pure PL-I (Fig. 1A, lane 2 and 1B). Electrospray ionization mass spectrometry revealed that there was a small amount of heterogeneity in the sample due most likely to multiple post-translational modifications, making the data analysis complicated. Therefore, to refine the raw data, regression analysis was performed using MassLynx 4.0 (Waters, Inc.), revealing that the most likely abundant species was 51454 Da (Fig. 1C). Considering that our previous estimate of the size of *Spisula* PL-I was in the range of 35000 Da (Ausió and Subirana 1982b), this result was initially unexpected.

The PL-I gene encodes the largest SNBP of bivalve molluscs

The complete gene sequence for the *Spisula solidissima* PL-I protein was obtained in a two-step process. Degenerate PCR using primers designed using the peptide sequence of the globular core was used to amplify a 306 bp genomic clone. This information was utilized to create non-degenerate PCR primers for genomic walking in both the 5' and 3' directions, after which a total of 1863 bp were obtained. This sequence contains a single open reading frame that encodes a protein of 453 amino acids (Fig. 2), flanked by 442 bp of upstream sequence and 45 bp of the 3' UTR. The calculated average mass for the entire PL-I protein is 51139 Da, slightly smaller than the 51454 Da measured by mass spectrometry. Protamines and sperm-specific histone H1s are known to be highly phosphorylated at the outset of spermiogenesis, and become rapidly dephosphorylated as histone replacement proceeds and the chromatin is compacted (Poccia and Green 1992; Papoutsopoulou et al. 1999). While the maximum phosphorylation state of PL-I is 124 as calculated by the Peptool protein analysis package (Biotools, Inc.) and 115 as predicted by NetPhos2 (Blom et al. 1999), it appears that the phosphorylation state of PL-I in mature sperm is four per molecule. The presence of four phosphoserines would increase the predicted mass of PL-I to 51459 Da, well within the margin of error of our regression analysis when compared to the measured 51454 Da.

Figure 2. Complete gene sequence (1863 bp) for the PL-I of *Spisula solidissima*. Coding region is in capital letters, flanking regions in lower case. Green bar indicates N-terminal SR repeat motif. Red bars denote conserved hexapeptide repeats (K/R)KRSAS. Dark orange bars indicate hexapeptide repeats with a single amino acid substitution. Dark and light yellow bars indicate hexapeptide repeats with two and three substitutions respectively. The histone H1-related winged-helix globular region is denoted with the blue dots. The single cysteine residue is indicated by the cyan box. The encoded protein PL-I protein sequence is 453 amino acids in length.

The 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 *Spisula* (Fig. 2, green bar). These motifs

have been described in a number of the PL proteins, such as the PL-II of *Mytilus* and the EM-1 of *Ensis*, as well as in protamines, such mammalian protamine 1 (Domenjoud et al. 1990) and the alligator Al-I and Al-II protamines (Hunt et al. 1996). Protamine 1 is phosphorylated in testis by the SR protein-specific kinase 1 (SRPK1) (Papoutsopoulou et al. 1999), a kinase that also targets the SR motifs of splicing factors (Gui et al. 1994). Since SR proteins are also specifically phosphorylated by topoisomerase I (Rossi et al. 1996), and the chromatin remodeling events during spermiogenesis require the activity of topoisomerase I to maintain sufficient DNA relaxation (Cobb et al. 1997), it is quite plausible that topoisomerase I may also be involved in the phosphorylation of SNBPs.

The amino-terminal tail of *Spisula* PL-I, at 306 amino acids in length, is significantly elongated when compared with 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 *Mytilus californianus* is 42 amino acids long. The N-terminal tail of *Spisula* is comprised mainly of highly basic but simple hexapeptide repeats, at least 39 in total (Fig. 2). Nineteen (19) of these are (K/R)KRSAS (Fig. 2, red bars) and a further 15 repeats follow this template with a single amino acid substitution (Fig. 2, dark orange bars). The remaining repeats contain two or three semi-conservative amino acid substitutions (Fig. 2, dark and light yellow bars). 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

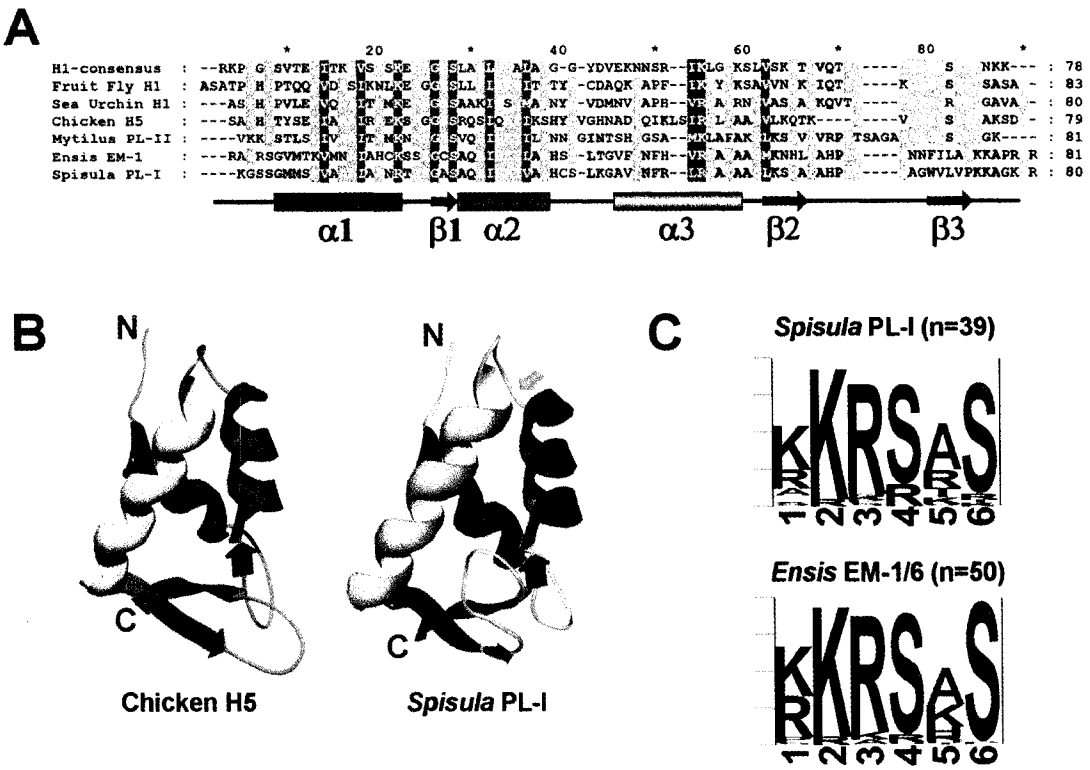


Figure 3. A. Multiple alignment of *Spisula* PL-I winged helix region with the winged helices of other histone H1 and H1-related SNBPs, with secondary structure highlighted below. The sequences were aligned using the CLUSTAL X multiple sequence alignment program (Thompson et al. 1997). The sequence accession numbers, if available, are: H1 consensus ((Wells and Brown 1991)); H1 Fruit Fly (P02255); H1 Urchin (P15869); H5 Chicken (P02259); PL-II *Mytilus* (A45317); EM-1 *Ensis* (AAA98076). **B.** Left: 3D rendering of the globular core of chicken erythrocyte H5 with coordinates as determined by (Ramakrishnan et al. 1993), Right: Theoretical 3D rendering of the globular core of *Spisula* PL-I, generated with the aid of the SWISS-MODEL server (Schwede et al. 2003). Secondary structures are colour-coded to match the secondary structure in A. Grey arrow indicates position of cysteine. **C.** The N-terminal hexapeptide repeats of *Spisula* PL-I and *Ensis* EM1/6 displayed in a Logos format (Schneider and Stephens 1990). 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 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 they represent.

PL protein from the sperm of *Ensis minor* has an N-terminal tail that bears the closest similarity to *Spisula* PL-I, with a length of 212 amino acids. It also possesses hexapeptide repeats, as does *Ensis* EM-6, which is an N-terminal cleavage product of a PL precursor (Bandiera et al. 1995). Comparison of the hexapeptide repeats from both *Spisula* and *Ensis* PL proteins is seen in Figure 3C, 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 *Spisula* 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.

Spisula 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 (Fig. 3A). The structure databases PFAM (Protein Families Database) (Bateman et al. 2002) and SMART (Simple Modular Architecture Research Tool) (Ponting et al. 1999) both 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 *Ensis minor* EM-1 is most similar at 69%.

Figure 3B shows the results obtained by using the solved H5 globular core (Ramakrishnan et al. 1993) to extrapolate the three dimensional structure of PL-I, with tools made available at the SWISS-MODEL server (Schwede et al. 2003). 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 towards 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 grey arrow in Figure 3B 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 (Zhang et al. 1999).

The PL-I gene has two genomic copies

Southern blot analysis was carried out in order to determine the copy number of the PL-I gene in *Spisula solidissima* (Fig. 4). The results indicate that the PL-I gene is present in two or a multiple of two copies. The data points strongly towards the existence of only two copies, however, as controls which contained single copy amounts of PL-I DNA displayed equivalent intensities to each of those in lanes EcoRI and XhoI in Figure 4 (data not shown). Previous studies have shown that there is a wide range of gene copy numbers for sperm

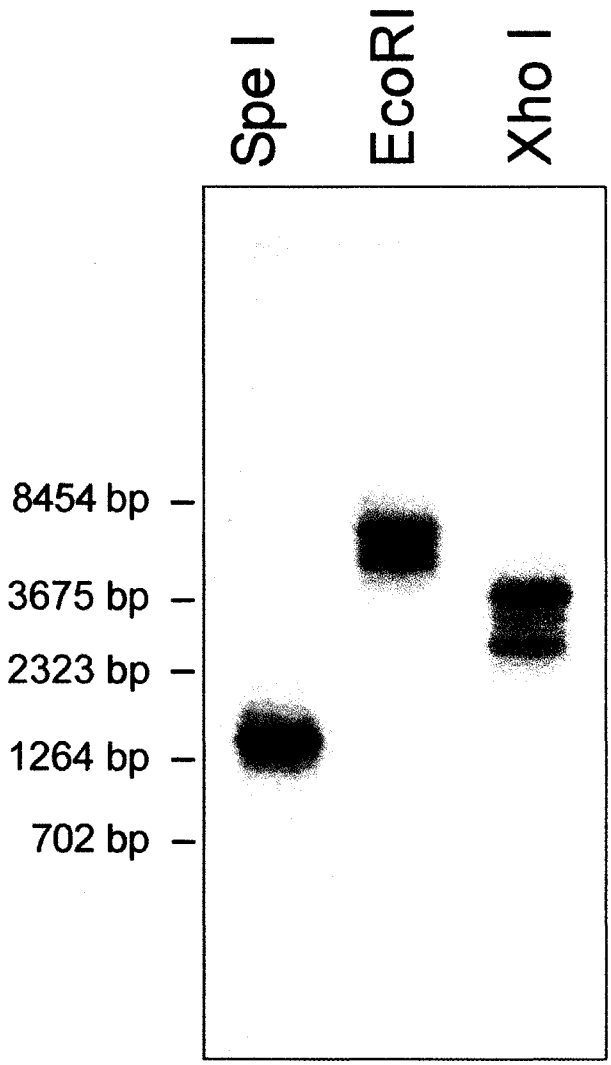


Figure 4. Southern blot of *Spisula solidissima* genomic DNA hybridized with the 306 bp PL-I probe. Each lane contains 10 µg of DNA digested with SpeI, EcoRI or XhoI as indicated. λDNA-*Bst*E II marker (New England Biolabs) positions are indicated at left .

nuclear basic proteins. 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 (Domenjoud et al. 1990). The gene for the protamine-like SNBP PL-III of *Mytilus californianus*, on the other hand, has many copies which are widely dispersed (Heath and Hilbish 1998) [see also Chapter 5]. The replacement of histones by protamines or protamine-like 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. Some organisms, such as humans, have solved this issue by building up large amounts of the requisite protamine mRNA more slowly and placing it under strict translational control. This is accomplished by the action of factors (MSY2 and MSY4) which specifically bind the 3' UTR and "silence" the mRNA until it is required for translation during spermiogenesis (Giorgini et al. 2002). Other organisms, however, seem to use a "brute force" method, and rely on high gene copy numbers to produce sufficient amounts of mRNA in the limited available time.

The PL-I protein has elongated through genomic duplication

The genomic sequence encoding the amino-terminal tail of *Spisula* PL-I provides some insight into the nature and origin of its numerous amino acid repeats. There are a number of regular and recurring sequences throughout the region coding for the N-terminal tail, with the nucleotides arranged into distinct patterns of 18, corresponding with the repeating amino acid hexapeptides. The extremely elongated nature of the N-terminal tail of *Spisula* PL-I, however, seems to have been the result of four distinct sequence

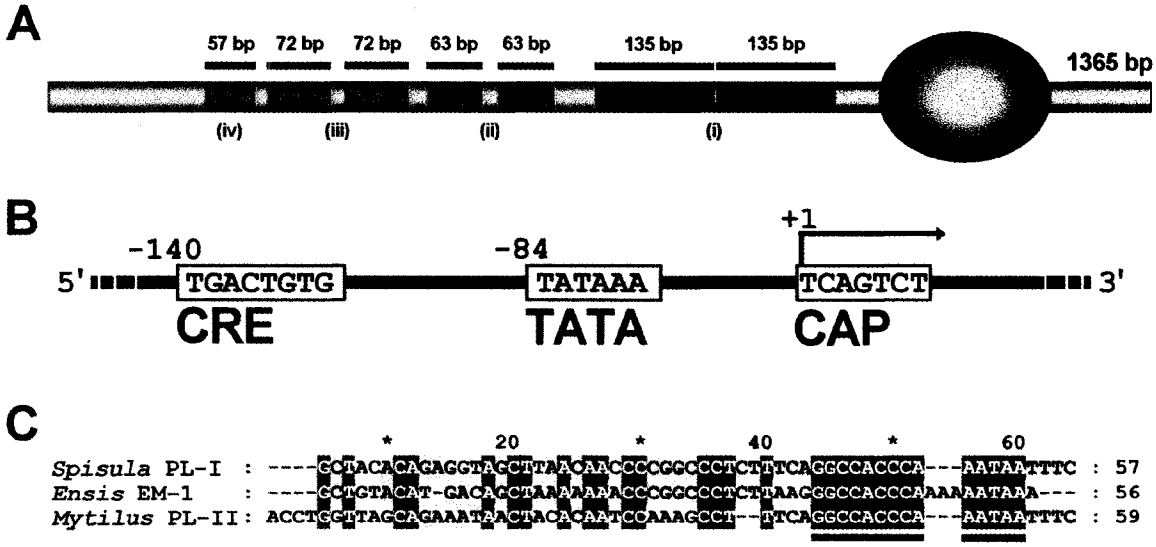


Figure 5. Diagram depicting the extensive sequence duplication within the coding region of the PL-I gene. Identical repeats appear as similar colours. Duplications (i), (ii), and (iii) are direct tandem repeats, while (iv) is a partial repeat of (iii). **A.** Putative identification of conserved structural features (CRE, TATA box, CAP site). There was a conspicuous lack of a consensus H4 box which is present in the genes of histone H5 and the H1s of sea urchin (Peretti and Khochbin 1997). **B.** Putatively identified conserved features of the promoter sequence of *Spisula solidissima* PL-I. **C.** Multiple alignment of 3' UTR of PLs from the related bivalve molluscs *Ensis minor* (L41834) and *Mytilus californianus* (see Chapter 5). The polyadenylation signal is indicated by the green bar. The red bar indicates an additional conserved sequence that may be the target of factors responsible for the repression of mRNA translation during spermiogenesis.

duplication events (Fig. 5A). The largest of these sequence duplications are indicated by the blue bars in Figure 5A and consist of two tandem identical repeats of 135 bp each, comprising nucleotides 610-744 and 745-849 (using the numbering of Figure 4). The second duplication is exactly 63 bp, consisting of the nucleotides 421-484 and 501-564. The last event depicted in Figure 5A is 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 (not shown in Fig. 5A) 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 winter flounder (Watson and Davies 1999) and the SNBPs of *Ensis minor* (Bandiera et al. 1995), 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 (Ohno and Becak 1993). 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 *Spisula* PL-I gene were compared with those of histone H1s, histone H5, and protamines. While there is a very distinctive TATA box (Fig. 5B), 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, based on the vertebrate consensus sequences (Oliva and Dixon 1990). 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 (Yiu and Hecht 1997).

While it is likely that the PL-I gene is acted upon by similar factors, the binding sites have diverged significantly enough so as not be recognizable.

.

The 3' UTR of the *Spisula* PL-I gene was compared to those from the closely related bivalves *Ensis minor* (EM-1/6) and *Mytilus californianus* (PL-II/IV), for which cDNA sequences are available (Fig. 5C). All three sequences possess a polyadenylation signal, which is modified from the consensus AATAAA to AATAAT in both *Spisula* and *Mytilus*. 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 (Hecht 1989; Steger 1999). Variations on the consensus polyadenylation signal are thought to be a means of additional regulation (Beaudoing et al. 2000). In addition, there is a well-conserved sequence of 9 nucleotides (GGCCACCCA) directly upstream from the polyadenylation signal (Fig. 5C, red bar). This sequence motif is in the same location as the nucleotide sequences that show a significant extent of similarity among protamine genes and binding of sequence-specific RNA-binding proteins to these regions has been suggested to play an important translational regulatory role (Steger 1999). Two recently identified proteins in mouse , MSY2 and MSY4, bind to a 5'-UCCAUCA-3' consensus sequence located in the 3' UTR of the protamine P1 mRNA (Giorgini et al. 2001). MSY4 in particular has been shown to play an active role in the translational repression of several mRNAs in differentiating spermatids (Giorgini et al. 2002). It is probable that related mRNA translation-repressing factors bind to this consensus sequence in the bivalve molluscs.

The evolution of sperm nuclear basic proteins

Our findings support the notion that the PL proteins of bivalve molluscs have arisen from a common histone H1 ancestor. We have found convincing evidence of a rapid expansion of the amino-terminal tail of a sperm-specific H1, resulting in an extended and regularly repeating structure consisting primarily of arginine, lysine and serine (26.1% + 25.2% + 30.4% = 81.7%). Sequences contained in the PL-I of *Spisula* bear close resemblance to both the H1-like and protamine-like SNBPs of *Ensis minor*, as well as the smaller and more protamine-like proteins PL-III and PL-IV of *Mytilus californianus* (see Chapter 5 for a more detailed examination). The size and number of protamine-like proteins present in the mature sperm can vary considerably between even closely related species, while the final resulting sperm chromatin structure is very condensed in each (Ausió 1986). The H1-like PL proteins of sperm, which co-exist with a significant fraction of germinal histones, may in fact form novel chromatin structures (Lewis and Ausió 2002), as they are evolving towards protamine structure and composition. Further genetic characterization of the SNBPs can capitalize on the significant amounts of genetic and protein-association data that is being amassed today and should further elucidate the evolutionary origins of these rapidly evolving proteins.

Genetic segregation of the sperm nuclear basic proteins of *Mytilus californianus*.

John D. Lewis, Leanne Howe and Juan Ausió§

Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6

§ to whom all correspondence should be addressed

Department of Biochemistry and Microbiology
University of Victoria
Petch Building, Room 220
Victoria, BC
Canada V8N 5Y2
Tel: 250-721-8863
e-mail: jausio@uvic.ca

Lewis et al. (2003) (manuscript in preparation)

ABSTRACT

Protamine-like (PL) proteins are DNA-condensing proteins that replace histones during spermiogenesis. Their composition and structural characteristics are intermediate to those of histones and protamines. Determination of the evolutionary relatedness of these proteins to either histones or protamines is complicated by the high degree of heterogeneity that is observed in both number and size of PL proteins, which varies considerably between even closely related species. We have hypothesized that the heterogeneity of these proteins is a manifestation of the evolutionary steps from a histone H1 ancestor to a protamine. To investigate this hypothesis, we have characterized the genes encoding the PL-II and PL-III proteins of the bivalve mollusc *Mytilus californianus*, and compared them to the recently characterized PL-I gene of another bivalve mollusc, *Spisula solidissima*. Comparative analysis of their promoters revealed key similarities between the genes of *Spisula* PL-I and *Mytilus* PL-II, indicating in all probability that these genes arose from a common gene ancestor. Comparison of the promoters for the two major protamine-like proteins from *Mytilus* (PL-II and PL-III) also showed some similarity, but to a much lower degree. These data strongly suggest that the PL-III gene from *Mytilus* has segregated from the gene encoding PL-II, completing a critical step in the evolution from a histone H1 to a protamine.

Lewis et al. (2003) (manuscript in preparation)

INTRODUCTION

Sperm nuclear chromosomal proteins can be grouped into three major types: histone (H type), protamine (P type), and protamine-like (PL-type) (Ausió 1995). These types of SNBP are widespread within different phylogenetic groups through the animal kingdom (Saperas et al. 1997). Organisms that replace their histones with protamines in the mature sperm are always found at the furthermost tips of the evolutionary branches (Ausió 1999), while the histone type of SNBPs are found in the sperm of more primitive organisms such as the sponge *Neofibularia* (Ausió et al. 1997). Despite suggestions that protamines may have a retroviral origin (Jankowski et al. 1986), we have hypothesized, based on our observations, that protamines have evolved through a PL type intermediate from a primitive histone ancestor (Ausió 1999). The bivalve molluscs present a unique opportunity to study this proposed evolutionary progression, as different species of molluscs possess examples of all three types of SNBP. The giant Pacific oyster, *Crassostrea gigas*, and the bay scallop, *Aequipecten irradians*, have SNBPs of the H type (Ausió 1986). The sperm of the octopus, *Eledone cirrhosa*, and the snail, *Gibbula divaricata*, possess P type SNBPs (Subirana et al. 1973; Gimenez-Bonafe et al. 2002), while the surf clam, *Spisula solidissima*, and the California mussel, *Mytilus californianus*, contain the PL type of SNBPs (Ausió 1986).

The sperm nuclear basic proteins (SNBP) of the bivalve mollusc *Mytilus* consist of three major sperm-specific proteins: PL-II, PL-III, and PL-IV (Ausió 1986) (Fig. 1A, lane Mc and Fig. 1B). These proteins replace much of the histone complement during spermiogenesis but co-exist with approximately 20% of the somatic-type histones in the

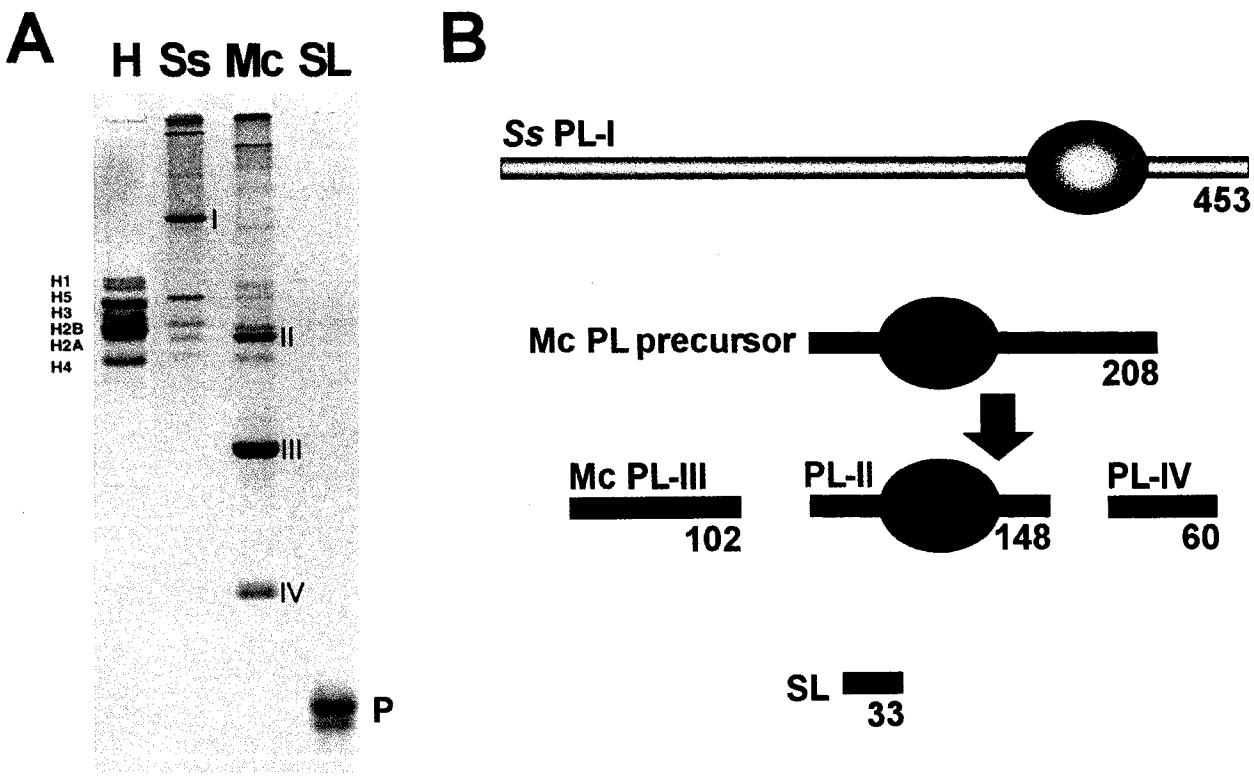


Figure 1. A. Urea (2.5M) - acetic acid (5%), polyacrylamide (15%) gel electrophoresis of 0.4N HCl extract from the sperm of: **Ss**, *Spisula solidissima*. **Mc**, *Mytilus californianus*. **SL**, salmon protamine (salmine). Marker used was chicken erythrocyte histones. Roman numerals refer to the PL protein designations. **B.** Schematic diagram of the PL proteins of *Spisula solidissima* (Ss), *Mytilus californianus* (Mc), and the salmon protamine. Numbers are protein length in amino acids. Post-translational cleavage of Mc precursor to PL-II and PL-IV is shown. Evolution is proposed to have occurred from an H1-like protein (PL-I), through PL proteins (PL-II, PL-III, PL-IV), to a protamine (such as SL).

mature sperm (Ausió 1986; Lewis and Ausió 2002). The PL-II protein of *Mytilus* is a member of the histone H1 family, containing a conserved globular core of 84 amino acid residues that has a high similarity to both the winged helix motif of histone H1 and to the core (also a winged helix motif) of the chromatin-condensing histone H5 of chicken erythrocyte nuclei (Jutglar et al. 1991). The PL-IV protein is very small at 6500 Da and is similar in composition to the lysine-rich carboxyl-terminal tail of histone H1s (Phelan et al. 1974). PL-IV is a product of the post-translational cleavage of a PL-II/PL-IV precursor (Carlos et al. 1993a). PL-III is present in the highest amounts of the three SNBPs in *Mytilus* sperm (Lewis and Ausió 2002). It's highly basic composition, rich in

both lysine (27.5% mol/mol) and arginine (22.5% mol/mol), is intermediate to that of histones and protamines, though, like protamines, it seems to lack any specific secondary structure *in vitro* (Rocchini et al. 1995a).

When the protein sequences of the *Mytilus* SNBPs are compared to the sperm nuclear protein of *Spisula solidissima*, a single large histone H1-like PL-I protein (Fig. 1, lane Ss), some significant similarities are observed. The 60 residue PL-IV is quite similar (48%) in primary structure to the 76 residue carboxyl-terminal tail of *Spisula* PL-I. *Mytilus* PL-II is very similar to the H1 winged helix globular core of PL-I (see Chapter 4, Figure 3A). While PL-I interacts with the sperm DNA as a single protein, PL-II and PL-IV are translated as a single protein and undergo post-translational cleavage at some point before deposition (Carlos et al. 1993a). The PL-III protein contains sequences that are similar to the hexapeptide repeats of the amino-terminal tail of the *Spisula* PL-I. The genetic organization and origin of the PL-III gene, therefore, are of great interest. If PL-III is an independent gene product under the control of an autonomous promoter, it most likely represents the initial genetic segregation of the N-terminal tail of a histone H1-like SNBP towards a protamine-like configuration.

With this goal in mind, we have isolated the genes for both the PL-II and PL-III of *Mytilus californianus*, and performed a comparative analysis with the gene sequence of the *Spisula solidissima* PL-I SNBP.

MATERIALS AND METHODS

Living Organisms

Specimens of *Mytilus californianus* were collected by the author as a part of the Science Venture Student Program from Point No Point (Sooke) on Vancouver Island.

Protein preparation, fractionation, and electrophoresis

Sperm nuclear basic proteins were routinely extracted with 0.4 NHCl following the procedures described previously (Subirana and Colom 1987). Reverse phase HPLC was performed on a 5mm Vydac C18 column (25 x 3 x 0.46 cm) with 0.1% trifluoroacetic acid as eluant with varying acetonitrile gradients (Ausió 1988). Acetic acid (5%)-urea (2.5 M) polyacrylamide gels were prepared as described in (Jutglar et al. 1991).

DNA Extraction

DNA was extracted from gonadal (0.1g) tissue according to the protocol described by Sambrook (Sambrook et al. 1989), with the same modifications as those in Chapter 4.

Genomic library construction and screening

A BamH1-digested genomic library of *Mytilus californianus* was constructed using the Lambda ZAP® II genomic library kit from Stratagene. Plaques were screened using the *Mytilus trossulus* PL-II cDNA (Genbank accession L02876) (Carlos et al. 1993a) as a probe. Plaques were lifted onto 22 cm x 22 cm Hybond N+ membranes. The 612 bp probe was labelled by nick translation according to (Sambrook et al. 1989). The labelled probe was purified from the free label using a microcon® 10 (Amicon). Hybridization

was performed according to the membrane manufacturer's instructions. Membranes were exposed for 24 hours and visualized using the PhosphorImager® System (Molecular Dynamics). Positive clones were subcloned into pBR322 and the DNA was sequenced by the dideoxynucleotide method (Sanger et al. 1977) using a Sequenase 2.0 kit (USB Corp).

Degenerate PCR

Degenerate primers for PCR were created based on the complete amino acid sequence of PL-III from *Mytilus trossulus* (Rocchini et al. 1995a). PCR was performed using the PCRSprint thermal cycler (Interscience) with genomic DNA as template. A touchdown profile was used for the amplification, with the annealing temperature decreasing from 65C to 45C over 20 cycles, followed by 10 cycles at 45C.

Inverse PCR and Genomic Walking

Inverse PCR was carried out as described in (Benkel and Fong 1996), using the primers MYTINV-F (5'- GTCCTCATCACCAAAGAAAAGGAG - 3') and MYTINV-R (5'- CTTTCCCCTTCTTGGGGTCTTGGAAC - 3'). Due to the large amounts of non-specific PCR products amplified using this method, Southern analysis was used to locate positive clones. Genomic walking was performed on *Mytilus* DNA using adaptors, adaptor primers, and protocols based on (Zhang and Gurr 2000). DNA was digested overnight with *Spe*I, *Nhe*I and *Xba*I (New England Biolabs). Adaptors were ligated at 16C for 6 hours, and PCR reactions were carried out using the adaptor-specific PCR primer PP1, and the gene-specific primer MYTWKF1 (5'- CAGCCTCCTCCCCCGGAAAGGCAGC - 3'). A 1/40 dilution was made of the

products of the first reaction, and 1µl of this was added to a second PCR reaction using the nested adaptor-specific PCR primer PP2, and the gene-specific primer MYTWKF2 (5'- CCAAAGAAAAGGAGGTCTGCTGGAAAG - 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 (Zhang and Gurr 2000).

Southern blot analysis of inverse PCR products

Half of each inverse PCR reaction was loaded onto a 1.0 % agarose gel containing ethidium bromide and visualized under UV. The gel was blotted onto Zeta-Probe® GT (BioRad) using the VacuGene® XL Vacuum Blotting System (Pharmacia Biotech) following each manufacturer's instructions for blotting and hybridization. The double-stranded 252 bp insert was labelled by nick translation according to (Sambrook et al. 1989). The labelled probe was purified from the free label using a microcon® 10 (Amicon). Blots were exposed for 24 hours and visualized using the PhosphorImager® System (Molecular Dynamics).

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 manufacturer's instructions, and transformed into TOP10 competent cells (Invitrogen). DNA was sequenced by the dideoxynucleotide method (Sanger et al. 1977) using a Sequenase 2.0 kit (USB Corp).

RESULTS AND DISCUSSION

Mytilus PL-III has a large number of pseudogenes

We utilized a number of techniques in our attempts to isolate the gene for the *Mytilus* PL-III SNBP. Due to the highly repetitive nature of the coding sequence, previous efforts to isolate even a cDNA clone resulted in only an incomplete nucleotide sequence of the PL-III from *Mytilus edulis* (Ruiz Lara et al. 1993). We utilized this cDNA sequence together with protein sequence information from the PL-III from *Mytilus trossulus* (Rocchini et al. 1995a) to generate partially degenerate PCR primers. Using genomic DNA of *Mytilus californianus* as a template, we were successful in the amplification of a 252 bp portion of the PL-III coding region (Fig 2A, lane c). This sequence was then used to design non-degenerate PCR primers for use with an inverse PCR methodology to obtain the

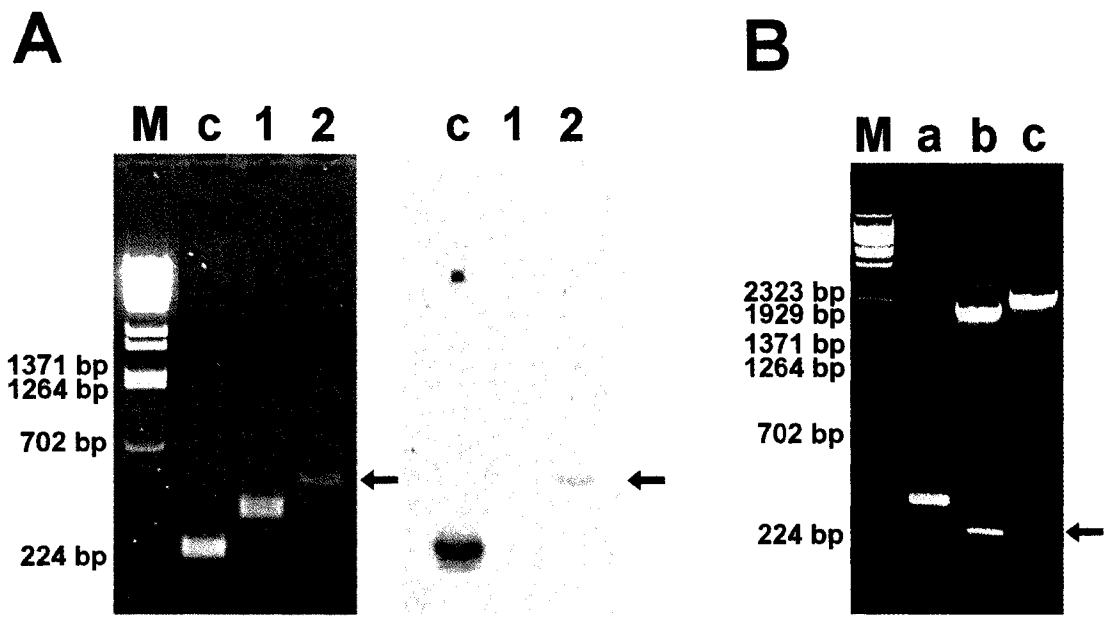


Figure 2. A. left side, 1.0% agarose gel containing inverse PCR results. Lane c, PL-III 252 bp control fragment, lane 1, EcoR1-digested iPCR products, lane 2, DdeI-digested iPCR products. Marker is λBstE marker (Gibco). right side, Southern blot of the gel from right, probed with 252 bp PL-III fragment. Black arrow indicates positive producton both sides. **B.** 1.0% agarose gel of 3' genome walking experiment. Positive fragment denoted by black arrow. Lane a, NheI, lane b, SpeI, lane c, XbaI-digested genomic DNA.

remaining flanking sequences.

There is some evidence that *Mytilus* PL-III genes are multicopy and closely associated with hypervariable regions of the genome (Heath and Hilbish 1998). It was likely due to this that our inverse PCR products contained many non-specific sequences. As a result of this, we utilized Southern analysis to screen our PCR products for those containing our gene of interest. Even with this screening in place, we cloned and sequenced more than ten sequences which contained portions of the *Mytilus* PL-III gene, but had nonsense mutations, deletions, or frame-shifting duplications. These sequences were obviously pseudogenes, and we sequenced at least 8 distinctly different pseudogenes. We were finally able, after many attempts, to identify and clone a 533 bp inverse PCR product (Fig. 2A, lane 2). This clone contained the entire coding region of PL-III in the 5' direction, plus an additional 314 bp of upstream nucleotide sequence. Unfortunately, in the 3' direction, the sequence was truncated approximately 10 nucleotides from the expected STOP codon.

We then decided to utilize a genomic walking technique to amplify the 3' end of the *Mytilus* PL-III gene. This proved to be much more successful, as we cloned and sequenced a 261 bp fragment (Fig. 2B) that contained the elusive remainder of the PL-III coding region and a further 162 bp of downstream nucleotide sequence. In total, the

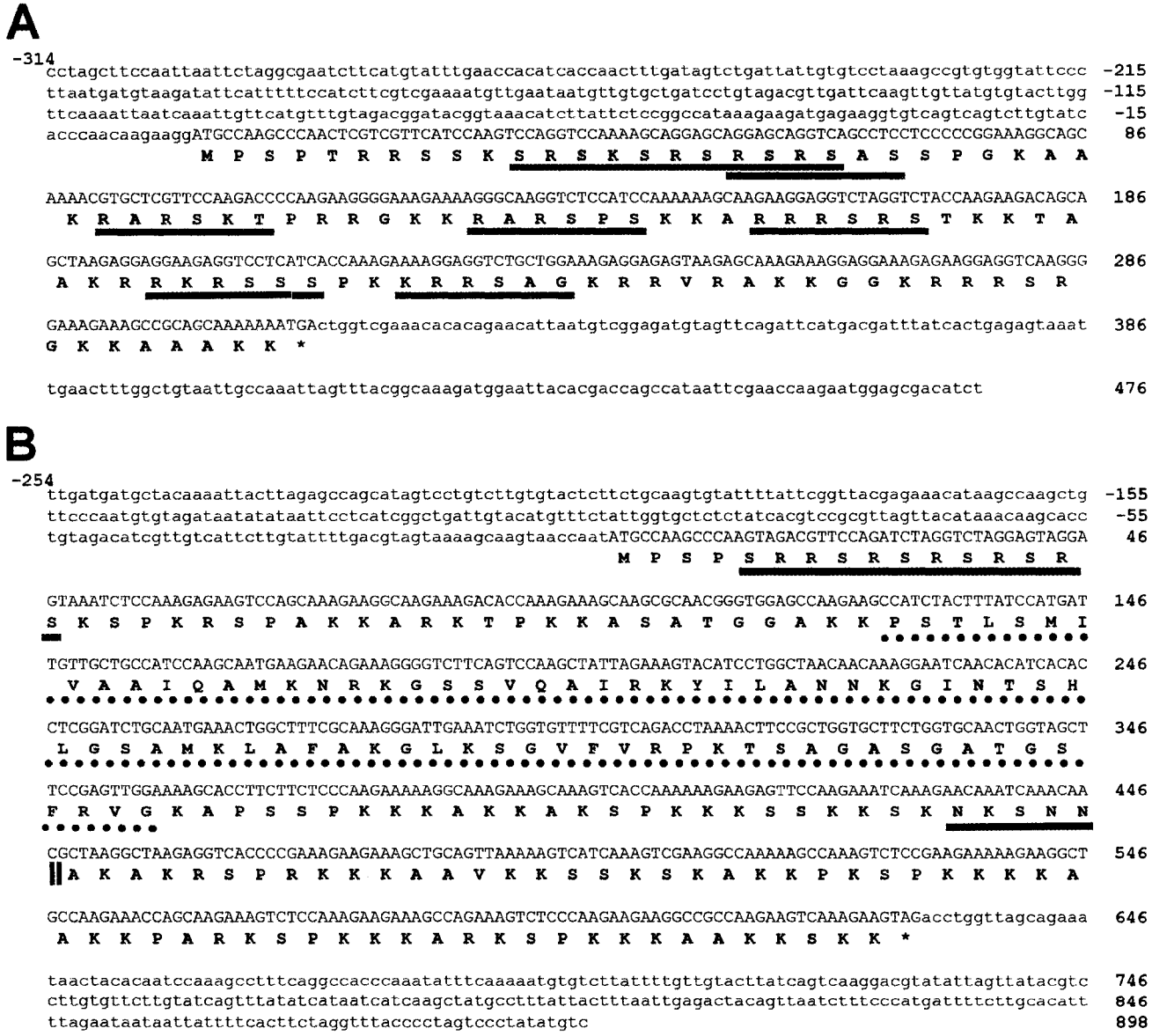


Figure 3. A. Complete sequence of the PL-III gene from *Mytilus californianus*. Green bar denotes conserved RS domain. Red bars indicate conserved hexapeptide repeats. Total length is 790 nt. **B.** Complete sequence of the PL-II/PL-IV gene from *Mytilus californianus*. Green bar denotes conserved RS domain. Blue dots indicates H1-like globular winged helix region. Purple bar indicates conserved pentapeptide involved in post-translational cleavage, this peptide is retained in the mature protein (Carlos et al. 1993a). Double red lines denote cleavage site.

length of the cloned region of the PL-III gene was 790 bp, which encoded a single open reading frame of 102 amino acid residues (Fig. 3A). The protein possesses a number of distinct features, including the presence of a conserved SR repeat domain that is characteristic of many PL proteins and mammalian protamines (Fig 3A, green bar). There are a number of repetitive hexapeptide amino acid sequence motifs, similar to but less conserved than those found in the PL-I of *Spisula solidissima* (see Chapter 4).

Characterization of the PL-II/IV gene of Mytilus

To obtain the gene sequence for the PL-II/IV sperm nuclear basic protein of *Mytilus californianus*, a genomic library was screened using a probe consisting of the previously isolated PL-II/IV cDNA (Carlos et al. 1993a). A positive subclone was obtained consisting of an 1152 bp sequence containing a single open reading frame encoding 208 amino acid residues corresponding to the PL-II/IV precursor protein. There was an additional 254 bp of 5' leading sequence and 274 bp of downstream nucleotide sequence Fig. 3B).

Mytilus PL-II is more similar to Spisula PL-I than to PL-III

The *Mytilus* SNBP genes were aligned with each other and to the PL-I gene of *Spisula solidissima* in order to assess their relatedness and to identify any conserved elements. The sequences were aligned using CLUSTAL X (Thompson et al. 1997) and then optimized by hand. Comparison of the gene structures of *Mytilus* PL-II and *Spisula* PL-I (Fig. 4A) reveals an overall similarity of 47%, with a number of common features. The overall similarity of the upstream sequence of PL-II with that of PL-III from the same

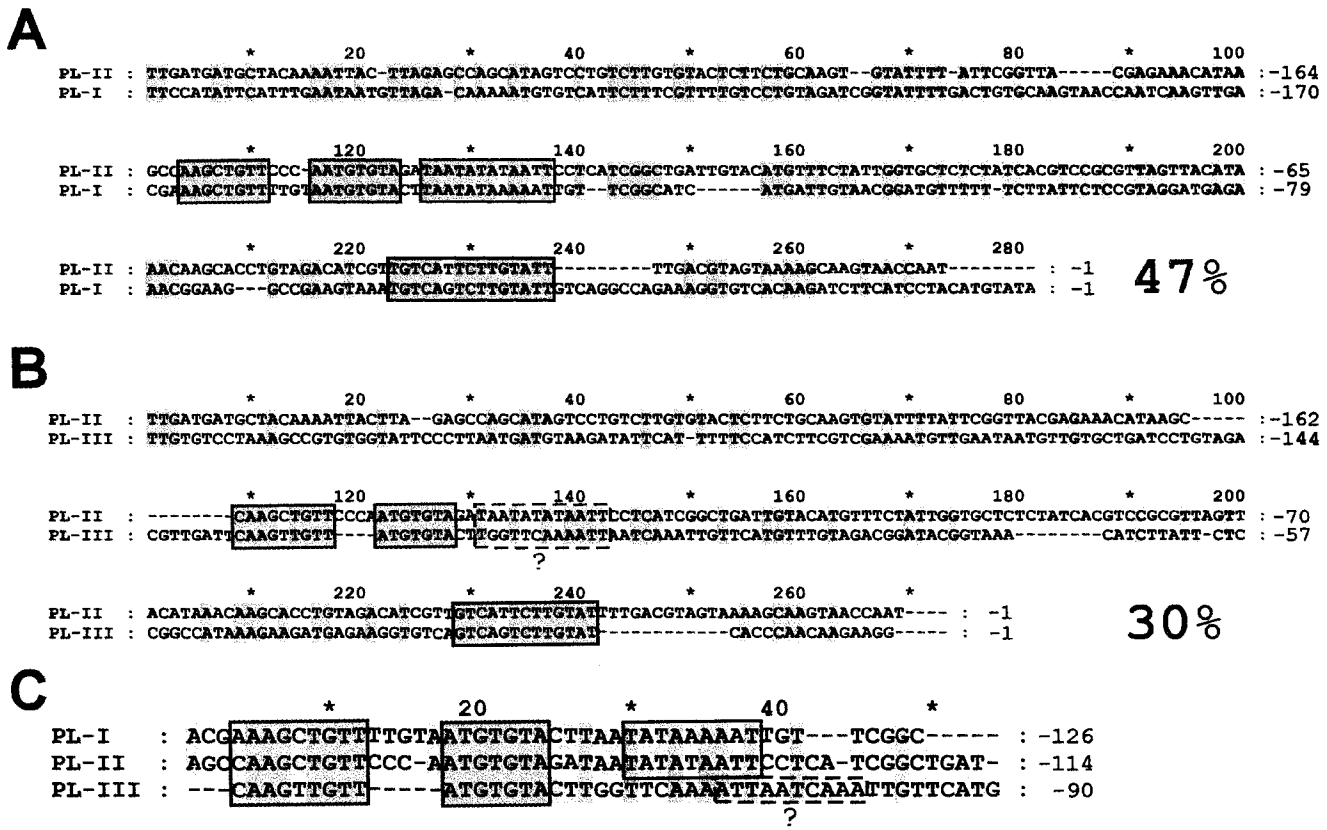


Figure 4. A. Alignment of the upstream regions of the PL-II/IV gene from *Mytilus californianus* with those of the PL-I from *Spisula solidissima*. Conserved putative initiation binding site in blue box, TATA box is indicated by the red box, and conserved putative H4-box-like region is in the green box. Red numbers correspond to overall nucleotide similarity. **B.** Alignment of the upstream regions of the PL-II/IV gene of *Mytilus* with PL-III gene from *Mytilus*. Dashed red box indicates lack of conserved TATA in PL-III gene. **C.** Alignment of all three sequences in the TATA box region, showing uniform conservation of conserved putative H4-box-like site, but not the TATA box itself.

organism (Fig. 4B) is only 30%. Approximately 40 bp upstream from the initiation codons of PL-I and PL-II, there is a conserved region of 14 bp over a 15 bp stretch of DNA that may represent the binding site of a common or related regulatory factor, and is most likely the region of the transcription initiation site. An almost identical similarity is seen in the same region between the PL-II and PL-III gene sequences, indicating that this element may be important for the regulation of SNBPs during spermiogenesis. Another region of similarity between the *Mytilus* PL-II gene and that of *Spisula* PL-I is the putative TATA box region (Fig. 4A and C, red box), where there are 10 of 12 identical nucleotides. The TATA region of the PL-III gene, however, is somewhat puzzling. There

are no typical TATA consensus sequences in any of the upstream sequence, although there is a short AT-rich region of nucleotides a few nucleotides downstream which may comprise a non-consensus TATA box (Fig. 4C). What is particularly interesting is the region directly upstream of the TATA box. Present in all three genes, there are a total of 16 nucleotides that are conserved between PL-I, PL-II, and PL-III. While, the PL-III conserved region is contiguous, the PL-I region is separated by an intervening five nucleotides, and the PL-II gene has four. In vertebrate H1 genes, this region contains the CAAT box, which is a well conserved binding site for tissue-specific transcription enhancing factors. In the vertebrate differentiation-specific H1 genes, the H1 genes of sea urchin, and chicken H5, there is a well-conserved H4 box in this region. The consensus H4 box sequence, however, bears no similarity to the putative conserved elements of *Spisula* and *Mytilus* SNBPs. It is very likely, however, that this element is the target of factors that either activate the transcription of sperm-specific proteins, or repress transcription until the appropriate time during spermiogenesis.

The evolution of the SNBPs of bivalve molluscs

These results of the comparison of the genes coding for sperm nuclear basic proteins in *Spisula* and *Mytilus* are compelling. While all three genes seem to share certain conserved elements, the PL-II gene of *Mytilus californianus* is much more similar to the PL-I gene of *Spisula solidissima* that to that of PL-III of the same organism. These results show also that the PL-III gene product is under distinct and autonomous control of a similar but separate gene. This result is not surprising, given the fact that PL-III is temporally expressed at a later point than PL-II and PL-IV (Bloch 1966). This analysis

supports very well the notion that H1-like sperm nuclear basic proteins are evolving towards protamines.

Lewis et al. (2003) (manuscript in preparation)

Protamine-like proteins: evidence for a novel chromatin structure.

John D. Lewis and Juan Ausió§

Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6

§ to whom all correspondence should be addressed

Department of Biochemistry and Microbiology
University of Victoria
Petch Building, Room 220
Victoria, BC
Canada V8N 5Y2
Tel: 250-721-8863
e-mail: jausio@uvic.ca

Lewis and Ausió (2002) Biochem Cell Biol **80**:353-61.

Abstract

Protamine-like proteins are DNA condensing proteins that replace somatic histones during spermatogenesis. Their composition suggests a function intermediate to that of histones and protamines. While these proteins have been well characterized at the chemical level in a large number of species, particularly in marine invertebrates, little is known about the specific structures arising from their interaction with DNA. Speculation concerning chromatin structure is complicated by the high degree of heterogeneity in both number and size of these proteins, which can vary considerably between even closely related species. After careful examination and comparison of the protein sequences available to date for the protamine-like proteins, we propose a model for a novel chromatin structure in the sperm of these organisms that is mediated by somatic histones, which are frequently found associated with these proteins. This structure supports the concept that the protamine-like proteins may represent various evolutionary steps between a sperm-specific histone H1 precursor and true protamines. Potential post-translational modifications and the control of PL protein expression and deposition are also discussed.

Lewis and Ausió (2002) Biochem Cell Biol **80**:353-61.

Introduction

The DNA of eukaryotic organisms is associated with a number of basic proteins in a complex macromolecular assembly, termed chromatin. At the structural level, the most important function of this assembly is to compact the lengthy DNA molecule inside the limited available nuclear space. In somatic cells, chromatin is a dynamic structure as DNA must be accessible for replication, repair and transcription. DNA is packaged in these cells via an assembly of core histones, linker histones and non-histone proteins. During spermiogenesis, however, there is a dramatic remodelling of chromatin that is characterized by considerable cellular morphological change, concurrent with modifications in the nature and content of the nuclear basic proteins. The chromatin of the mature sperm adopts in the majority of instances a highly compacted state in which gene expression is completely repressed.

An accumulation of data from recent studies involving these sperm nuclear basic proteins (SNBPs) has facilitated their classification into three main categories: histone type (H type), protamine type (P type), and protamine-like type (PL type) (Ausió 1995). The histone type is characterized by a persistence of somatic histones or by an appearance of sperm-specific histones that are compositionally and structurally related to the histones found in the nuclei of somatic cells. This group consists mainly of sperm-specific variants of H1 and H2B, such as spH1 and spH2B from the sperm of echinoderms (Zalenskaya et al. 1980; Poccia and Green 1992), and the mammalian histone H1t (Seyedin and Kistler 1980). The second group of SNBPs, the protamine type, consists of relatively small (generally $4000 \leq M_r \leq 10000$), arginine-rich (Arg $\geq 30\%$), highly basic proteins. During spermiogenesis, these proteins replace the majority of the

histone complement, either directly or subsequent to the appearance of transition proteins and/or protamine precursors. This group includes the protamines of mammals, marsupials, birds, fish and reptiles (for a review, see (Oliva and Dixon 1991)). The last group, the protamine-like type, are structurally quite heterogeneous, while maintaining a very consistent chemical composition; one intermediate to that of protamines and histones. While initially described in the bivalve molluscs, they are pervasive across the animal kingdom, having been identified in such phylogenetically diverse organisms as Cnidaria (Rocchini et al. 1995b; Rocchini et al. 1996), chordates (Saperas et al. 1992), and vertebrates (Saperas et al. 1994). There is a surprising variability in the size and number of different PL components present in the mature sperm of even closely related organisms, especially considering that the common function of these proteins is chromatin condensation. In molluscs, the PL proteins are generally sub-classified into four basic categories according to their relative electrophoretic mobilities; PL-I, PL-II, PL-III, and PL-IV (Ausió 1986). Like protamines, PL proteins are highly basic, with an arginine + lysine content of at least 35-50 mol%, with the notable presence of cysteine in a number of them (Zhang et al. 1999). They can vary in molecular weight from 6500 in the case of PL-IV of the blue mussel, *Mytilus californianus*, up to 200000 for the SNBPs of winter flounder (Watson and Davies 1998). For the majority of organisms studied thus far, PL proteins coexist in the mature sperm with a full histone complement, amounting to approximately 30-40% of the total SNBPs.

The direct analysis and visualization of the derivative structures arising from the interaction of sperm DNA with these proteins has proven difficult due to their highly packed nature, which results in chromatin that is relatively insoluble and generally

refractory to detailed conventional analysis by electron microscopy, analytical ultracentrifugation or nuclease digestion. Studies utilizing electron microscopy to examine the sperm of various bivalve molluscs have revealed the presence of densely packed chromatin fibers with a diameter in the range of 25 to 50 nm (Casas et al. 1993) depending on the protein composition. Early micrococcal nuclease digestion experiments of the sperm chromatin in these organisms has suggested the existence of a dual chromatin organization, whereby some of the genome appears to retain the somatic nucleosomal organization, and the remainder is highly compacted as a result of the association of the PL component (Azorin et al. 1983; Ausió and van Holde 1987), which binds in a highly cooperative fashion (Libertini et al. 1988). In spite of these observations, the detailed organization of the nucleoprotein structures arising from the interaction of protamine-like proteins with DNA has not been elucidated.

PL proteins are highly heterogeneous members of the histone H1 family

The high degree of heterogeneity of the protamine-like proteins has provided conceptually challenging questions concerning their individual roles in chromatin condensation during spermatogenesis. The isolation and molecular characterization of these proteins in a growing number of organisms has, however, begun to provide some valuable insight. We originally attempted to elucidate the structure of the PL-I of *Spisula solidissima*. The sperm of this bivalve mollusc possess a single, large PL protein that coexists with a fraction of somatic-like histones (Fig. 1A) (Ausió and Subirana 1982b). *Spisula* PL-I, with a length of over 450 amino acids and a molecular mass of $M_r \cong 50000$ (Ausió and Lewis, unpublished [see Chapter 2]), is much larger than any protamine or

histone previously described. 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 (Ausió et al. 1987) (Fig. 1B). Sequence analysis of the globular portion of PL-I revealed that this region shares high sequence similarity with the winged helix DNA-binding motif found in histones H1 and H5 (Ausió et al. 1987), and a number of transcription factors such as HNF3 (Cirillo et al. 1998) and the FOXn1 family of oncogenes (Schlake et al. 2000) (Fig. 2). In particular, the highest sequence similarity is found with histone H5, which is a member of the histone H1 family, and is present in the nucleated erythrocytes of certain vertebrates (Neelin 1968). Like spermatozoa, erythrocytes are terminally differentiated cells containing chromatin that is highly condensed.

The protamine-like protein complement of other related organisms displays a greater complexity. The SNBPs of the mussel, *Mytilus sp.*, consist of three major components: PL-II (M_r = 15600), PL-III (M_r = 11300), and PL-IV (M_r = 6500) (Fig. 1). Structural analysis of PL-II (Casas et al. 1993) has revealed that it too contains a globular core with significant sequence similarity to the H1 winged helix motif (Fig. 2), while PL-III (Rocchini et al. 1995a) and PL-IV are relatively smaller arginine/lysine-rich proteins. Sequence comparison from unpublished results from our lab between *Mytilus* PLs and the PL-I from *Spisula* indicates that PL-IV is most related to the C terminal tail of PL-I, and in fact has also been previously related to the C terminal tail of histone H1 (Phelan et al. 1974). PL-III shares sequence similarity with the N terminal region of *Spisula* PL-I (Fig. 1B). It was not surprising, therefore, that it was determined that PL-II and PL-IV are formed as the result of the post-translational cleavage of a common (PL-I) precursor

(Carlos et al. 1993a). PL-III, on the other hand, appears to be a completely independent gene product (Lewis and Ausió, unpublished [see Chapter 4]). Importantly, while the molar stoichiometry of PL-II to PL-IV appears to be 1:1, there seems to be a three fold

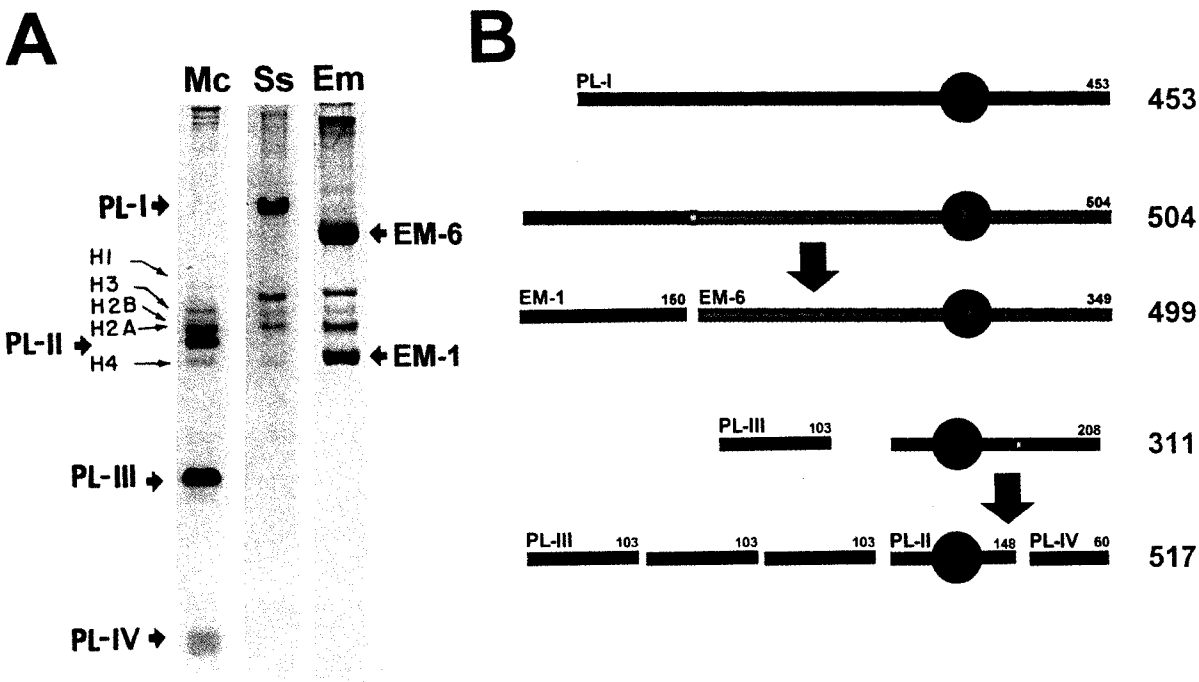


Figure 1. A. Acetic acid (5%)- urea (2.5 M) PAGE (15% acrylamide : 0.1% bisacrylamide) of SNBPs of the bivalve mollusks discussed in this paper: **Ss**, *Spisula solidissima*; **Mc**, *Mytilus californianus*; **Em**, *Ensis minor*. The direction of electrophoresis is from top (+) to bottom (-) on this and subsequent Figures. **B.** Comparison of the primary structures and post-translational modification of SNBPs from *Spisula solidissima*, *Mytilus californianus* and *Ensis minor*. Small black numbers indicate the length in amino acids of each protein. The larger red numbers on the right indicate the total length in amino acids of SNBPs in each organism. The small yellow boxes in the PLs of *Mytilus* and *Ensis* are sites of post-translational cleavage of the protein precursor, both distinguished by the sequence NKSNN at this cleavage point. While it appears that the total SNBP length in *Spisula* and *Ensis* is comparable at 453 and 499 amino acids respectively, the total length of *Mytilus* SNBPs is only 311. Examination of the relative amounts of protein present in the first lane in **A.** suggests that PL-III is present in stoichiometrically higher quantities than the other SNBPs. The presence of three PL-III molecules per molecule of PL-II and PL-IV would result in a total length of 517, a number which is quite comparable to that found in both *Spisula* and *Ensis*.

greater molar abundance of PL-III in the mature sperm of *Mytilus* (see Table 1).

The sperm of the razor clam, *Ensis minor*, contains two major PL components: a PL-II (EM-6) of M_r = 23600 which possesses a conserved H1 winged helix domain (Fig. 2), and a smaller PL-III (EM-1) (Bandiera et al. 1995) of M_r = 16800 (Fig. 1). These

proteins have also been shown to be the result of post-translational cleavage of a common PL-I precursor. In this case, however, it is the N terminal tail that is released, evidenced by the considerable sequence similarity of EM-1 to both the PL-III of *Mytilus* and the N terminal tail of *Spisula* PL-I. While the mechanism of the post-translational cleavage in both of these cases has not been elucidated, the signature pentapeptide NKSNN is retained in both at the C terminal end of each cleavage site (Carlos et al. 1993b; Bandiera et al. 1995).

It has been suggested that the heterogeneity of these proteins may represent an attempt at segregation driven by the evolutionary trend from an H1 precursor to the protamine type of SNBPs (Ausió 1999; Kasinsky et al. 2001). The fact that the PL-III of *Mytilus* appears to be a completely separate gene and is expressed in greater amounts relative to the other PLs supports this view. There is also evidence that PL-III may be temporally expressed at a later point than PL-II and PL-IV (Bloch 1966).

PL proteins contain multiple sites of phosphorylation

In the sperm of sea urchin, a number of well-conserved sites in the N terminal regions of spH1 and spH2B are phosphorylated at the beginning and dephosphorylated in the final stage of spermiogenesis, concurrent with extensive chromatin condensation (Poccia and Green 1992). These sites are then rapidly phosphorylated again immediately following fertilization. Phosphorylation effectively increases the negative charge, and is likely an important mechanism by which the affinity of these proteins to DNA can be precisely and temporally modulated during spermiogenesis. The heavily phosphorylated sites consist of tetrapeptide repeats with two basic residues flanking Ser-Pro residues

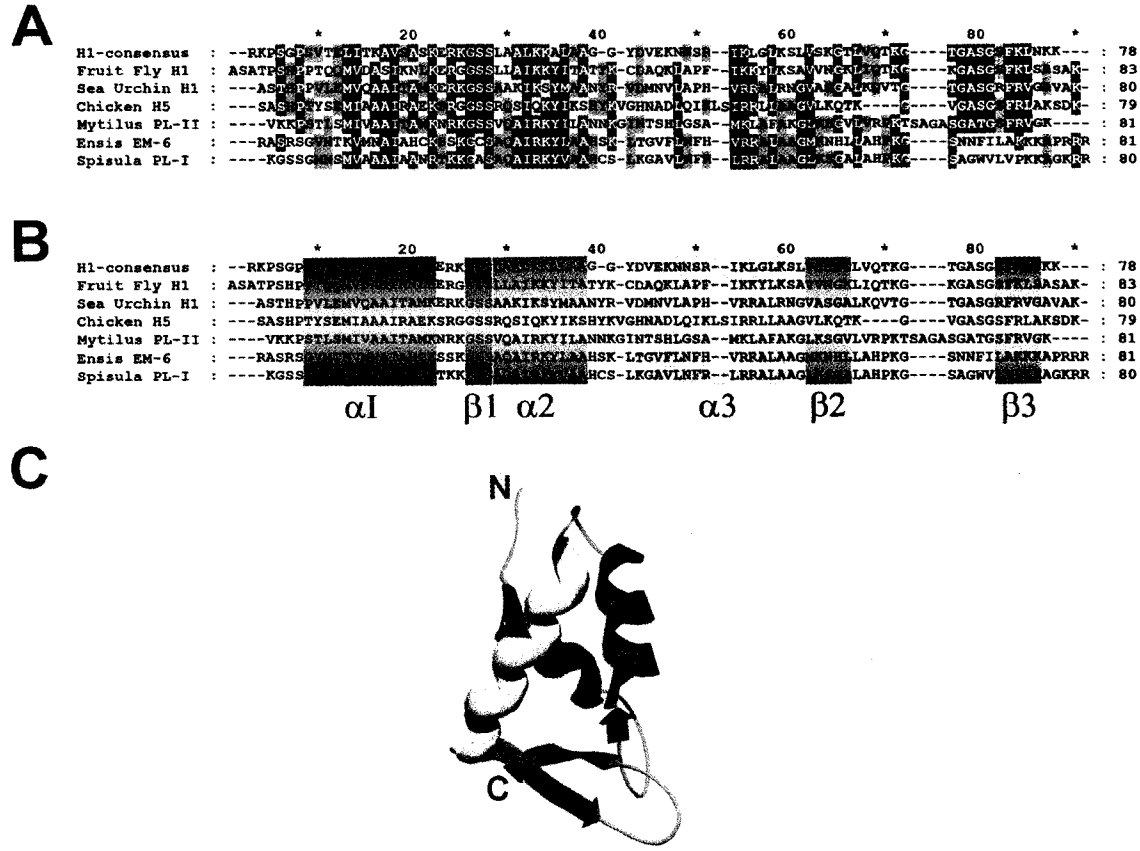


Figure 2. A. Sequence alignment of globular core regions of representative H1s and PL-I/PL-II SNBPs, generated with Clustal X (Thompson et al. 1997). Shading indicates the level of conservation between the sequences (dark blue, 100% conservation; medium blue, 75% conservation; light blue, 50% conservation). There is quite a significant level of conservation between the PL-I/PL-II globular regions and those of the winged helix motif H1, suggesting that the globular core of these proteins assumes a winged helix motif. The H1-consensus sequence is as determined by (Wells and Brown 1991). **B.** Same alignment as in **A.**, but overlaid with the secondary structures determined for the globular core of chicken erythrocyte H5. α1-α3; alpha helices 1-3 (cyan, maroon and orange respectively), β1-β2; beta sheet regions 1 and 2 (light blue). Comparing with alignment in **A.**, note the high level of conservation of amino acid sequences contained in conserved secondary structures. While helices 1 and 2 are very well conserved, it appears that there is some variability in alpha helix 3 and the beta sheet region. **C.** Three dimensional rendering of the globular core of chicken erythrocyte H5 with coordinates as determined by (Ramakrishnan et al. 1993). Secondary structures are colour-coded to match the highlighted sequences in **B**.

('SPKK' motifs), which resemble target sequences for cdc2/cdk2 kinase. It has been shown recently, however, that while a cdc2/cdk2-dependent protein kinase cascade is critical for histone phosphorylation and chromatin condensation in vivo, the cdc2/cdk2 kinase itself may only play an indirect role (Shimada et al. 1998). These and similar

motifs are found in the majority of PL proteins examined to date, and are well represented in the N terminal region of *Spisula* PL-I, the C terminal end of *Mytilus* PL-II and throughout the PL-IV protein of *Mytilus*. The high molecular weight basic nuclear proteins (HM_rBNPs) of winter flounder are heavily phosphorylated in regions which contain variations of the consensus X-S-X-S-P (where X is arginine or lysine) (Kennedy and Davies 1985). These consensus sequences are also found in the N terminal tails of the PLs of *Spisula*, *Ensis*, and *Mytilus*.

Another important potential target for phosphorylation is a domain containing a number of RS repeats found at the N terminus of many protamine-like proteins, including the PL-I of *Spisula*, PL-II and PL-III of *Mytilus*, and EM-1 of *Ensis*. These sequences are found only in a small subset of all proteins; the best characterized being the superfamily of RS domain-containing splicing factors (for a review, see (Fu 1995)). Similar motifs are also found in sperm nuclear basic proteins of the P type, including mammalian protamine 1 (Domenjoud et al. 1990) and the alligator Al-I and Al-II protamines (Hunt et al. 1996). It has recently been shown that protamine 1 is phosphorylated in testis by the SR protein-specific kinase 1 (SRPK1) (Papoutsopoulou et al. 1999), a kinase that has mainly been implicated in the phosphorylation of the SR motifs of splicing factors (Gui et al. 1994). SR proteins are also specifically phosphorylated by topoisomerase I (Rossi et al. 1996), an activity which is apparently functionally independent from its DNA cleavage (relaxation) ability (Labourier et al. 1999). The chromatin remodelling events during the early stages of spermiogenesis require the activity of topoisomerase I to maintain sufficient DNA relaxation (Cobb et al. 1997), so it is particularly interesting that topoisomerase I may also be involved in the regulation of PL proteins by phosphorylation

of their SR domains. While it is almost certain that phosphorylation/dephosphorylation of the protamine-like proteins plays an important role in their involvement in chromatin condensation, the specifics of the underlying mechanism and the precise components involved remain to be determined.

What does the structure of PLs say about their function?

While protamine-like proteins appear to be evolutionarily related to members of the histone H1 protein family, there are a number of significant differences between the primary structure of PLs and histone H1. Histone H1 has been described as the 'lysine-rich' histone, as the first linker histones to be purified and sequenced were found to be greatly enriched in lysine, particularly in the C terminal tail (Cole 1984). In fact, the C terminal tail of most typical somatic H1s consists of about 90% lysine, alanine and proline. The particular distribution of these amino acids in the C-terminus and the resulting positive charge density confers to histone H1 the unique ability to bind to the linker DNA (Subirana 1990). This binding is presumably achieved mainly by the screening of negative charges on the linker DNA connecting adjacent nucleosomes. Relative to the typical H1 linker histone (2-6 mol% arginine, 25-30 mol% lysine), protamine-like proteins are more highly enriched in arginine (8-30 mol%) as well as lysine (22-30 mol%). This trend to higher arginine content leads to increased charge density and greatly facilitates the compaction of sperm DNA. PL proteins are consequently very efficient at condensing DNA, and as a result, decondensation (during fertilization) requires the assistance of highly specific proteins from the egg, such as nucleoplasmin (Rice et al. 1995). The increased incidence of arginine in these proteins

may also promote their binding to DNA in the minor groove (Chikhirzhina et al. 1998), in addition to the major groove, which may serve to enhance chromatin compaction.

Protamine-like proteins also typically contain repetitive sequence motifs, particularly in the larger PL-I proteins found in *Spisula solidissima* (Fig. 1) and *Ensis minor* (Bandiera et al. 1995), and to an even greater extent in the high molecular weight basic nuclear proteins of winter flounder (*Pleuronectes sp.*) (Watson and Davies 1998). For example, in the PL-I of *Spisula*, there are 20 octapeptide repeats of the sequence KRSASKKR spanning a 244 amino acid stretch of the N terminal tail (Ausió and Lewis, unpublished [see Chapter 2]). Identical repeats are also observed in the PLs (EM-1, EM-6) of *Ensis*. While there is no physical data concerning these repeated sequences when bound to DNA, circular dichroism spectrophotometric analysis of these proteins in solution suggests the formation of an extended conformation (random coil) as opposed to alpha helix (Ausió and Subirana 1982a; Giancotti et al. 1983; Ausió et al. 1987). The sequence repeats from both *Spisula* and *Ensis* would, in an extended conformation, adopt a structure where the alternating serine residues would be placed on the same side of the chain, flanked by the basic side chains of lysine and arginine. This would be an ideal conformation to adequately spread out the positive charge density over a long stretch of DNA, effectively screening its charge to attain a high level of chromatin compaction.

If the presence of an adequate number of basic residues is sufficient for DNA charge screening and chromatin compaction (as in P type SNBPs), why do protamine-like proteins contain a globular H1 winged helix domain? The winged helix proteins comprise a family of DNA binding proteins whose DNA recognition helices are related in structure and function to the helix-turn-helix that was originally identified in bacterial

transcriptional regulatory proteins (Sauer et al. 1982). As first seen in the structures of histone H5 (GH5) (Ramakrishnan et al. 1993), the winged helix structure adopts a mixed α/β fold containing three α-helices and three β-strands (Fig. 2C). Recent studies utilizing a site-specific protein-DNA photo-crosslinking method with H5 on mixed sequence nucleosomes have determined that the winged helix domain forms a bridge between one

Table 1. Analysis of the chromatograms obtained by reversed phase HPLC and ion exchange chromatography of SNBPs from *Mytilus* and *Spisula*

	Area%[a]	ε_{230}[b]	$M_r(10^{-4})$	Area%/ε/M_r	ratio[g]
Mytilus californianus					
Histone octamers + H1	66.0	3.86[c]	13.0	1.3	**1.0**
PL-II + PL-IV	17.5	2.00[d]	2.1	4.0	**3.0**
PL-III	16.5	1.40[e]	1.1	11.2	**8.6**
Spisula solidissima					
Histone octamers + H1	54.4	3.86	13.0	1.1	**1.0**
PL-I	45.6	2.50[f]	5.0	3.6	**3.3**

[a] %Area corresponds to the average of several chromatograms such as those depicted in Figure 3A and B.
[b] extinction coefficients are expressed in $cm^2 \cdot mg^{-1}$
[c] this extinction coefficient was calculated from the extinction coefficient of the histone octamer (ε_{230}=4.2) (Stein 1979) and that of histone H1 (ε_{230}=2.0) (Camerini-Otero et al. 1976; Rocchini et al. 1995a) on the basis of 1 mol H1 per mol histone octamer using the molecular weights M_r shown in this table
[d] ε_{230} of histone H1 from (Camerini-Otero et al. 1976; Rocchini et al. 1995a)
[e] ε_{230} of *Mytilus* PL-III from (Ausió 1980)
[f] ε_{230} of *Spisula* PL-I (unpublished results)
[g] values normalized to 1 mol of histone octamer

terminus of chromatosomal DNA and the DNA in the vicinity of the dyad axis of symmetry of the core particle (Zhou et al. 1998). The winged helix of H5 possesses two distinct DNA-binding sites on opposing sides, placing the basic carboxy-terminal region of the globular domain in a position from which it could simultaneously bind the nucleosome-linking DNA strands that exit and enter the nucleosome. It is reasonable to envisage that protamine-like proteins are able to specifically interact with the DNA at the dyad axis of retained nucleosomes via their winged helix domain. Nucleosomes, due to

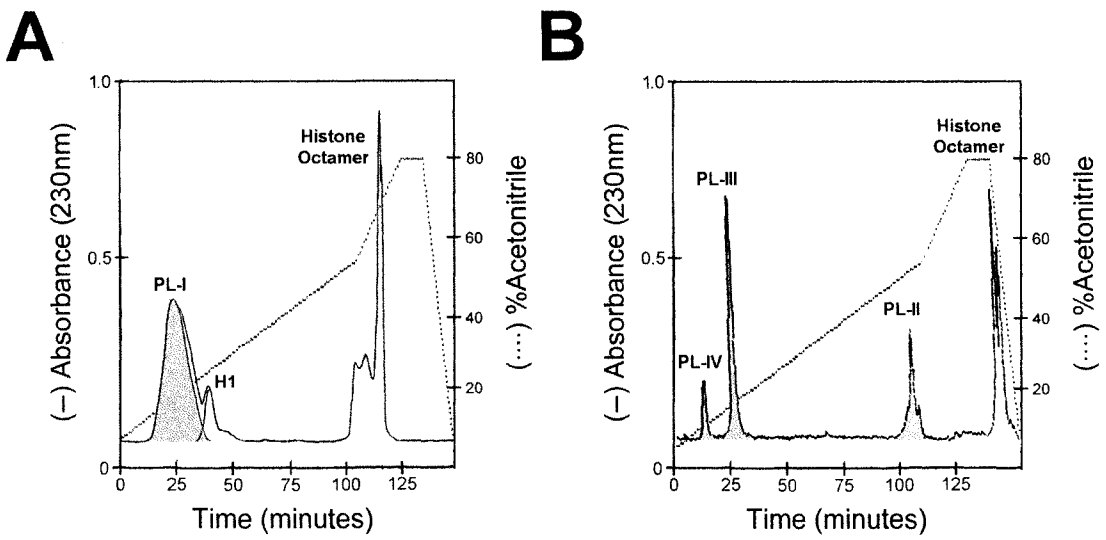


Figure 3. A. Reverse-phase HPLC fractionation of *Spisula* SNBPs (Vydac C_{18} column (25 x 0.46 cm) eluted with an acetonitrile gradient in the presence of 0.1% trifluoroacetic acid at a flow rate of 1 ml/min). Proteins were detected at 230 nm. The PL-I of *Spisula* (gray) co-eluted with histone H1 (white, minor peak), these peaks were manually deconvoluted. Histone octamers elute as a single large (white) peak. Areas below the trace and above the baseline were integrated and listed in Table 1. The data in Table 1 represents the average of several HPLC traces for each species. **B.** RP-HPLC fractionation of *Mytilus* SNBPs (as in **A.**). Peaks corresponding to SNBPs are shaded in gray. In this profile, H1 eluted with the histone octamers.

their pseudo-dyad axis of symmetry, potentially provide two possible binding sites for winged helix-containing proteins. In fact, there is experimental evidence suggesting that this can be the case (Nelson et al. 1979). Furthermore, in chicken erythrocytes where the molar stoichiometry of linker histones per nucleosome is 1.3:1 (0.9 for histone H5, 0.4 for histone H1) (Bates and Thomas 1981), it is likely that nucleosomes associated with both H1 and H5 could be found (Segers et al. 1991).

Model for a novel chromatin structure

With the extensive inter-species heterogeneity that is observed across the range of protamine-like proteins, the question remains: How do these organisms achieve a similar

apparent sperm chromatin structure with such different sizes and numbers of PLs? The size and number of protamine-like proteins present in the mature sperm can vary considerably between species (Fig. 1). Upon closer examination of the relative mass and stoichiometry of these proteins to one another and to the somatic-like histones in the sperm (see Table 1), a consistent pattern becomes apparent.

In contrast to organisms that utilize protamines (P type) to condense the sperm nuclear DNA, those that utilize protamine-like (PL type) proteins retain a significant amount of somatic-type nucleosomes following spermiogenesis (Ausió and van Holde 1987) (Fig. 1A and 3A, B). What is the origin of these nucleosomes in mature sperm? Are they synthesized *de novo* during spermatogenesis or are they remaining from their somatic cell precursors? To account for the limited experimental information available we would like to propose a hypothetical model in which somatic histone expression is shut off during meiosis II, resulting in an approximate 50% depletion of somatic nucleosomes, which disperse themselves evenly over the haploid genome of the maturing sperm. In this situation, the linker DNA length would be greatly extended, from an average of 60 bp found in somatic cells to an average of about 260 bp. In fact there is no experimental evidence to suggest a histone displacement mechanism by protamine-like proteins such as that observed with protamines (Oliva and Dixon 1991). In order to sufficiently compact the chromatin, therefore, protamine-like proteins must direct themselves to this extended linker and successfully screen the resultant negative charge arising from the phosphate moieties of the DNA backbone.

It seems likely that the winged helix domain of PL-I in *Spisula* and of PL-II/PL-IV precursor in *Mytilus* directs these molecules to the vicinity of the histone H1 which is

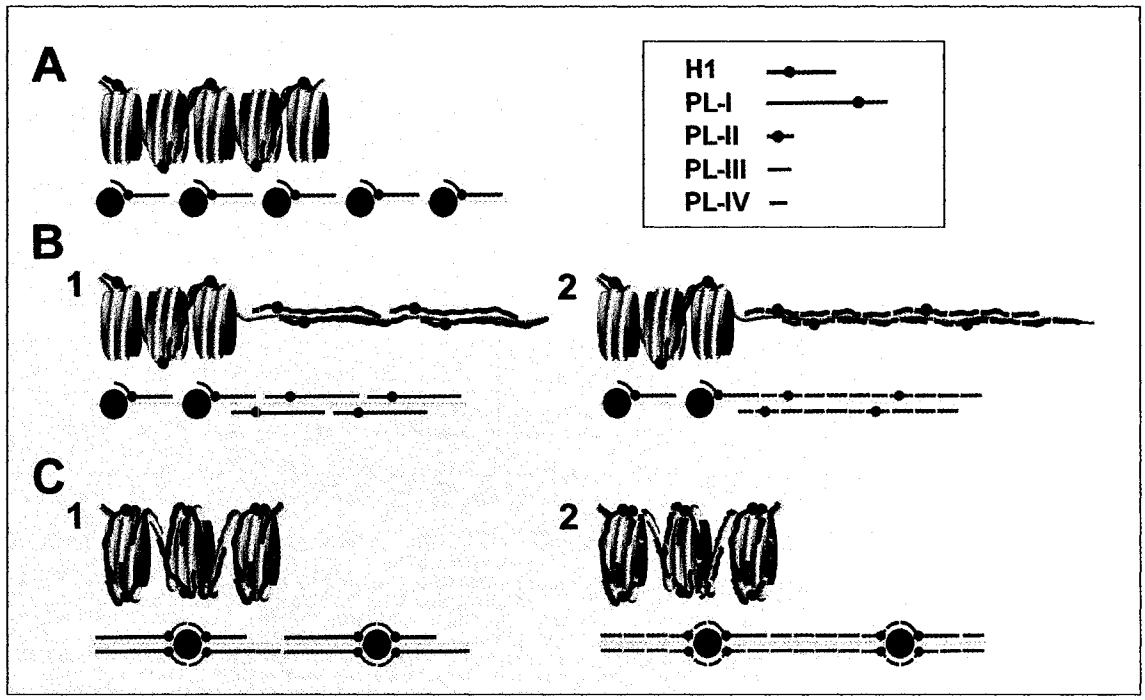


Figure 4. A. Above: three dimensional representation of the structure of somatic nucleosomes with histone H1. Histone octamer represented as purple cylinders, H1 is dark green, DNA is in orange. Model adapted from (Green 2001). Below: Schematic representation depicting the position of the nucleosomes and H1 on linearized DNA. Histone octamer represented as purple circles, H1 is dark green and DNA is orange. Linker DNA is approx. 60bp in length. **B. 1.** Hypothetical sperm chromatin structure of *Spisula solidissima* based on (Ausió and van Holde 1987). Spisula PL-I protein is in light purple. Normal nucleosomal structure is maintained in distinct regions while PL-I deposits in tandem linearly along naked DNA. **2.** Equivalent sperm chromatin structure in *Mytilus californianus*. Referring to Fig. 1, Fig. 3 and Table 1, there are most likely 3 PL-III molecules per molecule of PL-II/PL-IV. **C. 1.** Hypothetical sperm chromatin structure proposed in this paper for *Spisula solidissima*. Linker DNA is approx. 260bp in length. Winged helix of PL-I coordinates with winged helix of H1, allowing the PL-I N-terminal tail to neutralize the charge of significant length of linker DNA, allowing this DNA to further wrap around the nucleosome. The two extra coils generated in this way would create another 2 potential binding sites for PL-I, and result in an extremely compact but regular structure. **2.** Equivalent sperm chromatin structure in *Mytilus californianus*. Referring to Fig. 1, Fig. 3 and Table 1, there are most likely 3 PL-III molecules per molecule of PL-II/PL-IV, permitting a sperm chromatin structure very similar to that in *Spisula*.

associated with the remnant nucleosomes, bringing them to occupy the putative second binding site, while their long N- and C-terminal tails partially neutralize the charges on the adjacent long linker DNA regions. Association of protamines and highly charged basic proteins/peptides (including histone H1) has been shown to lead to the formation of highly organized toroidal structures as a result of the charge neutralization (Clark and Thomas 1988; Garcia-Ramirez and Subirana 1994). In the case PL-I/PL-II, this could

lead to a further wrapping of the linker DNA about the pre-existing nucleosome structure (see Fig. 4C), potentially creating two additional winged helix binding sites at the opposite site of the particle. In this way, four winged helix-containing proteins could be bound by each nucleosome, creating a highly condensed particle in which most of the charges would remain almost completely neutralized, as it occurs with protamine-DNA complexes (Oliva and Dixon 1991). The two extra turns of DNA around the nucleosome particle (Fig. 4C) would be stabilized by the charge neutralization of DNA provided by the highly charged N and C terminal PL tails.

Careful inspection and analysis of the chromatographic elution profiles of SNBPs from *Mytilus* and *Spisula* (see Fig. 3 and Table 1) reveals that the PL/histone octamer molar ratio is quite similar between these two species. Despite the approximate nature of this analysis it appears that there are about three PL-I or PL-II/PL-IV molecules per nucleosome in addition to the nucleosomal histone H1. This would support the model described above and depicted in Figure 4C. It is important to note that in the case of *Mytilus*, the existence of PL-III appears to account for the shorter N-terminal domain of PL-II/PL-IV. The stoichiometry of PL-III to PL-II in this analysis indicates that there are approximately three PL-III molecules per PL-II/PL-IV. This result is consistent with the observation in Figure 1 that all the PL-proteins observed in the bivalve molluscs analyzed to date seem to span a 500 amino acid length and the model shown in Figure 4C. This hypothesis also fits well with the determined protein/DNA ratio (0.83) of *Spisula* sperm chromatin (Ausió and van Holde 1987). This ratio predicts that the three PL proteins + histone H1 + histone octamer would protect 350-375bp (four to five complete turns) of DNA. Furthermore, assuming that all the positive charges of the protamine-like proteins

are involved in the neutralization of the DNA charges (as is the case with protamines), three PL-I proteins (from *Spisula*) or three PL-II/PL-IV + nine PL-III proteins (from *Mytilus*) provide approximately 750-800 positive charges that could completely neutralize the 350bp stretch of DNA.

Although in the absence of more experimental information, alternative models such as that depicted in Fig 4B which were early proposed by (Ausió and van Holde 1987) can not be discarded. They, however, do not explain the need for retention of the winged helix motif by PL proteins unless that there is a need for a four-way DNA junction recognition during the binding process (Zlatanova and van Holde 1998). Furthermore a structure such as that shown in Figure 4 is easier to reconcile with the regular chromatin (25-50 nm) fibers observed during spermiogenesis in these organisms (Casas et al. 1993). The proposed structure found in *Mytilus*, while apparently very similar to *Spisula* due to the DNA span covered by the overall PLs, has the potential for considerable flexibility both in the length of the linker DNA and the number of retained nucleosomes required in the mature sperm. In terms of evolution, it seems likely that the independent PL-III gene would evolve to become more arginine-rich and closer to protamines, and eventually obviate the requirement for nucleosomes to be a component of the chromatin structure in the sperm.

Conclusions

The protein and chromatin models we have proposed, in addition to their attractive structural features, support very readily the notion that the PL proteins represent an evolutionary step from a sperm-specific histone H1 to a protamine. This trend appears

to be characterized first by an expansion of the N-terminal tail, manifested by the appearance of well conserved, structurally significant peptide repeats and a significant increase in arginine content. Subsequent genetic events may result in the segregation of the N and/or C terminal tails from the globular winged helix domain of this extended H1, seen first at the post-translational level and later at the genetic level, where we see the emergence of independent protamine-like protein genes under distinct and autonomous control. For events such as these to occur, particularly at the rate that is observed between closely related species and sub-species, there is a requirement for a high rate of recombination at the genomic level. Evidence for abundant recombination events have been observed in the PL genes of *Mytilus* (Ausió, unpublished [see Chapter 4]; (Heath and Hilbish 1998)), winter flounder (Watson and Davies 1999), and *Spisula* (Ausió, unpublished [see Chapter 2]). In fact, the presence of repeated protein motifs may facilitate the elevated level of recombination in these gene regions. It has been suggested that proteins such as H1, which contain characteristic palindromic peptide repeats, may be partially responsible for their own evolution via autologous binding between these peptides and their own coding sequence (Ohno and Becak 1993). As our knowledge and understanding of the structure and evolution of these proteins increases, we will be in a better position to understand the complicated and critically important process of DNA condensation during spermiogenesis.

Lewis and Ausió (2002) Biochem Cell Biol **80**:353-61.

All roads lead to arginine: The squid protamine gene

John D. Lewis, Miriam E. de Jong, Sabira M. Bagha, Alpina Tang and Juan Ausió ¶

Department of Biochemistry and Microbiology, University of Victoria,
Victoria, B.C., V8W 3P6, Canada.

¶ Corresponding author: Department of Biochemistry and Microbiology,
University of Victoria,
P.O. Box 3055, Petch Building Room 220,
Victoria, B.C.
V8W 3P6, Canada

Phone: (250) 721-8863; Fax: (250) 721-8855
email: jausio@uvic.ca

Lewis et al. (2003) J Mol Evol (submitted)

ABSTRACT

The protamine of squid is one of the most arginine-rich protamines (77 % mol/mol). It possesses a leading sequence that is post-translationally removed during spermatogenesis in a manner that is analogous to that observed in some of its vertebrate protamine counterparts. In this paper we describe the gene sequence of the protamine of the squid *Loligo opalescens*. This represents the first complete gene sequence ever reported for an invertebrate protamine. Like those of vertebrate protamines, the messenger RNA is polyadenylated but the gene does not contain an intron. The promoter region contains the major transcriptional regulatory elements (CRE, TATA box and CAP) that are also characteristic of the vertebrate protamine genes. Although the codon usage varies among protamine genes, the squid protamine gene appears to be subject to the same purifying selection that maintains G in the second position of the codons of vertebrate protamine genes. The occurrence of protamines in species from both the deuterostome and protostome branches may thus represent the result of convergent molecular evolution.

Lewis et al. (2003) J Mol Evol (submitted)

INTRODUCTION

In 1961, David Bloch carried out the first biochemical characterization of the sperm nuclear basic proteins (SNBP) of the squid *Loligo opalescens* (Bloch 1962). He used starch gel electrophoresis and amino acid analysis for this characterization. It was shown that these proteins in the testes could be separated into four electrophoretic bands. The two more slowly migrating bands exhibited a very similar mobility to the two bands observed in somatic tissue (gills) from the same organism and had an amino acid composition corresponding to histones. The two bands with higher mobility were found to be arginine-rich. The faster of these two was classified as a protamine, and was shown to contain 70% arginine and a simple amino acid composition consisting only of eight different amino acids. The band preceding it contained 60% arginine by weight and a more complex amino acid composition. In contrast to the testes, the spermatophores consisted only of several fast moving bands, the major one of which had an electrophoretic mobility identical to that of the fastest (protamine) band observed in testes. The other minor bands present in the spermatophore all had higher electrophoretic mobility and most likely corresponded to degradation products of the protamine band. Bloch's study represents the first attempt not only to characterize the squid protamines but also to characterize the SNBP transitions occurring during spermatogenesis in squid.

A more detailed amino acid composition of the squid protamine found in mature spermatozoa of the squid *Loligo pealeii* was provided in 1973 by Subirana and coworkers (Subirana et al. 1973). The protamine was shown to contain 77.5 % (mol/mol) of

arginine, which was the highest of any yet described. Similar to Bloch's observation, only eight other amino acids were detected.

A recent and more detailed study of the SNBP transitions during spermatogenesis in squid (*Ilex argentinus*) (Kadura and Khrapunov 1988) revealed that in the course of testes maturation, histones are gradually replaced by two new proteins of higher electrophoretic mobility, called Ilexine I_1 and Ilexine I_2. The approximate molecular mass of each protein was estimated to be 9 kDa for I_1 and 7 kDa for I_2. In mature sperm only the fastest moving of them (I_2) could be observed. These two bands correspond to the arginine-rich bands that had been earlier described by Bloch. It was shown that I_2 exhibited some microheterogeneity, consisting of at least of two minor components of very close electrophoretic mobility ($I_{2\text{-}1}$ and $I_{2\text{-}2}$). In the course of spermatogenesis, I_1 appeared first, followed by I_2, which was the only protein found in the fully matured spermatozoa.

It was not until 1995 (Wouters Tyrou et al. 1995) that the structural relationship between the electrophoretically slower and the faster moving bands in squid testes was finally established. Similarly to what had been previously observed in cuttlefish (*Sepia officinalis*), another cephalopod (Wouters Tyrou et al. 1991), it was shown that a precursor-product relationship existed between these protein fractions. It was demonstrated that the fastest moving band present in the mature spermatozoa is the result of post-translational processing of the protein corresponding to the slower moving band. This processing involves the proteolytic cleavage of the precursor upon removal of an N-terminal leading peptide.

Protamines are an important SNBP type (Kasinsky 1989; Oliva and Dixon 1991; Ausió 1999) in invertebrates (Lewis et al. 2003b). However, beyond the cDNA sequence of the boll weevil protamine (*Anthonomous grandis*) (Trewitt et al. 1990), no information is available to date about any invertebrate protamine gene.

This paper describes the first characterization of an invertebrate protamine gene. The evolutionary significance of this novel protamine gene sequence is discussed.

MATERIALS AND METHODS

Living Organisms

Specimens of *Loligo opalescens* were a generous gift from Dr. William Gilly and Dr. Zora Lebaric from the John Hopkins Marine Station. The gonads from *L. opalescens* were collected from these organisms and frozen in liquid nitrogen, then transferred to our laboratory on dry ice where they were stored at -80 °C.

Protein Extraction

Chromosomal sperm proteins were extracted and isolated as described (Ausió 1986). Buffers used during the isolation of proteins contained Complete protease inhibitor cocktail tablets (Boehringer). The dried pellets were stored at -80 °C.

Protein Fractionation

Reverse phase HPLC was performed on a 5mm Vydac C18 column (25 x 3 x 0.46 cm) with 0.1% trifluoroacetic acid as eluant with varying acetonitrile gradients (Ausió 1988). 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.

Protein Gel Electrophoresis

Acetic acid (5%)-urea (2.5 M) polyacrylamide gels were prepared as described in (Jutglar et al. 1991).

Protein Sequencing

Protein sequencing was performed on an ABI Model 473 gas-phase protein sequencer at the Protein Microchemistry Center of the University of Victoria, British Columbia as previously described (Jutglar et al. 1991).

DNA Extraction

DNA was extracted from male gonadal (0.1g) tissue according to the protocol described by Sambrook with some modifications (Sambrook et al. 1989). The tissue was weighed, frozen in liquid nitrogen and ground to a powder using a mortar and pestle. The powder was suspended in 50 mL of 10 mM Tris-HCl (pH 8.0), 0.1 M EDTA (pH 8.0), 20 μg/mL pancreatic RNase I (Boehringer Mannheim), 0.5 % SDS and incubated for 1 hour at 37 °C. Proteinase K (Boehringer Mannheim) was then added to a final concentration of 100 μg/mL, followed by incubation for 3 hours at 50 °C with gentle swirling. This solution was cooled to room temperature and extracted twice in phenol:chloroform. A final extraction was performed with chloroform:isoamyl alcohol (49:1). The aqueous phase was then mixed with 0.2 volumes of 10 M ammonium acetate and the DNA was precipitated with 2 volumes of 95 % ethanol. A glass pipette hook was swirled in the solution to collect the high molecular weight precipitated DNA. The DNA was then immersed in 1 mL of 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0) (TE buffer) and tumbled at 4 °C until the DNA was fully resuspended (~48 hours).

RNA Extraction

RNA was extracted from 0.5 g of male gonadal tissue from *L. opalescens* using

TRIzol Reagent (Gibco Technologies), according to their protocol. RNA was analyzed on a 1.2 % formaldehyde/agarose gel as described below. Polyadenylated mRNA was purified from total RNA using Pharmacia Biotech's mRNA Purification Kit.

DNA/RNA Gel Electrophoresis

Horizontal 1.0 % agarose gels for analysis of DNA were prepared in 40 mM Tris-acetic acid (pH 8.0), 1.0 mM EDTA (pH 8.0) (Sambrook et al. 1989). The marker used was λDNA-*Bst*E II (New England Biolabs). Gels were run for 3 hours at 50 volts and visualized in the AlphaImager 2000 Documentation and Analysis System. PCR reactions were also analyzed on vertical 4% polyacrylamide native gels. The running buffer was 1 x E (40 mM Tris-HCl, pH 7.2, 1 mM EDTA, pH 8.0, 20 mM sodium acetate). HHA marker was prepared by digestion of pBR22 plasmid with *Hha*I. Gels were run at 60 volts for 1.5 hours and visualized using the AlphaImager 2000.

Electrophoretic analysis of RNA was performed with horizontal 1.2 % formaldehyde/agarose gels prepared in 1 x MOPS (0.2 M MOPS, 50 mM sodium acetate, and 10 mM EDTA, pH 8) and 1.11 % formaldehyde (Sambrook et al. 1989). Samples were heated at 65 °C for 15 minutes and quenched on ice before adding 2 μL of DEPC-treated RNA loading buffer (50 % glycerol, 1 mM EDTA, pH 8, 0.25 % bromophenol blue, and 0.25 % xylene cyanol FF). The marker used was a 0.24-9.5 kb RNA marker from Gibco Technologies. Gels were run at 40 volts for 3 hours and stained in an ethidium bromide for visualization using the AlphaImager 2000.

cDNA preparation

cDNA was prepared using purified poly A+ mRNA. Each reaction contained 40 units RNAsin (Promega), 10 mM DTT (Gibco Technologies), 1 x First Strand Buffer (Gibco Technologies), 0.2 mM dNTPs (Gibco Technologies), 1.2 μg polyA+ mRNA, 200 units Moloney murine leukemia virus reverse transcriptase (Gibco Technologies), and either 200 pmol of oligo dT primer (Promega), or 200 pmol of random primer (Promega) in a final volume of 40 μL. The reactions were incubated at 37 °C for 60 minutes, 70 °C for 10 minutes, and then cooled on ice and stored at -80 °C.

Degenerate PCR

Degenerate primers (Fig. 2) for PCR were created based on the determined amino acid sequence of the protamine of *L. opalescens*. PCR was performed using the PCRSprint thermal cycler (Interscience) with cDNA synthesized using the three different primer sets (oligo dT, random primers, and a mixture of the two) as template. Single primer controls were also carried out. Each reaction contained 0.75 μL cDNA, 1 x PCR buffer (Gibco Technologies), 1.5 mM $MgCl_2$ (Gibco Technologies), 0.2 mM dNTPs (Gibco Technologies), 3 pmol/μL of each primer (except for control reactions), and 1 unit Taq polymerase (Gibco Technologies) in a final volume of 20 μL. A touchdown profile was used for the amplification, with the annealing temperature decreasing from 65C to 45C over 20 cycles, followed by 10 cycles at 45C.

Hybridization probe preparation

Due to the highly repetitive nature of the 3' end of the 168 bp clone and the

resultant non-specific hybridization properties, it was necessary to create a shorter probe of 108 bp. This was accomplished by PCR amplification of the longer cDNA clone using the primers SQFORND3, 5'- GCTGAGAAGTTAGAATTGATGAAGGGCGG - 3', and MIDPROTREV, 5'- CCTGCGTCGGTATGGAGAA - 3'.

Northern Blot

A horizontal 1.2 % formaldehyde/agarose gel was run as described above with 2 μg of total RNA, 2 μg poly A+ mRNA, a positive control consisting of TOPO2.1 vector containing the 108 bp probe, and the 0.24-9.6 kb RNA marker (Gibco Technologies). The gel was blotted using the VacuGene® XL Vacuum Blotting System (Pharmacia Biotech) following the manufacturer's instructions. The gel was first washed twice in 20 x SSC for 20 minutes. Transfer was done for 1 hour to Zeta-Probe® GT (BioRad). The blot was washed for 5 minutes in 20 x SSC, air-dried for 30 minutes, and vacuum-dried for 30 minutes at 80 °C. The double-stranded 108 bp insert was labelled by nick translation (Sambrook et al. 1989). The insert was radiolabeled with 50 μCi of α^{32}P-dATP (Amersham/Pharmacia Biotech) in the presence of 10 ng/mL DNase I (Boehringer Mannheim), 2.5 units of *E. coli* DNA polymerase I (New England Biolabs), dNTPs-dATP (Gibco Technologies) for 1 hour at 16 °C. The resulting labelled probe was purified from the free label using a microcon® 10 (Amicon) following the manufacturer's protocol. 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), 7 % SDS at 65 °C for 30 minutes. The probe was heat-denatured at 95 °C for 5 minutes and added to the hybridization solution, 5 mL of 0.5 M Na_2HPO_4 (pH 7.2), 7 %

SDS. Hybridization was carried out for 18 hours at 65 °C.

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

Genomic Walking

Genomic walking was performed on *L. opalescens* genomic DNA using adaptors, adaptor primers, and protocols based on (Zhang and Gurr 2000), with a few modifications (Fig. 3). DNA was digested overnight with *Spe*I, *Nhe*I and *Xba*I (New England Biolabs). Adaptors were ligated at 16C for 6 hours, and PCR reactions were carried out using the adaptor-specific PCR primer PP1, and the gene-specific primers SQCD1-F (5'-ATGAAGGGCGGTAGGAGACGAAGACGAAGG-3') and SQCD1-R (5'-TGCGCCTGGAGGTACGCCTCCTCCTCCTCC-3'). A 1/40 dilution was made of the products of the first reaction, and 1µl of this was added to a second PCR reaction using the nested adaptor-specific PCR primer PP2, and the gene-specific primers SQCD2-F (5'-AGGTCTCGTTCTCCATACCGACGCAGGAG-3') and SQCD2-R (5'-CGGTATGGAGAACGAGACCTTCGGCGACG-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 (Zhang and Gurr 2000).

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 manufacturer's instructions, and transformed into TOP10 competent cells (Invitrogen). DNA was sequenced by the dideoxynucleotide method (Sanger et al. 1977) using a Sequenase 2.0 kit (USB Corp).

RESULTS AND DISCUSSION

Developmental SNBP changes during *L. opalescens* spermatogenesis result in the presence of a highly arginine-rich protamine in spermatozoa

Figure 1A shows an electrophoretic analysis of the SNBP transitions occurring during spermatogenesis in *L. opalescens*. The protein extracts were prepared from testes of squid at different stages of sexual maturity and from mature spermatozoa obtained from spermatophores. As can be seen, during the early stages of spermatogenesis the testes contain mainly histones (Fig. 1A, lane 2). As spermatogenesis proceeds, a protamine precursor appears (Fig. 1A, lanes 3,4) that is gradually transformed into the mature

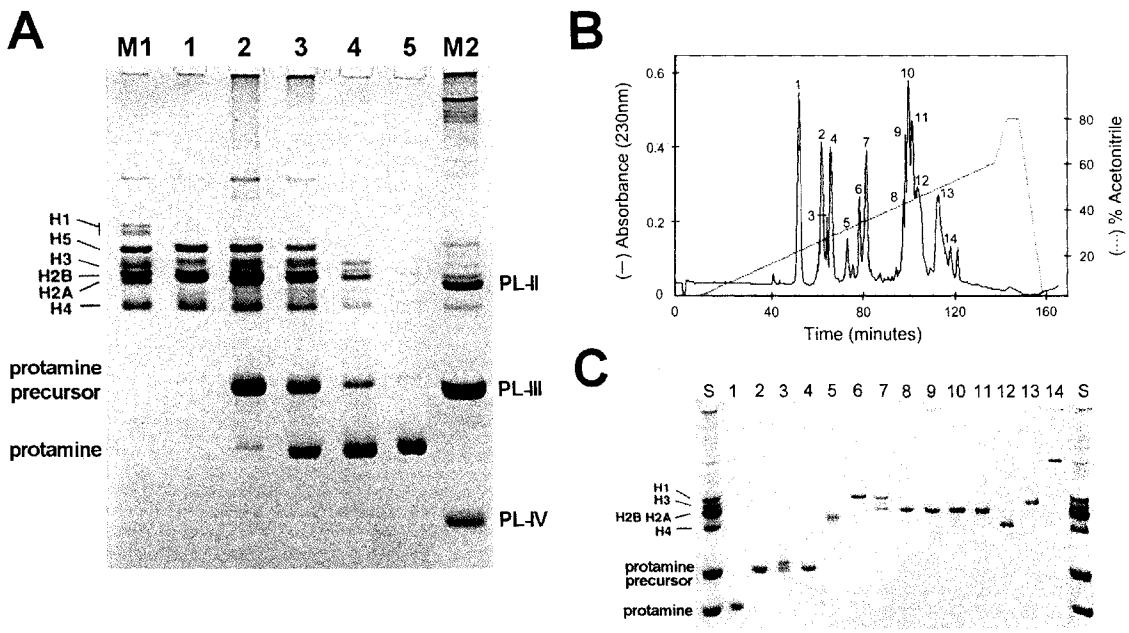


Figure 1. Characterization and fractionation of the squid SNBP. **A.** Acetic acid (5%)-urea (2.5 M) PAGE of SNBP extracted from squid testes at different stages of development. arranged in order of increasing in sexual maturity (lanes 1 to lane 8). Chicken erythrocyte histones (lane M1) and California mussel (*Mytilus californianus*) SNBP (Oliva and Dixon 1990; Carlos et al. 1993b) (lane M2) were used as protein markers. **B.** Reversed-phase HPLC chromatographic profile of SNBP from incompletely mature testis (lane 3 in panel A). Fractionation was on a Vydac C_{18} (25 x 0.46 cm) eluted at 1 mL/min with a 0.1 % TFA-acetonitrile gradient. **C.** Electrophoretic analysis of the peaks from B. The lane numbers correspond to the peaks shown in B.

protamine found in the spermatozoa (Fig. 1A, lane 5).

The occurrence of protamine precursors is not unique to cephalopod protamines (Wouters Tyrou et al. 1991). It has been described in other protostome (Caceres et al. 1999) invertebrates and in deuterostomes (Chiva et al. 1995; Lewis et al. 2003b), including mammalian P2 (Hecht 1989; Lewis et al. 2003b). Although the functional role of this precursor-product relationship is not clear, the sequential processing that occurs during spermatogenesis appears to play an important role in the structural transitions undergone by chromatin (Caceres et al. 1999; Lewis et al. 2003b) that lead to the highly compacted chromatin organization in spermatozoa.

In order to determine the sequence of the protamine precursor protein, an HCl nuclear extract obtained from testes at an intermediate stage of development was subjected to HPLC (Fig. 1B) and the peak corresponding to the protamine precursor (Fig. 1C, lane 2) was sequenced by Edman degradation microsequencing. The extent of purity of the fractionated sample and the amounts recovered were such that 71 amino acids could be sequenced in a single run, despite the high arginine-rich content of the protein (Fig. 2A). With this information sequence at hand it was possible to design several degenerate oligonucleotide primers (Fig. 2A), which then allowed us to obtain a partial cDNA sequence which clearly matched the protamine protein sequence.

The long quest for the squid protamine gene

We used the cDNA fragment of the squid protamine to screen a *L. opalescens* genomic library in search of the entire gene. Although many apparently positive colonies were obtained during this process, none of them contained the gene of interest. It became clear that most of the positive-testing clones contained highly repetitive $(GAA)_n$ DNA sequences which, although similar to the repetitive sequence of the squid cDNA, did not contain any significant open reading frames. We therefore decided to change our strategy and use a PCR-based genomic walking technique (Zhang and Gurr 2000) (see Fig. 3A). This approach resulted in two highly specific DNA fragments (see Fig 3B, lanes 5'b and 3'c') that provided the

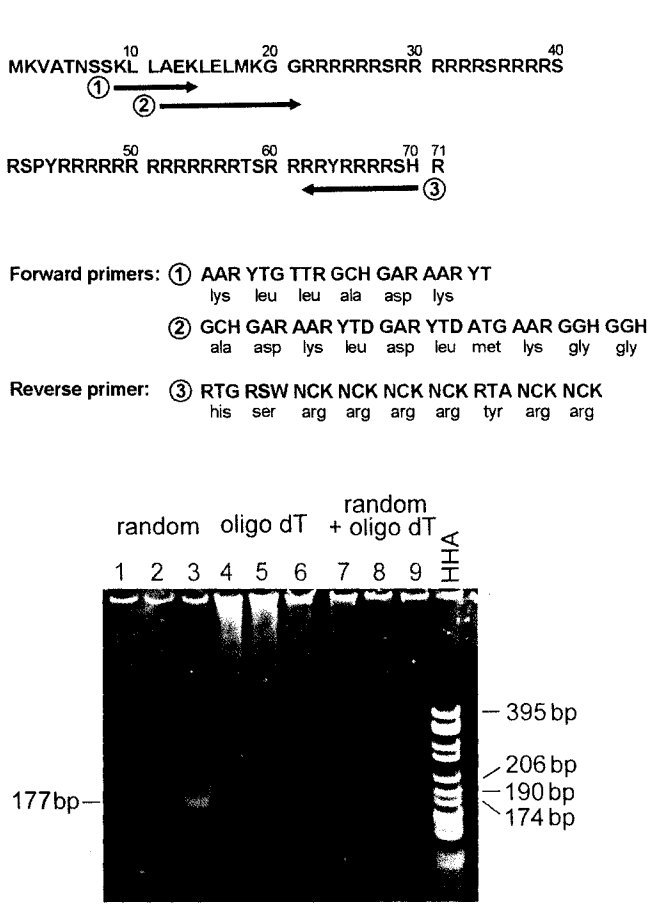
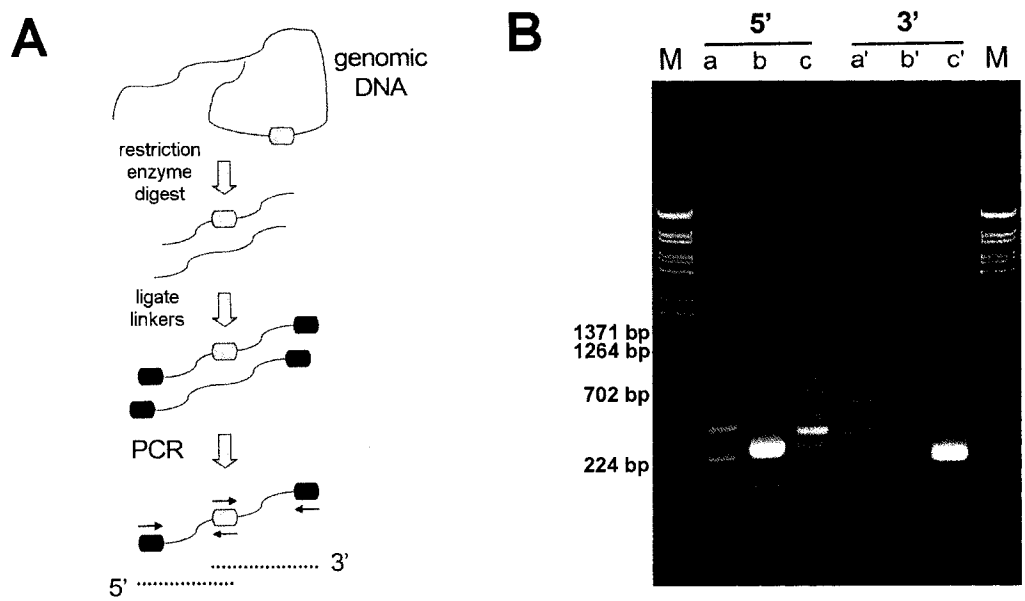


Figure 2. Strategy employed to obtain a squid protamine partial cDNA sequence. **A.** A squid protamine precursor purified as shown in lane 2 (Figures 1B and 1C), was sequenced by Edman degradation. The first 71 N-terminal amino acids shown were sequenced in a single run and this sequence information was used to design the degenerate oligonucleotide primers indicated. **B.** Electrophoretic analysis of PCR products using squid cDNA as template, produced by three different methods of cDNA preparation: random primers, oligo-dT primers and random + oligo dT primers. Lanes 1,4 and 7 are single forward primer controls. Lanes 2,5 and 8 are single reverse primer controls. Lanes 3, 6, and 9 contain reaction products utilizing forward primer #2 and the reverse primer shown in A. Also included in this gel (as a marker) is pBR322 plasmid DNA digested with *Hha* I. The products of the PCR reactions were run on a vertical 4% minislab native PAGE.



C

```
5' - ccagggttcaaatcaacttcattgtgttagatttgtatgtggttaaatgga -231
ggccacgttctatacgttggcaaatgagggatgactagctcgattagtaaaagcatttac -181
tgctcagtcccggctcttagatttgcaccattaacgacgtctgcattatttctccactgc -121
tccgcacaaacttatgacgtatcaattctagagtctaatactgcaatttaaactggccta -61
cttattgcaaaatataatggcagcactccaggtcccctcggcctcccgcaccagcccggc -1

atgaaggtcgctaccaacaggtctaagttgttggctgagaagttagaattgatgaagggc 60
 M  K  V  A  T  N  R  S  K  L  L  A  E  K  L  E  L  M  K  G

ggtaggagacgaagacgaaggtcaaggagacgacgaaggagaagtcgtcgccgaaggtct 120
 G  R  R  R  R  R  R  S  R  R  R  R  R  R  S  R  R  R  R  S

cgttctccataccgacgcaggaggaggaggaggaggaggaggaggcgtacctccaggcgc 180
 R  S  P  Y  R  R  R  R  R  R  R  R  R  R  R  R  T  S  R  R

aggcggtacaggagaagaaggagttactctaggcgaaggtacaggaggagaaggtagcta 240
 R  R  Y  R  R  R  R  S  Y  S  R  R  R  Y  R  R  R  R  *

ggtgttgacttctgcttagacaaaagatgggctctaacgcacaatacaataatagcctta 300
acgactgtatattngtcgatagatatttttaggctacaaactatgtaatttagataattt 360
cctaacgaatttgattgatagtataaataccataacggtccttaagcaaaacaacacgtt 420
tattgacaccaggtacacattcaatc - 3' 446
```

Figure 3. Characterization of the squid protamine gene. **A.** Schematic representation of the major steps involved in genomic walking (Zhang and Gurr 2000). **B.** Agarose (1.5 %) gel electrophoresis of the DNA fragments obtained by genomic walking. The fragments obtained by walking in the 3' and 5' direction are shown. Lanes a, b, and c, correspond to the fragments obtained from PCR using genomic DNA digested with *Spe*I, *Nhe*I and *Xba*I respectively. PCR was performed using the nested set of primers, SQCD2-R (gene-specific) and PP2 (adaptor-specific). Lanes a', b', and c' correspond to the fragments obtained from PCR using genomic DNA digested with *Spe*I, *Nhe*I and *Xba*I respectively. PCR was performed using the nested set of primers, SQCD2-F (gene-specific) and PP2 (adaptor-specific). Lane M is a *Bst* EII digest of λDNA, used as a marker. **C.** Nucleotide sequence of the squid protamine gene (Genbank accession AY269798). The nucleotides of the coding region are shown in bold and the amino acid sequence of the encoded protamine is shown in bold capital letters.

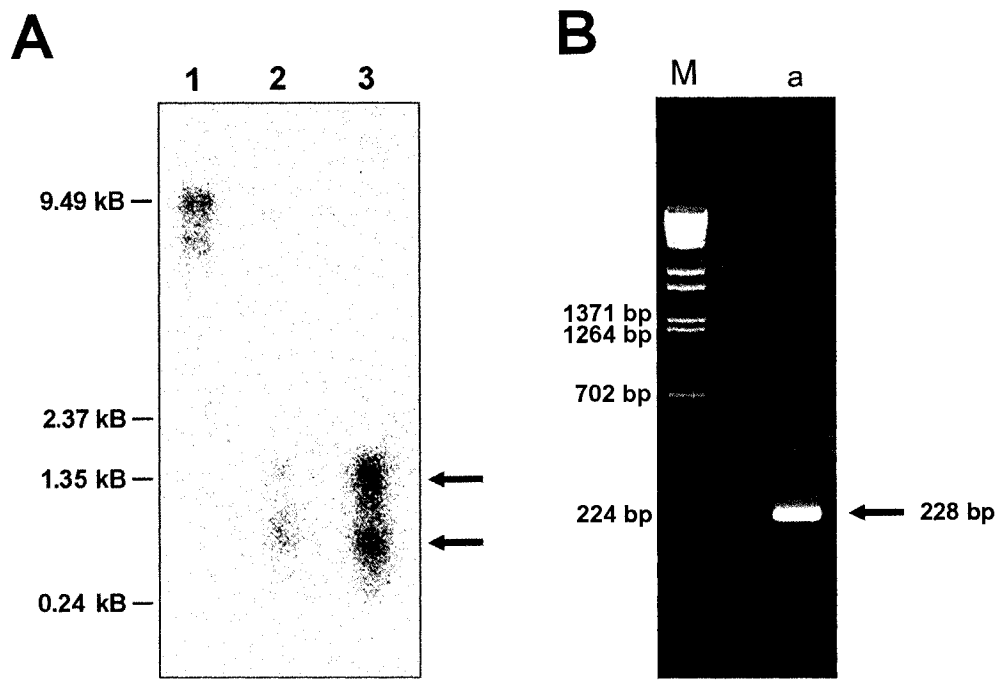


Figure 4. A. Northern blot of an agarose (1.2 %)/formaldehyde gel of total RNA (lane 2) and mRNA (lane 3) purified from immature squid testes. Lane 1 is a control consisting of 50 ng of TOPO2.1 vector (Invitrogen) containing the 108 bp probe. The two arrows point to the two different populations of mRNA present in this squid testes sample. **B.** Agarose (2 %) gel electrophoresis of a DNA fragment obtained from PCR of genomic squid DNA and primers corresponding respectively to the 5' (SQ-GENF, 5'- GTCGCTACCAACAGGTCTAAGTTG -3') and 3' (SQ-GENR, 5'-CCTTCTCCTCCTGTACCTTCGCCTAGAG -3') ends of the coding region of the squid protamine gene (lane a). Lane M contains the same DNA marker as that shown in B.

sequence information of the regions of the gene spanning approximately 300 nucleotides upstream from the initiation codon in the 5' direction and 200 nucleotides downstream from the termination codon in the 3' direction (see Fig. 3C).

Northern analysis carried out on RNA extracted from testes at intermediate stage of maturity using probes derived from the cDNA (see Materials and Methods) showed the presence of two almost equally intense broad distinct electrophoretic bands of 280 and 390 nucleotides in the mRNA fraction (see Fig. 4A). The total RNA fraction consisted mainly of the 280 nucleotide band (see Fig. 4A, lane 2). Treatment of the mRNA fraction with RNase H resulted in the almost complete disappearance of the 390 nucleotide band suggesting that the 390 and 280 nucleotide bands correspond to the poly A^+ and poly A^-

versions respectively. Similar size heterogeneity has been observed in other protamine mRNAs and has been shown to be the result of a progressive shortening of the poly A tail in the transition from round to elongated late spermatids (Hecht et al. 1985; Oliva et al. 1988).

PCR analysis carried out on genomic DNA using gene-specific primers corresponding to the beginning and ending of the coding region produced a single 228 bp fragment, indicating that the gene does not contain any introns (see Fig. 4B, lane a).

The squid protamine gene, a clear case of convergent molecular evolution

The predominantly striking property of the mature squid protamine is its highly arginine-rich composition, which makes it virtually indistinguishable from the vertebrate protamines. In addition, and perhaps even more interesting, is the existence of a precursor protein that possesses a leading sequence that is removed during the course of spermatogenesis (Lewis et al. 2003b) (see Fig. 5A). Thus, the overall similarity between this protostome protamine and deuterostome protamines is remarkable. Moreover, this similarity extends beyond the primary structure of the protein into the gene composition and organization. As shown in Fig. 5B, the proximal regions at the 5' and 3' of the gene appear to contain basic regulatory elements that are very similar to those characteristic of the vertebrate protamine genes (Oliva and Dixon 1990). Therefore, although it is not usually easy to convincingly provide support for a case of molecular evolutionary convergence (Doolittle 1994), we feel that the information available for the squid protamine strongly supports the notion that protostome and deuterostome genes may have evolved through a process of convergent evolution.

Furthermore, it appears that a selection mechanism similar to that responsible for maintaining a high content of arginine in vertebrate protamines (Rooney et al. 2000; Lewis et al. 2003b) has occurred during the evolution of the squid protamine. Vertebrate

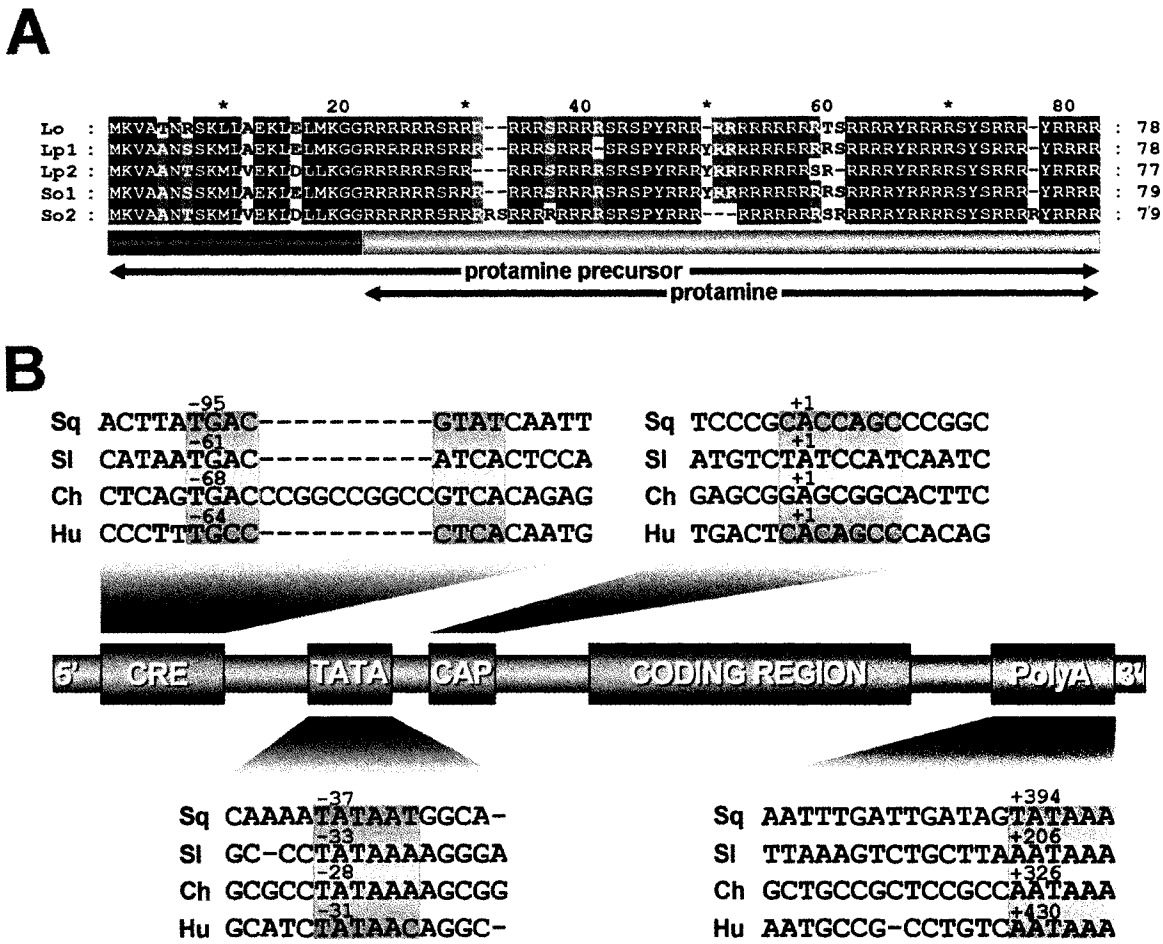


Figure 5. A. Comparative analysis of the Pacific squid (*L. opalescens*) protamine sequence (Lo) in comparison to the amino acid sequences of the protamines of other squid protamines. Lp: Atlantic squid (*Loligo pealeii*) (Wouters Tyrou et al. 1995) S56116; So: cuttlefish (*Sepia officinalis*) (Wouters Tyrou et al. 1991) A40973. Alignment was carried out by CLUSTAL X (1.64b) (Thompson et al. 1997). As indicated, the black cylinder indicates the region corresponding to the leading sequence and the grey cylinder corresponds to the protamine which is found in the mature sperm. **B.** Alignment of the nucleotide sequences corresponding to the constitutively conserved structural features (CRE, TATA box, CAP site, and polyadenylation signal) of the consensus gene organization of vertebrate protamine gene (Oliva and Dixon 1990). The sequence of the protamine gene of squid (*L. opalescens*) (Sq) is shown in comparison to the sequences of the P1 protamine genes of salmon (*Oncorhynchus keta*) (Sl), chicken (*Gallus domesticus*) (Ch) and human (*Homo sapiens*) (Hu). The GenBank sequence accession numbers are: Salmon P1 (X07511); Chicken P1 (M28100); Human P1 (Z46940).

protamine genes exhibit an unusual form of purifying selection that maintains the occurrence of the nucleotide guanine (G) at a high frequency in the second position of any given codon (see Fig. 6A) (Rooney et al. 2000). As shown in this figure, there is evidence that this mechanism has also taken place in the protostome protamines, to a greater extent in the squid as compared to the weevil. This behaviour is observed despite a significant difference in overall codon usage displayed by vertebrate and invertebrate protamine genes (Fig. 6B).

The driving force behind the common selection process in the evolution of protamine genes appears to be the maintenance of high arginine content, which is required for the efficient and tight packaging of DNA in sperm chromatin.

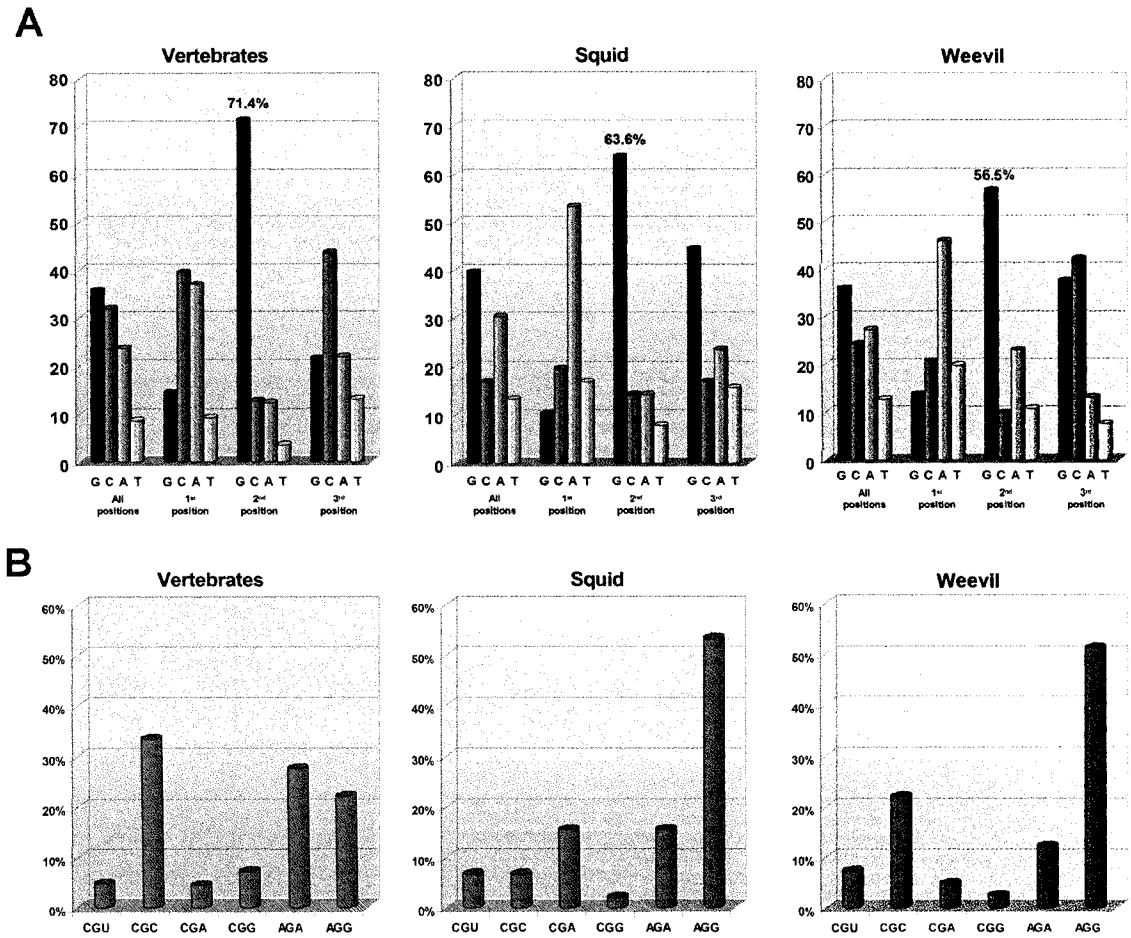


Figure 6. A. Codon nucleotide composition of the 'consensus' vertebrate protamine P1 gene in comparison to the codon nucleotide composition of squid and boll weevil (X52058), the only two invertebrate protamines for which there is information available. The sequences using to generate the vertebrate consensus are; Fish: Salmon P1 from *O. keta* (X07511); Trout P1 from *O. mykiss* (X01599); Dogfish protamine from *S. canicula* (X04517); Amphibians: Newt P1 (D85427) from *C. pyrrhogaster* and Toad P1 (X56529) from *B. japonicus*; Reptiles: *A. mississipiensis* (PI and PII) (Ausió et al. unpublished results); Birds: Chicken P1 (M28100) from *G. domesticus* and Quail P1 *C. coturnix* (M30275); and Mammals: Platypus P1 from *O. anatinus* (Z26849); Opossum P1 from *D. marsupialis* (X74044); Human P1 from *H. sapiens* (Z46940); Killer whale P1 from *O. orca* (Z11496).. The nucleotide compositions were determined as described in (Rooney et al. 2000) using the computer program MEGA, version 2.1 (Kumar et al. 2001). **B.** Comparison of arginine codon usage between vertebrates (determined from the same organisms as A) and the two invertebrates, squid and boll weevil.

Acknowledgments

We are very indebted to Dr. William F. Gilly from the Department of Biological Sciences at Stanford University, Hopkins Marine Station, Pacific Grove, CA, for collecting the squid tissue used in this research. This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant OGP 0046399.

Histone H1 and the origin of protamines

John D. Lewis†*, Núria Saperas§*, Yue Song†, Mª Jose Zamora¶, Manel Chiva¶ and Juan Ausió†‡

† Department of Biochemistry and Microbiology, University of Victoria, P.O. Box 3055, Petch Building, Victoria, B.C., Canada, V8W 3P6

§ Departament d'Enginyeria Química, ETSEIB, Universitat Politècnica de Catalunya, Diagonal 647, E-08028, Barcelona, Spain

¶ Departament de Ciències Fisiològiques II, Facultat de Medicina, Campus de Bellvitge, Universitat de Barcelona, L'Hospitalet de Llobregat, Barcelona E-08907, Spain

* These two authors have contributed equally to this work

‡Corresponding author: Department of Biochemistry and Microbiology,
University of Victoria,
P.O. Box 3055, Petch Building Room 220,
Victoria, B.C.
V8W 3P6, Canada

Phone: (250) 721-8863; Fax: (250) 721-8855
email: jausio@uvic.ca

ABSTRACT

During the final stages of spermatogenesis, the compaction of DNA in many organisms is accomplished by the replacement of histones with a class of arginine-rich proteins called protamines. In other organisms, however, condensation of sperm DNA can occur with comparable efficiency in the presence of somatic-type histones or, alternatively, an intermediate class of proteins called protamine-like (PL) proteins. The idea that the highly specialized sperm chromosomal proteins (protamines) and somatic chromosomal proteins (histones) could be related dates back almost to the discovery of these proteins. While this notion has frequently been revisited since that time, there has been a complete lack of supporting experimental evidence. Here we show that the emergence of protamines in chordates occurred very quickly, as a result of the conversion of a lysine-rich histone H1 to an arginine-rich protamine. We have characterized the sperm nuclear proteins of the tunicate *Styela montereyensis*, which we show to consist of both a protamine and a novel sperm-specific histone H1 with a protamine tail. Comparison of the genes encoding these proteins to that of a sister protochordate, *Ciona intestinalis*, has indicated that this rapid and dramatic change is the result of a frameshift mutation in the tail of the sperm-specific histone H1. By establishing an evolutionary link between the chromatin-condensing histone H1s of somatic tissues and the chromatin-condensing proteins of the sperm, these results provide unequivocal support to the notion that vertebrate protamines evolved from histones.

Lewis et al. (2003) PNAS (in press)

INTRODUCTION

The canonical structure of histone H1 consists of a tripartite organization, with a globular core containing a conserved winged helix motif, flanked by less well conserved lysine-rich amino- and carboxy-terminal tails. Somatic-type histone H1s typically contain little or no arginine. Histone H1-like sperm nuclear proteins have been identified in a diverse range of organisms, including marine invertebrates (Ausió 1992), amphibians (Kasinsky et al. 1985; Itoh et al. 1997), and fish (Saperas et al. 1994; Watson and Davies 1998). They are generally characterized by both an elevated arginine and lysine content (Lys + Arg > 35%), and the presence of the conserved winged-helix motif that is a defining characteristic of histone H1s (Kasinsky et al. 2001). Beyond these simple constraints, however, these proteins show significant compositional heterogeneity and can range in size from 6500 Da up to 200000 Da. In contrast, protamines are relatively small (4000 Da to 12000 Da), are composed of over 50% arginine, and contain little or no lysine. From the time that these nuclear proteins were first characterized, it was suggested that histones of somatic cells and protamines of germ cells were evolutionarily related (Stedman and Stedman 1947; Felix 1960). It was hypothesized in 1973 (Subirana et al. 1973), based on compositional amino acid analysis, that protamines had evolved from a primitive somatic-like histone precursor via a protamine-like (PL) intermediate through a mechanism of vertical evolution. This theory contrasted with the hypothesis of the horizontal evolution of protamines, according to which protamines have a retroviral origin (Jankowski et al. 1986). The hypothesis of retroviral horizontal transmission was proposed to account for the apparently random distribution of protamines in fish and was based on the observation that the flanking regions of the protamine genes from rainbow

trout exhibit a large degree of similarity to the long terminal repeats of avian retroviruses (Jankowski et al. 1986). However, a detailed systematic analysis of the distribution of SNBPs in fish provided additional support for the vertical evolution hypothesis by revealing that the sporadic distribution of protamines was not random and could be traced phylogenetically (Saperas et al. 1997). The principal difficulty with the theory of vertical evolution of protamines, however, was the absence of a mechanism by which the lysine-rich histones could be converted to the highly arginine-rich protamines.

To address this difficulty, we have studied the sperm nuclear basic proteins of the primitive chordates *Styela montereyensis* and *Styela plicata*, revealing that each possesses both an arginine-rich protamine and a novel histone H1 with an extremely arginine-rich protamine tail. A closely related tunicate, *Ciona intestinalis*, was found to possess a single major SNBP, a lysine-rich sperm-specific histone H1. The surprising result of a genetic comparison of the genes encoding these proteins has indicated that this wholesale compositional change is not the result of a gene fusion event (retroviral or otherwise), but rather a frameshift mutation in the tail of the sperm-specific histone H1. This observation provides direct evidence for an evolutionary mechanism linking the histones of somatic tissues with the sperm nuclear basic proteins.

MATERIALS AND METHODS

Protein extraction and sequencing. Chromosomal sperm proteins were extracted and isolated as described (Jutglar et al. 1991). Buffers used during the isolation of proteins contained Complete protease inhibitor cocktail tablets (Boehringer). The dried pellets were stored at -80 °C. Protein sequencing was performed on an ABI Model 473 gas-phase protein sequenator at the Protein Microchemistry Center of the University of Victoria, British Columbia as previously described (Jutglar et al. 1991).

Gel Electrophoresis. Acetic acid (5%)-urea (2.5 M) polyacrylamide gels were prepared as described in (Jutglar et al. 1991).

Degenerate PCR and RACE. Degenerate primers for PCR were created based on the amino acid sequence of the P1 protein from *Styela plicata*, with the sequences: 5' primer (TAYAAYGTHATGGTHAARMG), 3' primer (TTRTTYTTRTADATRAANCCNCC). PCR was performed on genomic DNA from *Styela montereyensis* using the PCRSprint thermal cycler (Interscience). A touchdown profile was used for the amplification, with an annealing temperature range of 65°C to 45°C over 20 cycles, followed by 10 cycles at 45°C. RACE was performed using the Marathon cDNA amplification kit (Clontech), with primers based on the genomic fragment obtained from degenerate PCR. The primers were: 5' primer (CCTTCGAATACGACCTACTATAGGGCG), 3' primer (GCGGCCTCTTCGCTATTACGCCAGC).

Codon nucleotide analysis. Nucleotide composition was determined as described in (Rooney et al. 2000) using the computer program MEGA, version 2.1 (Kumar et al. 2001). The mammalian protamine codon nucleotide frequency was generated as an

average of protamine P1 sequences from human (Z46940), opossum (X74044), platypus (Z26849), and killer whale (Z11496).

RESULTS

The sperm of the ascidians, *Styela montereyensis* and *Styela plicata*, contains two major nuclear basic proteins (Fig. 1) (Saperas et al. 1992). These proteins coexist in the mature sperm with 20-25% of a full somatic-type histone complement. Amino acid microsequencing of the smaller of the two proteins established its identity as a protamine of 93 amino acids (*Styela* P2) (Fig. 2d). Containing 51.6% arginine residues arranged in tracts of 4 to 8 residues each, the *Styela* protamine also displays an unusually high lysine content (20.4%) in comparison with other protamines. Unlike the human protamines P1 and P2, it does not contain cysteine or histidine respectively (Lewis et al. 2003b). Amino acid microsequencing data was then obtained for the second major nuclear basic protein (*Styela* P1), a PL protein which is a much larger 170 amino acids and consists of two distinct domains (Fig. 2b). The leading 78 residues of this protein show a remarkable similarity to the amino terminal tail and globular region of histone H1 (Fig. 2c), and to the sperm-specific H1s and protamine-like H1 proteins of other invertebrates (Zhang et al. 1999; Kasinsky et al.

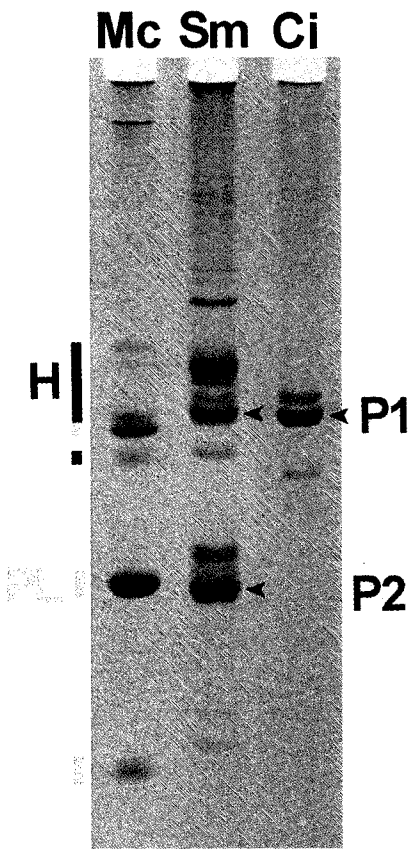


Figure 1. Acetic acid (5%)-urea (2.5 M) polyacrylamide gel electrophoresis of sperm nuclear proteins extracted from testes from: **Mc**. California mussel (*Mytilus californianus*), used as a protein marker. *Mytilus* protamine-like (PL) proteins are indicated by the orange bars. **Sm**. *Styela montereyensis*. **Ci**. *Ciona intestinalis*. Somatic-type histones are indicated by the blue bars. Sperm nuclear proteins referred to in the text are indicated by red arrows.

2001). The carboxy-terminal tail, surprisingly, is comprised of a 93 amino acid sequence (amino acids 78-170) identical to that of the protamine (P2) (Figs. 2b and 2d). This protein represents the first direct tangible link between histone H1 and the evolution of protamines.

Using degenerate primers based on the *Styela plicata* amino acid sequence, a genomic fragment of the larger sperm nuclear protein (P1) was obtained from *Styela montereyensis*. This partial sequence was then used as a template to design PCR primers for RACE in order to obtain the full length cDNA of the P1 from *Styela montereyensis*. The sequence thus obtained codes additionally for a 16 amino acid leading peptide that is not present in the mature PL protein (Fig. 2a). Many protamines are processed post-translationally, including the human P1 and squid P1 (Lewis et al. 2003b).

The sperm nuclear basic proteins of the closely related tunicate *Ciona intestinalis* were extracted, revealing the existence of a single species that was similar in mobility to the larger of the *Styela* SNBPs (Fig. 1). The draft genomic sequence of *Ciona* has recently been completed so we mined the database to see if a similar sperm nuclear basic protein was present in that organism. In addition to two hits that represent the somatic histone H1s in *Ciona*, an H1-like protein with significant similarity to that of the *Styela* P1 was identified. Like the *Styela* P1, it possesses a leading sequence that is presumably cleaved post-translationally.

At the primary structural level, the putative *Ciona* P1 protein exhibits a 75% conservation and 53% identity to the P1 from *Styela* over the entire sequence, but it is clear that the similarity is not uniform (Fig. 2b). While the amino-terminal region of 91 amino acids exhibits 64% identity, the carboxy-terminal tail is only 41% identical, and

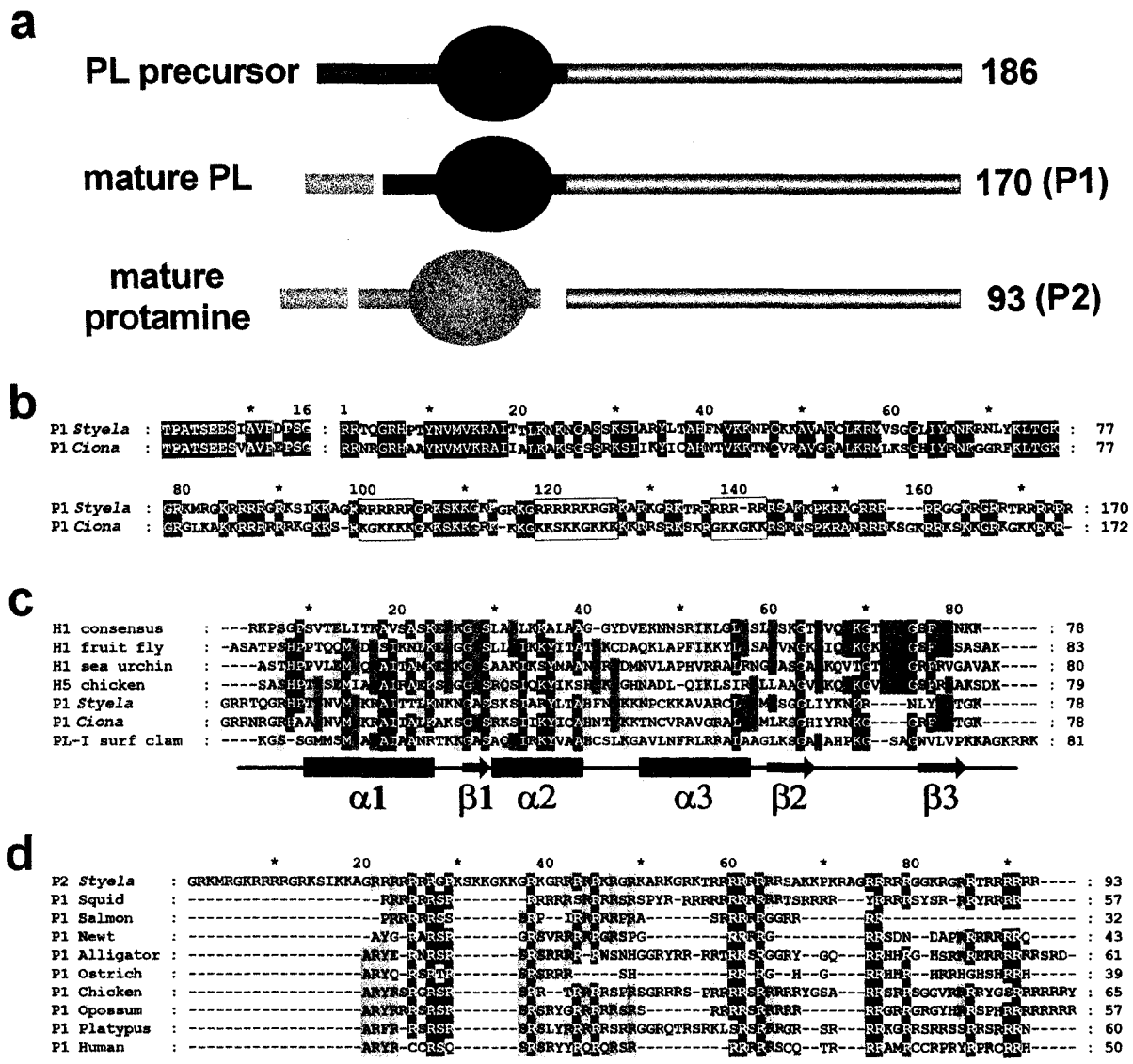


Figure 2. a. Diagram of sperm protein structure in tunicates. Leading sequence is red, winged helix region is blue, carboxy terminal tail is orange. Numbers indicate protein length in amino acids. Mature protein designations are in brackets. **b**. Alignment of *Styela* P1 with *Ciona* P1. Red open boxes indicate areas containing poly-arginine and poly-lysine tracts respectively. Amino acids in red, blue and orange are as in Figure 2a. **c**. Multiple alignment of *Styela* P1 amino-terminal winged helix region with other winged helix-containing proteins, with secondary structure highlighted below. The sequence accession numbers, if available, are: H1 consensus ((Wells and Brown 1991)); H1 Fruit Fly (P02255); H1 Urchin (P15869); H5 Chicken (P02259); P1 *Styela* (AY332242); P1 *Ciona* (BP019154); PL-I Surf Clam (Ausió et al. unpublished). **d**. Multiple alignment of *Styela* P2 with protamines. The sequence accession numbers, if available, are: Squid P1 (AY269798); Salmon P1 (X07511); Newt P1 (D85426); Alligator P1 (Ausió et al. unpublished); Ostrich P1 ((Ausió et al. 1999)); Chicken P1 (M28100); Opossum P1 (X74044); Platypus P1 (Z26849); Human P1/P2 (Z46940).

shows a striking but simple sequence divergence from the *Styela* P1. Where the *Styela* P1

tail is composed of greater than 50% arginine and 20% lysine, the *Ciona* P1 tail contains

only 25.3% arginine and almost 50% lysine. Side-by-side analysis of the two proteins shows that in regions where *Styela* possesses poly-arginine tracts, the *Ciona* P1 has poly-lysine tracts (Fig. 2b). As mentioned previously, *Ciona intestinalis* expresses only a single sperm nuclear basic protein, with a size corresponding with that predicted by the putative P1 obtained from the sequence database (Fig. 1). With the exception of the genera *Styela*, all of the other members of the class Ascidiacea studied to date, like *C. intestinalis*, express only P1 (and no P2) in their sperm nuclei (Chiva et al. 1992).

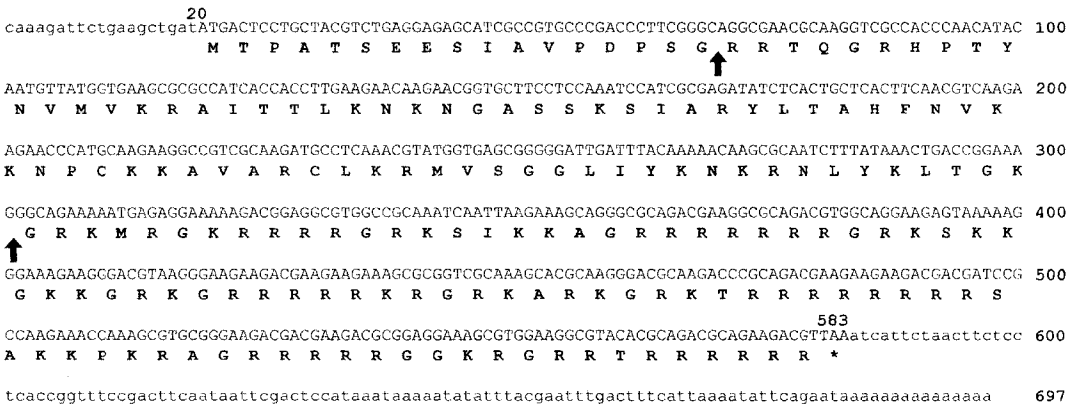


Figure 3. Full-length nucleotide sequence of the *Styela montereyensis* P1 cDNA, Genbank accession AY332242. Sites of likely post-translational cleavage are indicated by the arrows.

DISCUSSION

How is it possible that two homologous proteins from very closely related organisms could manifest such a wholesale switch in amino acid composition? Analysis of the codon usage in the carboxy-terminal tails of the *Ciona* and *Styela* PLs (Fig. 4a) shows that there is significant codon bias in each. Comparison of the lysine codon usage reveals that although each utilizes a nearly equivalent number of AAA codons to encode lysine, 26 of the 27 lysines (96%) lost by the *Ciona* P1 (as compared with *Styela* P1) are AAG codons. Of the six possible codons that encode arginine, 15 of the 24 arginines (63%) gained by the *Styela* P1 are AGA codons. Save for the gain of seven CGC codons, the remaining arginine codons remain essentially constant.

The conversion of lysine AAG codons to arginine AGA codons requires two point mutations. If evolution between these two proteins was to have occurred by point mutation alone, a minimum of 90 independent nucleotide substitutions would be required to achieve such a radical shift in lysine and arginine content. Examination of the nucleotide sequences of both the *Styela* and *Ciona* P1 gene indicate that it is much more likely that a frameshift mutation occurred in the carboxy-terminal tail of the *Styela* P1. Deletion of a single nucleotide at position 342 and two nucleotides at position 437 of the coding region of the *Ciona intestinalis* P1 creates a frameshift mutation with surprising consequences (Fig. 4b). The arginine content of the carboxy-terminal tail effectively increases from 25.3% to 42.6%, with the net conversion of 15 lysine residues to arginine residues. In turn, this mutation decreases the lysine composition from 46.3% to 25.5%, reflecting the observed compositional differences between the *Styela* and *Ciona* P1

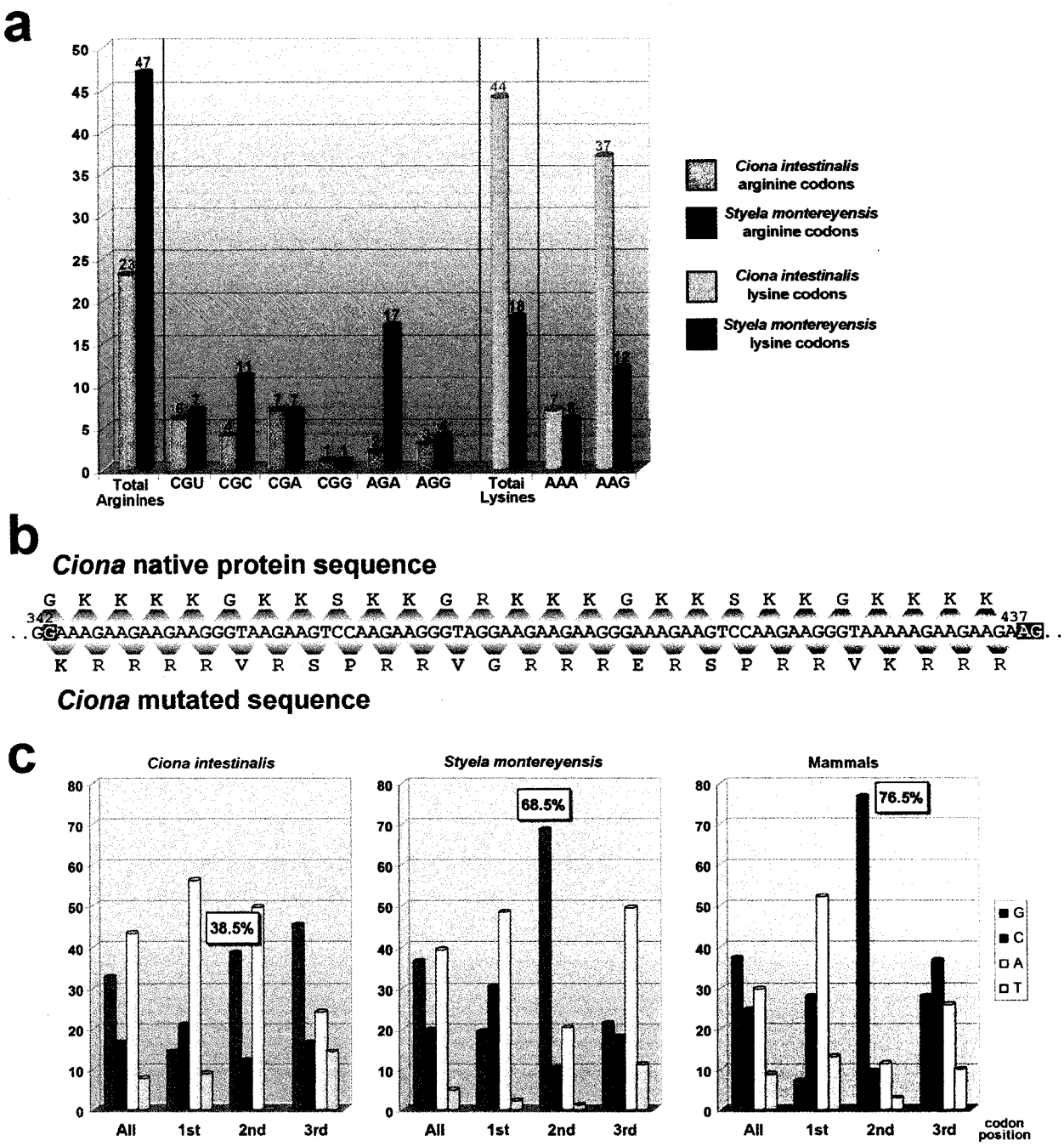


Figure 4. a. Codon frequencies of arginine and lysine codons contained within the carboxy terminal 93 amino acids of *Styela* P2 and *Ciona* P1. **b**. Deletion of nucleotide 342 results in a frameshift mutation that converts 15 lysine codons to arginine codons. **c**. Codon nucleotide frequencies generated for the carboxy terminal 93 amino acids of *Styela* P2 and *Ciona* P1. Frequencies shown for each position, and an average for all positions. Percentages highlighted for nucleotide G occurring in the second position.

proteins. This phenomenon represents a novel type of rapid protein evolution, particularly in its implications. While it has been suggested that there may be frameshift evolutionary

relationships between protein sequence families (Pellegrini and Yeates 1999), this is the first example of such a significant and functional conversion due to frameshift.

Another important shift that this mutation would initiate is a marked change in codon nucleotide composition bias (Fig. 4c). Protamines, like many of the reproductive proteins, are amongst the most rapidly evolving proteins in the animal kingdom (Wyckoff et al. 2000; Swanson and Vacquier 2002). The evolution of protamines is driven by a unique form of purifying selection that permits a relatively high rate of nonsynonymous substitution as long as the proportion of arginine residues is conserved (Rooney et al. 2000). This form of selection, as a direct result of a preference for arginine, favours the inclusion of codons with the nucleotide G in the second position. The *Styela* protamine displays a nucleotide codon bias that is characteristic of vertebrate protamines, with a 68.5% incidence of the nucleotide G in the second codon position (significantly higher than the calculated 18% average for all tunicate proteins). The carboxy-terminal tail of the *Ciona* P1 protein, however, shows a much lower bias with a frequency of only 38.5%. When a frameshift mutation occurs such as that depicted in Fig. 4b, the nucleotide bias in the second codon position is elevated to 54.4%. We submit that this form of purifying selection becomes relevant as a direct result of the frameshift mutation event described above.

Despite the sporadic occurrence of protamines in the animal kingdom, it is clear that selection for arginine in these proteins is at least in part due to its DNA-binding function. While lysine could also conceivably also perform this function, it has been reported that an increase in arginine content at the expense of lysine residues increases the affinity of a protein for DNA (Puigdomenech et al. 1976; Ausió et al. 1984). A

"commitment" to an arginine-rich protamine could also be the result of its involvement at the time of sperm-egg fertilization. It has been shown that proteins containing poly-arginine tracts have the ability to activate casein kinase II (Ohtsuki et al. 1996), an important regulator of cellular metabolism in the developing egg.

In summary, the results presented in this paper provide unequivocal support to the notion that at least in chordates, the line that gave rise to the vertebrate protamines has in fact evolved from histones (Subirana et al. 1973). Most importantly, they establish a hitherto elusive evolutionary link between the chromatin-condensing histone H1s of somatic tissues and the chromatin-condensing proteins of the sperm.

ACKNOWLEDGEMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant OGP 0046399 to JA and grant BMC2002-04081-C02-02 (Ministerio Ciencia y Tecnologia/FEDER) to MC and NS.

CONCLUSIONS

My research has concentrated mainly on the investigation of the proposal that protamines have evolved vertically from an ancestral histone H1 by characterizing the SNBPs and their genes from a diverse range of organisms that employ histones, protamines, or protamine-like proteins to achieve sperm chromatin compaction. The results obtained therein have provided a great deal of insight into the answers to the questions posed in the introduction of this thesis regarding the plausibility of the theory of vertical evolution of SNBPs.

Many organisms are closely related phylogenetically yet contain very different SNBP numbers and sizes. By what mechanism does this rapid change occur?

The characterizations of the *Spisula* and *Mytilus* sperm nuclear basic proteins have provided valuable insight into probable mechanisms of evolution between the various SNBP types. Evidence for tandem genetic duplication in *Spisula* PL-I that is presented in Chapter 4 provides a mechanism for the general increase in protein length that is seen between somatic-type histone H1s and the larger PL proteins. The evidence presented in *Mytilus* that post-translational cleavage is occurring in a PL precursor to produce multiple mature PL proteins, particularly when compared to the *Spisula* PL protein, indicates that post-translational cleavage is a mechanism that results in multiple proteins that originate from equivalent genetic structures. In addition, the data presented in Chapter 5 indicates that while the *Mytilus* PL-II and PL-III proteins both display similarity with portions of the PL-I protein of *Spisula*, they are expressed from independent genes. This gene segregation represents a critical step in the evolution of histone H1-related proteins to

protamines.

What are the differences at the genetic level of the different types of sperm nuclear basic proteins? Extensive characterization of the vertebrate protamine genes has provided good insight into the regulation of these proteins, but with limited scope.

The research presented in this thesis represent the first genes characterized for PL proteins (*Spisula* PL-I, *Mytilus* PL-II, PL-III). Comparison of these genetic sequences has revealed a number of putative conserved regions that are unique to invertebrate PL proteins. Most significant is the identification of a conserved region upstream from the polyadenylation signal that may be a target of factors involved in the translational repression of PL mRNA during spermiogenesis. Furthermore, the characterization of the squid protamine gene is the first protamine gene described in invertebrates. This gene possesses elements that are also present in the vertebrate protamine genes, and therefore seems to have arisen as a result of convergent evolution.

All of these SNBPs fulfill the common function of sperm chromatin condensation. How can PL proteins of such variability in size and number adequately perform this function in such a structurally indistinguishable way?

In Chapter 6, a novel model of sperm chromatin structure is proposed to account for both SNBP variability and the observation that these proteins condense sperm DNA with comparable efficiency. The important observation is made that regardless of the number and size of PL proteins present in the sperm of each organism examined, the total length of protein is essentially equivalent in each case. The model presented also accommodates

very well the presence of the winged helix motif in the H1-related PL proteins, and allows for considerable flexibility both in the length of the linker DNA and the number of retained nucleosomes required in the mature sperm. This proposed novel chromatin structure supports very readily the notion that the PL proteins represent an evolutionary step from a sperm-specific histone H1 to a protamine.

How have somatic linker histones that are by definition lysine-rich, evolved so rapidly into the highly arginine-rich protamines of the sperm?

This has historically been the most puzzling question concerning the evolution of protamines. Histone H1s typically have a high lysine content, with little or no arginine. While the PL proteins are both arginine and lysine-rich, protamines possess an extremely high arginine content with little or no lysine. The characterizations of PL proteins in Chapters 3,4,5 and 7 of this thesis have revealed a number of mechanisms for the evolution of histone H1 to PL proteins and the evolution within the heterogeneous group of PL proteins. It is the research presented in Chapter 8 concerning the SNBPs of tunicates, however, which has uncovered a compelling mechanism for the wholesale conversion of lysine to arginine through frameshift mutation. This mechanism, in fact, may facilitate the evolution of an arginine-rich protein from lysine-rich members of the H1, PL-I, or PL groups of SNBPs outlined in the introduction.

Taken as a whole, the genetic data presented in this thesis, beyond providing valuable information concerning the regulation and organization of the heterogeneous family of sperm nuclear basic proteins, has provided unequivocal support to the hypothesis that the

chromatin-condensing protamines of the sperm have evolved vertically from the chromatin-condensing histones of somatic cells.

REFERENCES

Allan, J., P.G. Hartman, C. Crane Robinson and F.X. Aviles (1980). The structure of histone H1 and its location in chromatin. *Nature* **288**: 675-9.

Allan, J., T. Mitchell, N. Harborne, L. Bohm and C. Crane Robinson (1986). Roles of H1 domains in determining higher order chromatin structure and H1 location. *J Mol Biol* **187**: 591-601.

Ando, T., M. Yamasaki and K. Suzuki (1973). Protamines. Isolation, characterization, structure and function. *Mol Biol Biochem Biophys* **12**: 1-114.

Arents, G., R.W. Burlingame, B.C. Wang, W.E. Love and E.N. Moudrianakis (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc Natl Acad Sci U S A* **88**: 10148-52.

Arents, G. and E.N. Moudrianakis (1995). The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc Natl Acad Sci U S A* **92**: 11170-4.

Ausió, J. (1980). Caracterizacion de las protaminas de los moluscos bivalvos. *Mytilus edulis* y *Spisula solidissima* y estudio de sus interacciones con el ADN. Faculty of Sciences. Barcelona, Spain, University of Barcelona.

Ausió, J. (1986). Structural variability and compositional homology of the protamine-like components of the sperm from the bivalve mollusks. *Comp. Biochem. Physiol.* **85B**: 439-449.

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**: 10141-50.

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**: 163-72.

Ausió, J. (1995). Histone H1 and the evolution of the nuclear sperm-specific proteins. *In* Advances in Spermatozoal Phylogeny and Taxonomy. B. G. M. Jamieson, J. Ausió and J. L. Justine. Paris, Memoires du Museum National d'Histoire Naturelle. **166:** 447-462.

Ausió, J. (1999). Histone H1 and evolution of sperm nuclear basic proteins. *J Biol Chem* **274**: 31115-8.

Ausió, J., K.O. Greulich, E. Haas and E. Wachtel (1984). Characterization of the fluorescence of the protamine thynnine and studies of binding to double-stranded DNA. *Biopolymers* **23**: 2559-71.

Ausió, J. and R. McParland (1989). Sequence and characterization of the sperm-specific protein phi 3 from *Mytilus californianus*. *Eur J Biochem* **182**: 569-76.

Ausió, J., J.T. Soley, W. Burger, J.D. Lewis, D. Barreda and K.M. Cheng (1999). The histidine-rich protamine from ostrich and tinamou sperm. A link between reptile and bird protamines. *Biochemistry* **38**: 180-4.

Ausió, J. and J.A. Subirana (1982a). Conformational study and determination of the molecular weight of highly charged basic proteins by sedimentation equilibrium and gel electrophoresis. *Biochemistry* **21**: 5910-8.

Ausió, J. and J.A. Subirana (1982b). A high molecular weight nuclear basic protein from the bivalve mollusc *Spisula solidissima*. *J Biol Chem* **257**: 2802-5.

Ausió, J. and J.A. Subirana (1982c). Nuclear proteins and the organization of chromatin in spermatozoa of *Mytilus edulis*. *Exp Cell Res* **141**: 39-45.

Ausió, J., A. Toumadje, R. McParland, R.R. Becker, W.C. Johnson and K.E. van Holde (1987). Structural characterization of the trypsin-resistant core in the nuclear sperm-specific protein from *Spisula solidissima*. *Biochemistry* **26**: 975-82.

Ausió, J. and K.E. van Holde (1987). A dual chromatin organization in the sperm of the bivalve mollusc *Spisula solidissima*. *Eur J Biochem* **165**: 363-71.

Ausió, J., M.L. Van Veghel, R. Gomez and D. Barreda (1997). The sperm nuclear basic proteins (SNBPs) of the sponge *Neofibularia nolitangere*: implications for the molecular evolution of SNBPs. *J Mol Evol* **45**: 91-6.

Azorin, F., C. Olivares, A. Jordan, L. Perez Grau, L. Cornudella and J.A. Subirana (1983). Heterogeneity of the histone-containing chromatin in sea cucumber spermatozoa. Distribution of the basic protein phi 0 and absence of non-histone proteins. *Exp Cell Res* **148**: 331-44.

Balhorn, R., M. Corzett and J.A. Mazrimas (1992). Formation of intraprotamine disulfides in vitro. *Arch Biochem Biophys* **296**: 384-93.

Balhorn, R., S. Reed and N. Tanphaichitr (1988). Aberrant protamine 1/protamine 2 ratios in sperm of infertile human males. *Experientia* **44**: 52-5.

Bandiera, A., U.A. Patel, G. Manfioletti, A. Rustighi, V. Giancotti and C. Crane-Robinson (1995). A precursor-product relationship in molluscan sperm proteins from *Ensis minor*. *Eur J Biochem* **233**: 744-9.

Barra, J.L., L. Rhounim, J.L. Rossignol and G. Faugeron (2000). Histone H1 is dispensable for methylation-associated gene silencing in *Ascobolus immersus* and essential for long life span. *Mol Cell Biol* **20**: 61-9.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths_Jones, K.L. Howe, M. Marshall and E.L. Sonnhammer (2002). The Pfam protein families database. *Nucleic Acids Res* **30**: 276-80.

Bates, D.L. and J.O. Thomas (1981). Histones H1 and H5: one or two molecules per nucleosome? *Nucleic Acids Res* **9**: 5883-94.

Baxevanis, A.D., G. Arents, E.N. Moudrianakis and D. Landsman (1995). A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res* **23**: 2685-91.

Beaudoing, E., S. Freier, J.R. Wyatt, J.M. Claverie and D. Gautheret (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**: 1001-10.

Benkel, B.F. and Y. Fong (1996). Long range-inverse PCR (LR-IPCR): extending the useful range of inverse PCR. *Genet Anal* **13**: 123-7.

Black, J.A. and G.H. Dixon (1967). Evolution of protamine: a further example of partial gene duplication. *Nature* **216**: 152-4.

Bloch, D.P. (1962). Symposium: Synthetic processes in the cell nucleus: Histone synthesis in non-replicating chromosomes. *J. Hist. Cyt.* **7**: 137-144.

Bloch, D.P. (1966). Histone differentiation and nuclear activity. *Chromosoma* **19**: 317-339.

Bloch, D.P. (1969). A catalog of sperm histones. *Genetics* **61**: Suppl:93-111.

Blom, N., S. Gammeltoft and S. Brunak (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* **294**: 1351-62.

Brewer, L.R., M. Corzett and R. Balhorn (1999). Protamine-induced condensation and decondensation of the same DNA molecule. *Science* **286**: 120-3.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273**: 1058-73.

Burri, M., W. Schlimme, B. Betschart, U. Kampfer, J. Schaller and H. Hecker (1993). Biochemical and functional characterization of histone H1-like proteins in procyclic *Trypanosoma brucei brucei. Parasitol Res* **79**: 649-59.

Caceres, C., P. Gimenez-Bonafe, E. Ribes, D. Wouters Tyrou, A. Martinage, M. Kouach, P. Sautiere, S. Muller, J. Palau, J.A. Subirana, et al. (1999). DNA-interacting proteins in the spermiogenesis of the mollusc *Murex brandaris. J Biol Chem* **274**: 649-56.

Camerini-Otero, R.D., B. Sollner-Webb and G. Felsenfeld (1976). The organization of histones and DNA in chromatin: evidence for an arginine-rich histone kernel. *Cell* **8**: 333-47.

Caplan, E.B. (1975). A very rapidly migrating f1 histone associated with gene-sized pieces of DNA in the macronucleus of *Oxytricha* sp. *Biochim Biophys Acta* **407**: 109-13.

Carlos, S., D.F. Hunt, C. Rocchini, D.P. Arnott and J. Ausió (1993a). Post-translational cleavage of a histone H1-like protein in the sperm of *Mytilus. J Biol Chem* **268**: 195-9.

Carlos, S., L. Jutglar, I. Borrell, D.F. Hunt and J. Ausió (1993b). Sequence and characterization of a sperm-specific histone H1-like protein of *Mytilus californianus. J Biol Chem* **268**: 185-94.

Casas, M.T., J. Ausió and J.A. Subirana (1993). Chromatin fibers with different protamine and histone compositions. *Exp Cell Res* **204**: 192-7.

Charlesworth, M.C. and R.W. Parish (1977). Further studies on basic nucleoproteins from the cellular slime mold *Dictyostelium discoideum. Eur J Biochem* **75**: 241-50.

Chikhirzhina, E.V., E.I. Kostyleva, E.I. Ramm and V.I. Vorobiev (1998). [Chromatin compactification using a model system of DNA-protein complexes]. *Tsitologiia* **40**: 883-8.

Chiva, M., M. Daban, E. Rosenberg and H.E. Kasinsky (1991). Protamines in polyplacophors and gastropods as a model for evolutionary changes in molluscan

sperm proteins. *In* Comparative spermatology 20 years after. B. Baccetti. New York, Raven Press: 27-30.

Chiva, M., F. Lafargue, E. Rosenberg and H.E. Kasinsky (1992). Protamines, not histones, are the predominant basic proteins in sperm nuclei of solitary ascidian tunicates. *J. Exp. Zool.* **263**: 338-349.

Chiva, M., N. Saperas, C. Caceres and J. Ausió (1995). Nuclear basic proteins from the sperm of tunicates, cephalochordates, agnathans and fish. Paris, Editions du Museum national d'Histoire naturelle.

Cho, C., W.D. Willis, E.H. Goulding, H. Jung Ha, Y.C. Choi, N.B. Hecht and E.M. Eddy (2001). Haploinsufficiency of protamine-1 or -2 causes infertility in mice. *Nat Genet* **28**: 82-6.

Choudhary, S.K., S.M. Wykes, J.A. Kramer, A.N. Mohamed, F. Koppitch, J.E. Nelson and S.A. Krawetz (1995). A haploid expressed gene cluster exists as a single chromatin domain in human sperm. *J Biol Chem* **270**: 8755-62.

Churchill, M.E. and A.A. Travers (1991). Protein motifs that recognize structural features of DNA. *Trends Biochem Sci* **16**: 92-7.

Cirillo, L.A., C.E. McPherson, P. Bossard, K. Stevens, S. Cherian, E.Y. Shim, K.L. Clark, S.K. Burley and K.S. Zaret (1998). Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J* **17**: 244-54.

Clark, A.G. and A. Civetta (2000). Evolutionary biology. Protamine wars. *Nature* **403**: 261, 263.

Clark, D.J. and J.O. Thomas (1988). Differences in the binding of H1 variants to DNA. Cooperativity and linker-length related distribution. *Eur J Biochem* **178**: 225-33.

Cobb, J., R.K. Reddy, C. Park and M.A. Handel (1997). Analysis of expression and function of topoisomerase I and II during meiosis in male mice. *Mol Reprod Dev* **46**: 489-98.

Cole, R.D. (1984). A minireview of microheterogeneity in H1 histone and its possible significance. *Anal Biochem* **136**: 24-30.

Daban, M., A. Martinage, M. Kouach, M. Chiva, J.A. Subirana and P. Sautiere (1995). Sequence analysis and structural features of the largest known protamine isolated

from the sperm of the archaeogastropod *Monodonta turbinata*. *J Mol Evol* **40**: 663-70.

Dacks, J. and H.E. Kasinsky (1999). Nuclear condensation in protozoan gametes and the evolution of anisogamy. *Comp. Biochem. Physiol. A* **124**: 287-295.

Dacks, J. and A.J. Roger (1999). The first sexual lineage and the relevance of facultative sex. *J Mol Evol* **48**: 779-83.

de Yebra, L., J.L. Ballesca, J.A. Vanrell, L. Bassas and R. Oliva (1993). Complete selective absence of protamine P2 in humans. *J Biol Chem* **268**: 10553-7.

Domenjoud, L., G. Nussbaum, I.M. Adham, G. Greeske and W. Engel (1990). Genomic sequences of human protamines whose genes, PRM1 and PRM2, are clustered. *Genomics* **8**: 127-33.

Doolittle, R.F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem Sci* **19**: 15-8.

Doolittle, W.F. (1999). Phylogenetic classification and the universal tree. *Science* **284**: 2124-9.

Duschak, V.G. and J.J. Cazzulo (1990). The histones of the insect trypanosomatid, *Crithidia fasciculata*. *Biochim Biophys Acta* **1040**: 159-66.

Eickbush, T.H. and E.N. Moudrianakis (1978). The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry* **17**: 4955-64.

Faguy, D.M. and W.F. Doolittle (1999). Lessons from the *Aeropyrum pernix* genome. *Curr Biol* **9**: R883-6.

Felix, K. (1960). Protamines. *Adv Prot Chem* **15**: 1-56.

Folco, H.D., M. Freitag, A. Ramon, E.D. Temporini, M.E. Alvarez, I. Garcia, C. Scazzocchio, E.U. Selker and A.L. Rosa (2003). Histone H1 Is required for proper regulation of pyruvate decarboxylase gene expression in *Neurospora crassa*. *Eukaryot Cell* **2**: 341-50.

Fu, X.D. (1995). The superfamily of arginine/serine-rich splicing factors. *RNA* **1**: 663-80.

Garcia-Ramirez, M. and J.A. Subirana (1994). Condensation of DNA by basic-proteins does not depend on protein-composition. *Biopolymers* **34**: 285-292.

Gardiner Garden, M., M. Ballesteros, M. Gordon and P.P. Tam (1998). Histone- and protamine-DNA association: conservation of different patterns within the beta-globin domain in human sperm. *Mol Cell Biol* **18**: 3350-6.

Garel, A., A. Mazen, M. Champagne, P. Sautiere, D. Kmiecik, O. Loy and G. Biserte (1975). Chicken erythrocyte histone H5; I. Amino terminal sequence (70 residues). *FEBS Lett* **50**: 195-9.

Gatewood, J.M., G.R. Cook, R. Balhorn, E.M. Bradbury and C.W. Schmid (1987). Sequence-specific packaging of DNA in human sperm chromatin. *Science* **236**: 962-4.

Gatewood, J.M., G.R. Cook, R. Balhorn, C.W. Schmid and E.M. Bradbury (1990). Isolation of four core histones from human sperm chromatin representing a minor subset of somatic histones. *J Biol Chem* **265**: 20662-6.

Giancotti, V., E. Russo, M. Gasparini, D. Serrano, D. Del Piero, A.W. Thorne, P.D. Cary and C. Crane Robinson (1983). Proteins from the sperm of the bivalve mollusc *Ensis minor*. Co-existence of histones and a protamine-like protein. *Eur J Biochem* **136**: 509-16.

Gimenez-Bonafe, P., M. Laszczak, H.E. Kasinsky, M.J. Lemke, J.D. Lewis, M. Iskandar, T. He, M.G. Ikonomou, F.M. White, D.F. Hunt, M. Chiva and J. Ausió (2000). Characterization and evolutionary relevance of the sperm nuclear basic proteins from stickleback fish. *Mol Reprod Dev* **57**: 185-93.

Gimenez-Bonafe, P., E. Ribes, P. Sautiere, A. Gonzalez, H. Kasinsky, M. Kouach, P.E. Sautiere, J. Ausió and M. Chiva (2002). Chromatin condensation, cysteine-rich protamine, and establishment of disulphide interprotamine bonds during spermiogenesis of *Eledone cirrhosa* (Cephalopoda). *Eur J Cell Biol* **81**: 341-9.

Giorgini, F., H.G. Davies and R.E. Braun (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**: 7010-9.

Giorgini, F., H.G. Davies and R.E. Braun (2002). Translational repression by MSY4 inhibits spermatid differentiation in mice. *Development* **129**: 3669-79.

Green, G.R. (2001). Phosphorylation of histone variant regions in chromatin: unlocking the linker? *Biochem Cell Biol* **79**: 275-87.

Gui, J.F., W.S. Lane and X.D. Fu (1994). A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369**: 678-82.

Gusse, M. and P. Chevaillier (1978). [Ultrastructural and chemical study of the chromatin during spermiogenesis of the fish *Scyliorhinus caniculus* (L.) (author's transl)]. *Cytobiologie* **16**: 421-43.

Hackstadt, T., W. Baehr and Y. Ying (1991). *Chlamydia trachomatis* developmentally regulated protein is homologous to eukaryotic histone H1. *Proc Natl Acad Sci U S A* **88**: 3937-41.

Hansen, J.C. and J. Ausió (1992). Chromatin dynamics and the modulation of genetic activity. *Trends Biochem Sci* **17**: 187-91.

Hausmann, K. and N. Hülsmann (1996). Protozoology, 2nd Edition. New York, Georg Thieme Verlag.

Hayashi, T., H. Hayashi and K. Iwai (1987). Tetrahymena histone H1. Isolation and amino acid sequence lacking the central hydrophobic domain conserved in other H1 histones. *J Biochem (Tokyo)* **102**: 369-76.

Heath, D.D. and T.J. Hilbish (1998). *Mytilus* protamine-like sperm-specific protein genes are multicopy, dispersed, and closely associated with hypervariable RFLP regions. *Genome* **41**: 587-96.

Hecht, N.B. (1989). Mammalian protamines and their expression. Boca Raton, FL, CRC Press Inc.

Hecht, N.B., P.A. Bower, K.C. Kleene and R.J. Distel (1985). Size changes of protamine 1 mRNA provide a molecular marker to monitor spermatogenesis in wild-type and mutant mice. *Differentiation* **29**: 189-93.

Hoffmann, J.A., R.E. Chance and M.G. Johnson (1990). Purification and analysis of the major components of chum salmon protamine contained in insulin formulations using high-performance liquid chromatography. *Protein Expr Purif* **1**: 127-33.

Hunt, J.G., H.E. Kasinsky, R.M. Elsey, C.L. Wright, P. Rice, J.E. Bell, D.J. Sharp, A.J. Kiss, D.F. Hunt, D.P. Arnott, et al. (1996). Protamines of reptiles. *J Biol Chem* **271**: 23547-57.

Isenberg, I. (1978). Histones. *In* The Cell Nucleus. H. Busch. New York, Academic Press: 135-154.

Itoh, T., J. Ausió and C. Katagiri (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**: 181-90.

Iwai, K. (1964). Histones of rice embryos and of *Chlorella*. *In* The Nucleohistones. J. Bonner and P. Ts'o. San Francisco, Holden-Day: 59-65.

Jankowski, J.M., J.C. States and G.H. Dixon (1986). Evidence of sequences resembling avian retrovirus long terminal repeats flanking the trout protamine gene. *J Mol Evol* **23**: 1-10.

Jardine, N.J. and J.L. Leaver (1978). The fractionation of histones isolated from *Euglena gracilis*. *Biochem J* **169**: 103-11.

Johns, E.W. (1971). The preparation and characterization of histones. *In* Histones and Nucleohistones. D. M. P. Phillips. London, Plenum Press: 1-45.

Jutglar, L., J.I. Borrell and J. Ausió (1991). Primary, secondary, and tertiary structure of the core of a histone H1-like protein from the sperm of *Mytilus*. *J Biol Chem* **266**: 8184-91.

Kadura, S.N. and S.N. Khrapunov (1988). Displacement of histones by sperm-specific proteins at different stages of spermatogenesis of squid. *Eur J Biochem* **175**: 603-7.

Kadura, S.N., S.N. Khrapunov, V.N. Chabanny and G.D. Berdyshev (1983). Changes in chromatin basic proteins during male gametogenesis of grass carp. *Comp Biochem Physiol B* **74**: 343-50.

Kasinsky, H.E. (1989). Specificity and distribution of sperm basic proteins. Boca Raton, FL, CRC Press Inc.

Kasinsky, H.E., S.Y. Huang, M. Mann, J. Roca and J.A. Subirana (1985). On the diversity of sperm histones in the vertebrates: IV. Cytochemical and amino acid analysis in Anura. *J Exp Zool* **234**: 33-46.

Kasinsky, H.E., J.D. Lewis, J.B. Dacks and J. Ausió (2001). Origin of H1 linker histones. *FASEB J* **15**: 34-42.

Kawarabayasi, Y., Y. Hino, H. Horikawa, S. Yamazaki, Y. Haikawa, K. Jin no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, et al. (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res* **6**: 83-101, 145-52.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, et al. (1998). Complete sequence and

gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res* **5**: 55-76.

Kennedy, B.P. and P.L. Davies (1985). Sites of phosphorylation on the high molecular weight basic nuclear proteins of the winter flounder. *J Biol Chem* **260**: 4338-44.

Kimura, Y. and R. Yanagimachi (1995). Mouse oocytes injected with testicular spermatozoa or round spermatids can develop into normal offspring. *Development* **121**: 2397-405.

Kistler, W.S., K. Henriksen, P. Mali and M. Parvinen (1996). Sequential expression of nucleoproteins during rat spermiogenesis. *Exp Cell Res* **225**: 374-81.

Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson, et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**: 364-70.

Kosikov, K.M., A.A. Gorin, X.J. Lu, W.K. Olson and G.S. Manning (2002). Bending of DNA by asymmetric charge neutralization: all-atom energy simulations. *J Am Chem Soc* **124**: 4838-47.

Kossel, A. (1896a). Uber die basischen Stoffe des Zellkerns. *S-B Kgl Preuss Akad Wiss* **18**: 403-408.

Kossel, A. (1896b). Uber die basischen Stoffe des Zellkerns. *Z Physiol Chem* **22**: 176-287.

Kossel, A. (1928). The protamines and histones. London, Longmans Green and Co.

Kumar, S., K. Tamura, I.B. Jakobsen and M. Nei (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**: 1244-5.

Labourier, E., J.F. Riou, M. Prudhomme, C. Carrasco, C. Bailly and J. Tazi (1999). Poisoning of topoisomerase I by an antitumor indolocarbazole drug: stabilization of topoisomerase I-DNA covalent complexes and specific inhibition of the protein kinase activity. *Cancer Res* **59**: 52-5.

Landsman, D. (1996). Histone H1 in *Saccharomyces cerevisiae*: a double mystery solved? *Trends Biochem Sci* **21**: 287-8.

Lennard, A.C. and J.O. Thomas (1985). The arrangement of histone H5 molecules in extended and condensed chicken erythrocyte chromatin. *EMBO J.* **4**: 3455-3462.

Lewis, J.D., D.W. Abbott and J. Ausió (2003a). Sex and chromatin; modified and variant histones in meiosis and spermatogenesis. *Recent Res Devel Mol Cell Biol* **3**: 439-453.

Lewis, J.D. and J. Ausió (2002). Protamine-like proteins: evidence for a novel chromatin structure. *Biochem Cell Biol* **80**: 353-61.

Lewis, J.D., Y. Song, M.E. De Jong, S.M. Bagha and J. Ausió (2003b). A walk through vertebrate and invertebrate protamines. *Chromosoma* **111**: 473-82.

Libertini, L.J., J. Ausió, K.E. van Holde and E.W. Small (1988). Highly cooperative binding to DNA by a histone-like, sperm-specific protein from *Spisula solidissima*. *Biopolymers* **27**: 1459-77.

Luger, K., A.W. Mader, R.K. Richmond, D.F. Sargent and T.J. Richmond (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-60.

Maier, W.M., G. Nussbaum, L. Domenjoud, U. Klemm and W. Engel (1990). The lack of protamine 2 (P2) in boar and bull spermatozoa is due to mutations within the P2 gene. *Nucleic Acids Res* **18**: 1249-54.

Maleszewski, M., S. Kuretake, D. Evenson, H. Yanagimachi, J. Bjordahl and R. Yanagimachi (1998). Behavior of transgenic mouse spermatozoa with galline protamine. *Biol Reprod* **58**: 8-14.

Mende, L.M., J.H. Waterborg, R.D. Mueller and H.R. Matthews (1983). Isolation, identification, and characterization of histones from plasmodia of the true slime mold *Physarum polycephalum* using extraction with guanidine hydrochloride. *Biochemistry* **22**: 38-51.

Miescher, F. (1874). Das Protamin, eine neue organische Base aus den Samenf den des Rheinlachses. *Ber Dtsch Chem Gesellschaft* **7**: 376-379.

Mogensen, C., S. Carlos and J. Ausió (1991). Microheterogeneity and interspecific variability of the nuclear sperm proteins from *Mytilus*. *FEBS Lett* **282**: 273-6.

Morris, R.L., A.P. Salinger and P.J. Rizzo (1999). Analysis of lysine-rich histones from the unicellular green alga *Chlamydomonas reinhardtii*. *J Eukaryot Microbiol* **46**: 648-54.

Nakano, M., T. Tobita and T. Ando (1975). Studies on a protamine (galline) from fowl sperm. 2. The amino acid sequences of two components of galline. *Int J Pept Protein Res* **7**: 31-46.

Neelin, J.M. (1968). Variability in cell-specific and common histones of avian erythrocytes. *Can J Biochem* **46**: 241-7.

Neelin, J.M., P.X. Callahan, D.C. Lamb and K. Murray (1964). The histones of chicken erythrocyte nuclei. *Can. J. Biochem.* **42**: 1743-1752.

Nelson, P.P., S.C. Albright, J.M. Wiseman and W.T. Garrard (1979). Reassociation of histone H1 with nucleosomes. *J Biol Chem* **254**: 11751-60.

Ohno, S. and M.L. Becak (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**: 7341-5.

Ohtsuki, K., Y. Nishikawa, H. Saito, H. Munakata and T. Kato (1996). DNA-binding sperm proteins with oligo-arginine clusters function as potent activators for egg CK-II. *FEBS Lett* **378**: 115-20.

Oliva, R. (1995). Sequence, evolution and transcriptional regulation of avian-mammalian P1 type protamines. *In* Advances in spermatozoal phylogeny and taxonomy. B. G. M. Jamieson, J. Ausió and J. L. Justine. Paris, Editions du Museum National d'Histoire Naturelle**:** 537-548.

Oliva, R. and G.H. Dixon (1990). Vertebrate protamine gene evolution I. Sequence alignments and gene structure. *J Mol Evol* **30**: 333-46.

Oliva, R. and G.H. Dixon (1991). Vertebrate protamine genes and the histone-to-protamine replacement reaction. *Prog Nucleic Acid Res Mol Biol* **40**: 25-94.

Oliva, R., J. Mezquita, C. Mezquita and G.H. Dixon (1988). Haploid expression of the rooster protamine mRNA in the postmeiotic stages of spermatogenesis. *Dev Biol* **125**: 332-40.

Papoutsopoulou, S., E. Nikolakaki, G. Chalepakis, V. Kruft, P. Chevaillier and T. Giannakouros (1999). SR protein-specific kinase 1 is highly expressed in testis and phosphorylates protamine 1. *Nucleic Acids Res* **27**: 2972-80.

Patterton, H.G., C.C. Landel, D. Landsman, C.L. Peterson and R.T. Simpson (1998). The biochemical and phenotypic characterization of Hho1p, the putative linker histone H1 of *Saccharomyces cerevisiae*. *J Biol Chem* **273**: 7268-76.

Pellegrini, M. and T.O. Yeates (1999). Searching for frameshift evolutionary relationships between protein sequence families. *Proteins* **37**: 278-83.

Peretti, M. and S. Khochbin (1997). The evolution of the differentiation-specific histone H1 gene basal promoter. *J Mol Evol* **44**: 128-34.

Phelan, J.J., J. Colom, C. Cozcolluela, J.A. Subirana and R.D. Cole (1974). A lysine-rich protein from spermatozoa of the mollusc *Mytilus edulis*. *J Biol Chem* **249**: 1099-102.

Poccia, D.L. and G.R. Green (1992). Packaging and unpackaging the sea urchin sperm genome. *Trends Biochem Sci* **17**: 223-7.

Ponting, C.P., J. Schultz, F. Milpetz and P. Bork (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* **27**: 229-32.

Puigdomenech, P., P. Martinez, J. Palau, E.M. Bradbury and C. Crane Robinson (1976). Studies on the role and mode of operation of the very-lysine-rich histones in eukaryote chromatin. Nuclear-magnetic-resonance studies on nucleoprotein and histone phi 1-DNA complexes from marine invertebrate sperm. *Eur J Biochem* **65**: 357-63.

Ramakrishnan, V. (1997). Histone H1 and chromatin higher-order structure. *Crit Rev Eukaryot Gene Expr* **7**: 215-30.

Ramakrishnan, V., J.T. Finch, V. Graziano, P.L. Lee and R.M. Sweet (1993). Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**: 219-23.

Ramon, A., M.I. Muro Pastor, C. Scazzocchio and R. Gonzalez (2000). Deletion of the unique gene encoding a typical histone H1 has no apparent phenotype in *Aspergillus nidulans*. *Mol Microbiol* **35**: 223-33.

Raukas, E. and R.H. Mikelsaar (1999). Are there molecules of nucleoprotamine? *Bioessays* **21**: 440-8.

Retief, J.D. and G.H. Dixon (1993). Evolution of pro-protamine P2 genes in primates. *Eur J Biochem* **214**: 609-15.

Retief, J.D., R.J. Winkfein and G.H. Dixon (1993a). Evolution of the monotremes. The sequences of the protamine P1 genes of platypus and echidna. *Eur J Biochem* **218**: 457-61.

Retief, J.D., R.J. Winkfein, G.H. Dixon, R. Adroer, R. Queralt, J. Ballabriga and R. Oliva (1993b). Evolution of protamine P1 genes in primates. *J Mol Evol* **37**: 426-34.

Reynolds, W.F. and S.L. Wolfe (1984). Protamines in plant sperm. *Exp Cell Res* **152**: 443-8.

Rhim, J.A., W. Connor, G.H. Dixon, C.J. Harendza, D.P. Evenson, R.D. Palmiter and R.L. Brinster (1995). Expression of an avian protamine in transgenic mice disrupts chromatin structure in spermatozoa. *Biol Reprod* **52**: 20-32.

Rice, P., R. Garduno, T. Itoh, C. Katagiri and J. Ausió (1995). Nucleoplasmin-mediated decondensation of Mytilus sperm chromatin. Identification and partial characterization of a nucleoplasmin-like protein with sperm-nuclei decondensing activity in *Mytilus californianus*. *Biochemistry* **34**: 7563-8.

Rizzo, P.J., W. Bradley and R.L. Morris (1985). Histones of the unicellular algae *Olisthodiscus luteus*. *Biochemistry* **24**: 1727-1732.

Rocchini, C., R.M. Marx, J.S. Carosfeld, H.E. Kasinsky, E. Rosenberg, F. Sommer and J. Ausió (1996). Replacement of nucleosomal histones by histone H1-like proteins during spermiogenesis in Cnidaria: evolutionary implications. *J Mol Evol* **39**: 240-246.

Rocchini, C., P. Rice and J. Ausió (1995a). Complete sequence and characterization of the major sperm nuclear basic protein from *Mytilus trossulus*. *FEBS Lett* **363**: 37-40.

Rocchini, C., F. Zhang and J. Ausió (1995b). Two highly specialized histone H1 proteins are the major chromosomal proteins of the sperm of the sea anemone *Urticina* (*Tealia*) *crassicornis*. *Biochemistry* **34**: 15704-12.

Rooney, A.P. and J. Zhang (1999). Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? *Mol Biol Evol* **16**: 706-10.

Rooney, A.P., J. Zhang and M. Nei (2000). An unusual form of purifying selection in a sperm protein. *Mol Biol Evol* **17**: 278-83.

Rossi, F., E. Labourier, T. Forne, G. Divita, J. Derancourt, J.F. Riou, E. Antoine, G. Cathala, C. Brunel and J. Tazi (1996). Specific phosphorylation of SR proteins by mammalian DNA topoisomerase I. *Nature* **381**: 80-2.

Ruiz Lara, S., E. Prats, M.T. Casas and L. Cornudella (1993). Molecular cloning and sequence of a cDNA for the sperm-specific protein phi 1 from the mussel *Mytilus edulis*. *Nucleic Acids Res* **21**: 2774.

Sala Rovira, M., M.L. Geraud, D. Caput, F. Jacques, M.O. Soyer Gobillard, G. Vernet and M. Herzog (1991). Molecular cloning and immunolocalization of two variants of the major basic nuclear protein (HCc) from the histone-less eukaryote *Crypthecodinium cohnii* (Pyrrhophyta). *Chromosoma* **100**: 510-8.

Sambrook, J., E.F. Fritsch and T. Maniatus (1989). Molecular Cloning: A laboratory manual (2nd Ed.). New York, Cold Spring Harbour.

Sanger, F., S. Nicklen and A.R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-7.

Saperas, N., J. Ausió, D. Lloris and M. Chiva (1994). On the evolution of protamines in bony fish: alternatives to the "retroviral horizontal transmission" hypothesis. *J Mol Evol* **39**: 282-95.

Saperas, N., M. Chiva and J. Ausió (1992). Purification and characterization of the protamines and related proteins from the sperm of a Tunicate. *Comp Biochem Physiol* **103B**: 969-974.

Saperas, N., M. Chiva, D.C. Pfeiffer, H.E. Kasinsky and J. Ausió (1997). Sperm nuclear basic proteins (SNBPs) of agnathans and chondrichthyans: variability and evolution of sperm proteins in fish. *J Mol Evol* **44**: 422-31.

Sauer, R.T., R.R. Yocum, R.F. Doolittle, M. Lewis and C.O. Pabo (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature* **298**: 447-51.

Sautiere, P., D. Kmiecik, O. Loy, G. Briand, G. Biserte, A. Garel and M. Champagne (1975). Chicken erythrocyte histone H5 II. Amino acid sequence adjacent to the phenylalanine residue. *FEBS Lett* **50**: 200-3.

Schlake, T., M. Schorpp and T. Boehm (2000). Formation of regulator/target gene relationships during evolution. *Gene* **256**: 29-34.

Schneider, T.D. and R.M. Stephens (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**: 6097-100.

Schwede, T., J. Kopp, N. Guex and M.C. Peitsch (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* **31**: 3381-5.

Segers, A., S. Muyldermans and L. Wyns (1991). The interaction of histone H5 and its globular domain with core particles, depleted chromatosomes, polynucleosomes, and a DNA decamer. *J Biol Chem* **266**: 1502-8.

Seyedin, S.M. and W.S. Kistler (1980). Isolation and characterization of rat testis H1t. An H1 histone variant associated with spermatogenesis. *J Biol Chem* **255**: 5949-54.

Shimada, A., K. Ohsumi and T. Kishimoto (1998). An indirect role for cyclin B-Cdc2 in inducing chromosome condensation in *Xenopus* egg extracts. *Biol Cell* **90**: 519-30.

Smith, D.R., L.A. Doucette Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, et al. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J Bacteriol* **179**: 7135-55.

Smith, T.F. and M.S. Waterman (1981). Identification of common molecular subsequences. *J Mol Biol* **147**: 195-7.

Stedman, E. and E. Stedman (1947). The chemical nature and functions of the components of cell nuclei. *Cold Spring Harbour Symp. Quant. Biol.* **12**: 224-236.

Steger, K. (1999). Transcriptional and translational regulation of gene expression in haploid spermatids. *Anat Embryol (Berl)* **199**: 471-87.

Stein, A. (1979). DNA folding by histones: the kinetics of chromatin core particle reassembly and the interaction of nucleosomes with histones. *J Mol Biol* **130**: 103-34.

Stiller, J.W. and B.D. Hall (1999). Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol* **16**: 1270-9.

Subirana, J.A. (1983). Nuclear proteins in spermatozoa and their interactions with DNA. *In* Fourth international symposium on spermatology. J. Andre. The Hague, Matinus Nijhoff: 197-214.

Subirana, J.A. (1990). Analysis of the charge distribution in the C-terminal region of histone H1 as related to its interaction with DNA. *Biopolymers* **29**: 1351-7.

Subirana, J.A. (1991). Protein-DNA interactions in spermatozoa. *In* Comparative spermatology 20 years after. B. Baccetti. New York, Raven Press: 89-92.

Subirana, J.A. and J. Colom (1987). Comparison of protamines from freshwater and marine bivalve molluscs: evolutionary implications. *FEBS Lett* **220**: 193-196.

Subirana, J.A., C. Cozcolluela, J. Palau and M. Unzeta (1973). Protamines and other basic proteins from spermatozoa of molluscs. *Biochim Biophys Acta* **317**: 364-379.

Suzuki, M. and T. Wakabayashi (1988). Packaging of DNA in cricket sperm. A compact mode of DNA packaging. *J Mol Biol* **204**: 653-61.

Swanson, W.J. and V.D. Vacquier (2002). The rapid evolution of reproductive proteins. *Nat Rev Genet* **3**: 137-44.

Takamune, K., H. Nishida, M. Takai and C. Katagiri (1991). Primary structure of toad sperm protamines and nucleotide sequence of their cDNAs. *Eur J Biochem* **196**: 401-6.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-82.

Toro, G.C. and N. Galanti (1988). H 1 histone and histone variants in *Trypanosoma cruzi*. *Exp Cell Res* **174**: 16-24.

Trewitt, P.M., L.J. Heilmann and A.K. Kumaran (1990). Boll weevil testis-specific cDNA. *Nucleic Acids Res* **18**: 3646.

Ushinsky, S.C., H. Bussey, A.A. Ahmed, Y. Wang, J. Friesen, B.A. Williams and R.K. Storms (1997). Histone H1 in *Saccharomyces cerevisiae*. *Yeast* **13**: 151-61.

van Holde, K.E. (1989). Chromatin. New York, Springer-Verlag.

Vaughn, J.C., J. Chaitoff, R. Deleon, C. Garland and L. Thomson (1969). Changing nuclear histone patterns during development. II. Isolation and partial characterization of "decapodine" from sperm cells of the crab *Emerita analoga*. *Exp Cell Res* **54**: 362-6.

Vernet, G., M. Sala Rovira, M. Maeder, F. Jacques and M. Herzog (1990). Basic nuclear proteins of the histone-less eukaryote *Crypthecodinium cohnii* (Pyrrhophyta): two-dimensional electrophoresis and DNA-binding properties. *Biochim Biophys Acta* **1048**: 281-9.

Ward, W.S. and D.S. Coffey (1991). DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells. *Biol Reprod* **44**: 569-74.

Ward, W.S. and A.O. Zalensky (1996). The unique, complex organization of the transcriptionally silent sperm chromatin. *Crit Rev Eukaryot Gene Expr* **6**: 139-47.

Watson, C.E. and P.L. Davies (1998). The high molecular weight chromatin proteins of winter flounder sperm are related to an extreme histone H1 variant. *J Biol Chem* **273**: 6157-62.

Watson, C.E. and P.L. Davies (1999). Recent and rapid amplification of the sperm basic nuclear protein genes in winter flounder. *Biochim Biophys Acta* **1444**: 337-45.

Wells, D. and D. Brown (1991). Histone and histone gene compilation and alignment update. *Nucleic Acids Res* **19 Suppl**: 2173-88.

Winkfein, R.J., S. Nishikawa, W. Connor and G.H. Dixon (1993). Characterization of a marsupial sperm protamine gene and its transcripts from the North American opossum (*Didelphis marsupialis*). *Eur J Biochem* **215**: 63-72.

Wolffe, A.P. (1992). Chromatin: structure and function. New York, Academic Press.

Wolffe, A.P. (1994). Structural and functional properties of the evolutionarily ancient Y-box family of nucleic acid binding proteins. *Bioessays* **16**: 245-51.

Wouters Tyrou, D., M.C. Chartier Harlin, A. Martin Ponthieu, C. Boutillon, A. Van Dorsselaer and P. Sautiere (1991). Cuttlefish spermatid-specific protein T. Molecular characterization of two variants T1 and T2, putative precursors of sperm protamine variants Sp1 and Sp2. *J Biol Chem* **266**: 17388-95.

Wouters Tyrou, D., A. Martin Ponthieu, N. Ledoux Andula, M. Kouach, M. Jaquinod, J.A. Subirana and P. Sautiere (1995). Squid spermiogenesis: molecular characterization of testis-specific pro-protamines. *Biochem J* **309 (Pt 2)**: 529-34.

Wouters Tyrou, D., A. Martinage, P. Chevaillier and P. Sautiere (1998). Nuclear basic proteins in spermiogenesis. *Biochimie* **80**: 117-28.

Wu, J. and M. Grunstein (2000). 25 years after the nucleosome model: chromatin modifications. *Trends Biochem Sci* **25**: 619-23.

Wu, M., C.D. Allis, M.T. Sweet, R.G. Cook, T.H. Thatcher and M.A. Gorovsky (1994). Four distinct and unusual linker proteins in a mitotically dividing nucleus are derived from a 71-kilodalton polyprotein, lack p34cdc2 sites, and contain protein kinase A sites. *Mol Cell Biol* **14**: 10-20.

Wyckoff, G.J., W. Wang and C.I. Wu (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**: 304-9.

Wykes, S.M. and S.A. Krawetz (2003). The structural organization of sperm chromatin. *J Biol Chem* **278**: 29471-7.

Yiu, G.K. and N.B. Hecht (1997). Novel testis-specific protein-DNA interactions activate transcription of the mouse protamine 2 gene during spermatogenesis. *J Biol Chem* **272**: 26926-33.

Yoshinobu, K., T. Kondo, M. Takai, C. Katagiri, H. Tou, S.I. Abe and K. Takamune (1997). Primary structures of sperm-specific basic nuclear proteins and gene expression in Japanese newt, *Cynops pyrrhogaster*. *Mol Reprod Dev* **46**: 243-51.

Zalenskaya, I.A., N.A. Odintsova, A.O. Zalensky and V.I. Vorobiev (1982). [Nucleosomal organization of chromatin from sperm of the bivalve mollusk *Swiftopecten swifti*]. *Mol Biol (Mosk)* **16**: 335-44.

Zalenskaya, I.A., E.O. Zalenskaya and A.O. Zalensky (1980). Basic chromosomal-proteins of marine-invertebrates. 2. Starfish and holothuria. *Comp. Biochem. Physiol. B* **65**: 375-378.

Zalensky, A.O., M.J. Allen, A. Kobayashi, I.A. Zalenskaya, R. Balhorn and E.M. Bradbury (1995). Well-defined genome architecture in the human sperm nucleus. *Chromosoma* **103**: 577-90.

Zhang, F., J.D. Lewis and J. Ausió (1999). Cysteine-containing histone H1-like (PL-I) proteins of sperm. *Mol Reprod Dev* **54**: 402-9.

Zhang, Z. and S.J. Gurr (2000). Walking into the unknown: a 'step down' PCR-based technique leading to the direct sequence analysis of flanking genomic DNA. *Gene* **253**: 145-50.

Zhou, Y.B., S.E. Gerchman, V. Ramakrishnan, A. Travers and S. Muyldermans (1998). Position and orientation of the globular domain of linker histone H5 on the nucleosome. *Nature* **395**: 402-5.

Zlatanova, J. and K.E. van Holde (1998). Binding to four-way junction DNA: a common property of architectural proteins? *FASEB J* **12**: 421-31.