The biocalcification of mollusk shells and coral skeletons: Integrating molecular, proteomics and bioinformatics methods
Sequeira dos Ramos Silva, P.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 2

Novel Molluskan Biomineralization Proteins Retrieved from Proteomics: a Case Study with Upsalin

The work presented in this chapter was published in ChemBioChem, 2010, 13(7), 1067–78. doi:10.1002/cbic.201100708
Chapter 2 Novel Molluskan Proteins Retrieved from Proteomics: a Case Study with Upsalin

Authors:

Paula Ramos-Silva\textsuperscript{a,b}, Sana Benhamada\textsuperscript{a}, Nathalie Le Roy\textsuperscript{a}, Benjamin Marie\textsuperscript{a,c}, Nathalie Guichard\textsuperscript{a}, Isabelle Zanella-Cléon\textsuperscript{a}, Laurent Plasseraud\textsuperscript{d}, Marion Comeliat\textsuperscript{e}, Gérard Alcaraz\textsuperscript{f}, Jaap Kaandorp\textsuperscript{b} and Frédéric Marin\textsuperscript{a}

\textsuperscript{a} UMR CNRS 6282, Biogéosciences, Université de Bourgogne, 6 boulevard Gabriel, 21000 Dijon, France

\textsuperscript{b} Section Computational Science, Faculty of Science University of Amsterdam Science Park 904, 1098XH Amsterdam, The Netherlands

\textsuperscript{c} UMR CNRS 7245, Muséum National d'Histoire Naturelle, 75005 Paris, France

\textsuperscript{d} Service de Spectrométrie de Masse, Institut de Biologie et Chimie des Protéines, Université Lyon 1 7 Passage du Vercors, 69367 Lyon, France

\textsuperscript{e} UMR CNRS 5260, ICMUB, Université de Bourgogne 21000 Dijon, France

\textsuperscript{f} UPSP PROXISS, Département Agronomie Environnement, AgroSupDijon 26 boulevard Dr. Petitjean, 21079 Dijon, France
2.1 Introduction

The formation of the molluskan shell is regulated by an array of extracellular proteins secreted by the calcifying epithelial cells of the mantle. These proteins remain occluded within the recently formed biominerals. To date, many shell proteins have been retrieved, but only a few of them, such as nacreins, have clear identified functions. In the present case, by combining molecular biology and biochemical approaches, we performed the molecular characterization of a novel protein, named Upsalin, associated with the nacreous shell of the freshwater mussel *Unio pictorum*. The full sequence of the upsalin transcript was obtained by RT-PCR and 5'/3' RACE, and the expression pattern of the transcript was studied by PCR and qPCR. Upsalin is a 12 kDa protein with a basic theoretical pI. The presence of Upsalin in the shell was demonstrated by extraction of the acetic-acid-soluble nacre matrix (ASM), purification of a shell protein fraction by monodimensional preparative SDS-PAGE, and by submitting this fraction, after trypsic digestion, to nano-LC-MS/MS. *In vitro* experiments with the purified protein showed that it interferes poorly with the precipitation of calcium carbonate. Homology searches also could not affiliate Upsalin to any other protein of known function, leaving open the question of its exact role in shell formation. An antibody raised against an immunogenic peptide of Upsalin was found to be specific and was subsequently assayed for immunogold localization of the target protein in the shell, revealing the ubiquitous presence of Upsalin in the nacreous and prismatic layers. Recently, with the increasing request of high-throughput proteomic studies to the organic fractions composing the shells, the number of candidate proteins without clear functions has been increasing exponentially.
The Upsalin example highlights the crucial need for the scientific community dealing with biomineralization in general, to dedicate the coming years in exploring new experimental approaches, such as gene silencing and functional assays, that focus specifically on the functions of mineral-associated proteins.

2.2 Background

Mollusks are known for their ability to synthesize the shell, a calcified structure formed outside their living tissues. The mollusk shell, which is mainly composed of CaCO$_3$ crystals, plays an essential role in supporting the soft body and in protecting it against predation and desiccation. The shell formation process is initiated at the early stage of larval development (trochophore) and the shell continues to grow after metamorphosis, throughout the entire life of the animal [21].

Like other biomineralization processes, the shell formation is regulated by an extracellular organic matrix (OM), which is a complex mixture of proteins, glycoproteins, polysaccharides, pigments and presumably lipids, all secreted by the mantle tissue [21]. This matrix controls crystal formation, nucleation and growth and gets occluded within the mineral during its growth. This process results in a stable and well-packed organo-mineral assembly [20], in which proteins were early recognized as key components for shell formation [76], and for modulating calcium carbonate crystal shapes [48]. Consequently, a lot of efforts have been put on their identification and the characterization of their primary structure [1, 5–7], in order to elucidate the mechanisms by which proteins interact in shell formation [79] and to use these proteins to generate tailored composite biomaterials [80].
Classically, shell proteins are retrieved by dissolving the mineral phase of the shell with weak acid [81]. This extraction gives rise to two organic fractions, the soluble matrix and the insoluble one [21,81]. Proteins of the soluble matrix can be further fractionated by electrophoresis or HPLC, and submitted to sequence analysis via mass spectrometry [82] or Edman degradation [83]. In order to obtain the full-length primary structures, degenerate primers are designed from partial protein sequences for amplifying shell protein-encoding transcripts, extracted from mantle tissues. This one-per-one protein approach enriched the information on primary structures of many shell proteins in the last decade [21]. Today, not all shell proteins have the same status: some of them have been fully characterized at both transcriptional and protein levels, and are firmly established as shell proteins: typical examples include Prismalin-14 [84] or Pearlin [85]. Some proteins are true shell proteins, but because they have been sequenced directly, information at the transcript level is missing [82,83]. Reversely, some proteins have been characterized only at the transcript level but the demonstration of their presence in the shell remains to be done [86].

Lately, the introduction of high-throughput approaches on mollusks is changing this picture. Although genomic resources on mollusks are still scarce [87], the last three years have seen the development of molluskan mantle transcriptome projects. The generated EST datasets were recently combined with proteomic data from shell extracts [88], enabling the identification of biomineralization-related proteins at record rate. So far, this methodology has been applied to some molluskan groups: clams [89], mussels [90], oysters [17, 20] and the abalone [7].

Owing to proteomics, it was possible to identify new homologous proteins belonging to the set of sequences that exhibit conserved domains such as N66 and
Nacrein-like homologous sequences with carbonic anhydrase domains [88,90]. Papilin homologues with Kunitz-like domains [7] or even Perlucin, with a C-type lectin domain [90], are just a few examples. Also proteins characterized by the predominance of one or two amino acids or by repeated short motifs in a domain were identified, corroborating earlier studies on single proteins. Among the most popular examples are the aspartic acid-rich proteins, such as Aspein [88], which strongly interact in vitro with calcium carbonate crystals [92] and are usually considered as high capacity, low affinity calcium binding proteins [93]. Proteins with basic domains constitute another example, such as the KRMP and the glycine-rich shematrins [88], which may play a role by interacting with the negatively charged bicarbonate ions, or by anchoring the acidic proteins [15, 24]. In addition to the identification of homologous proteins, proteomics on shell extracts has revealed a set of novel unknown proteins for which structure-function relations are far from being elucidated. Such examples include MUSP-2 and 3 (edible mussel) [90], MRNP34 (pearl oyster) [95] or IMSP-1,-2,-3 (manila clam) [89]. The growing number of these shell proteins ‘orphan of function’ points out the need of their complete characterization, in order to understand their role in biomineralization.
Section 2.3 Materials and Methods

In this scope, we report the characterization – both at transcriptional and protein level - of a novel biomineralization protein, which does not exhibit any homology with previous known shell proteins. This protein, called Upsalin, is 125 residues long (~ 12 kDa) and was retrieved from mantle tissues of the freshwater mussel *Unio pictorum*. The cDNA from Upsalin was identified by a classical molecular biology approach based on degenerate primers derived from sequences of peptides described in earlier proteomic work [96]. The effective occurrence of Upsalin in the shell was confirmed by proteomics, while a specific antibody showed that Upsalin was associated to the two shell layers of *U. pictorum*: prisms and nacre. In spite of several *in vitro* and functional characterizations, we could not assign a clear function to Upsalin, in relation to shell biomineralization and homology searches only revealed the tip of the veil.

The Upsalin example emphasizes the need of studying ‘orphan proteins’ in order to identify novel functions in biomineralization that have not yet been listed in previous works.

2.3 Materials and Methods

2.3.1 Sample collection and characterization

Living adult *Unio pictorum* specimens with shells between 30–80 mm in length were collected in the stream of La Varaude (Izeure, Cote d’Or, France). The soft tissues were removed from the shell. The mantle, foot, gills, and muscles were sampled under RNAsse-free conditions and transferred to liquid nitrogen, and were stored at -80 °C until further experiments.
The fresh shells were cleaned, fragmented, and etched with EDTA (1 %, w/v) for direct microscopic observations with a Hitachi TM-1000 table-top scanning electron microscope (SEM) with an acceleration voltage of 15 kV, under a back-scattered electron mode.

### 2.3.2 General strategy for the identification of Upsalin

A previous proteomic study on the bulk nacre shell matrix of *U. pictorum* and on SDS-fractionated protein bands allowed the identification of several peptides [96]. Twenty-six of them were identified from a 12 kDa and 16 kDa bands by *de novo* sequencing. They were used for the design of degenerate primers (Figure 2.1), in order to obtain the full sequence of the corresponding transcripts. By using this approach we could identify one full transcript corresponding to the 12 kDa band. Based on the sequence, we designed a polyclonal antibody from an immunogenic peptide in order to purify the protein from nacre extracts and to localize it in shell tissues. The predicted protein sequence was confirmed by proteomic analysis of the purified extract and subsequently used for *in vitro* functional assays.
2.3.3 Identification of a cDNA fragment

The cDNA from the mantle was first screened by PCR with the degenerate primers. Amplification of cDNA was performed with GoTaq Flexi DNA Polymerase (1.25 U, Promega, USA) and an automated MJ Mini Gradient Thermal Thermocycler (USA, BioRad) by the following program: 5 min of initial denaturation at 95 °C, 1 min of denaturation at 95 °C (35x), 1 min of annealing at 55 °C (35x), 5 min of extension at 72 °C (35x) and a final extension of 5 min at 72 °C. A set of 16 degenerate primers based on eight peptide sequences (sense and antisense) obtained from our previous study was designed (Figure 2.1).

![Figure 2.1: SDS-PAGE fractionation of nacre ASM.](image)
The peptides were chosen because of their presence in the protein bands of 12 and 16 kDa obtained by SDS-PAGE analysis of the ASM. From the set of degenerate primers, one pair gave a highly specific PCR product of 250 bp (sense, 5’-GARTG YWSNG AYCCN GARGT-3’ and antisense, 5’-GGRCA RCANA CNCCR TARAA YTC-3’).

2.3.4 Rapid amplification of cDNA ends (5’- and 3’-RACE)

The cDNA ends were identified by use of the 3’-RACE (ref. 18373–019) and 5’-RACE (ref. 18374–058) Systems for Rapid Amplification of cDNA ends (Invitrogen). The procedure was carried out according to the manufacturer’s instructions with gene-specific primers based on the nucleotide sequence of the Upsalin cDNA fragment. The 3’-end product was obtained by PCR amplification of the reverse transcribed cDNA sequence with the gene-specific primer (GSP) 5’-CGTCC GTATG CGTGC CCTAG-3’, with an annealing temperature of 52 ºC. To obtain the 5’-end product, the total RNA was reverse-transcribed by use of a first antisense gene-specific primer (GSP1) 5’-AATGG ATATG CATAA GGAC-3’ and amplification of the cDNA was performed by use of a nested antisense specific primer (GSP2) 5’-GTGGC GGAAA ACCTT CAGGA-3’ with an annealing temperature of 50.4 ºC.
2.3.5 Purification, amplification and sequencing

All PCR products were run on agarose gel (1.5%, TBE 1.0x), purified with the Wizard SV gel and PCR clean-up system (Promega, ref. A9281, USA) then cloned in a pGEM®-T Easy vector system I (Promega, USA) using JM109 Competent Cells (Promega, USA). Transformed colonies, which contain the cloning vector, were selected on LB agar plates with ampicillin (100 μg/mL), at 37 °C for 12 h. From each single colony, a small portion was sampled and scattered in new numbered Petri dishes and grown in the same previously described conditions for 20 h. For selection of the positive clones, a small portion of each grown colony was taken and put in nuclease free water (10 μL). The solution was heated at 99°C for 20 min to release the plasmids and a PCR reaction was performed by using the pGEM(R)-T Easy vector specific primers: T7 5'-TAATACGACTCACTATAGGG-3' and SP6 5'-CATTTAGGTGACACTATAG-3'. This allowed identifying the positive recombinant plasmids, which were subsequently purified using the QIAprep(R) miniprep kit (Qiagen, ref. 27106, USA) and sequenced by Eurofins MWG Operon sequencing service (Eurofins MWG Operon sequencing service, Ebersberg, Germany).
2.3.6 Amplification of the full nucleotide sequence and quantitative real-time PCR

To confirm that the overlap of the sequences obtained by RACE corresponded to the predicted complete nucleotide sequence, the full-length cDNA was amplified using two specific primers based on the 5’ and 3’-ends. In order to check its presence in different tissues, RNA was extracted from mantle, gills, foot and adductor muscles. RNA was treated with DNase (Invitrogen, USA) to remove DNA contaminants then converted into cDNA. PCR and quantitative PCR were performed with the primers sense 5’- GCCGGATAGGACACCTTGAG-3’, upstream the coding region, and anti-sense 5’- ACCCGACCCTTAATGGGCAA-3’, coding for the C-terminal of the predicted protein. Standard PCR was performed with the following program: 2 min of initial denaturation at 95 °C, 1 min of denaturation at 95 °C (35x), 1 min of annealing at 60 °C (35x), 1 min of extension at 72 °C (35x) and a final extension of 4 min at 72 °C. Quantitative PCR was carried out using a iQ™ SYBR® Green Supermix kit (BioRad, USA) with an iCycler iQ Real Time PCR Detection System (Bio-Rad, USA). First, fusion curve and primers efficiency were tested by using sequentially diluted cDNAs (1, 1:10, 1:100, 1:1000) from the four tissues. Consequently, each real-time PCR reaction was performed in triplicate with non-diluted cDNA (1 µL) and the mean of three independent biological replicates was calculated. All results were normalized to the β-actin mRNA levels and calculated using the $2^{-ΔΔCt}$ method [97].
2.3.7 \textit{In silico} analysis of the deduced amino acid sequence

Homology searches with Upsalin nucleotide sequence were performed with the program tblastx available from NCBI, against the nucleotide databases \textit{nr} and \textit{est}. The default algorithm parameters were used with exception of the e-value, which was set to $10^{-3}$. The predicted amino acid sequence was also used as query against UniprotKB using the blastp program. The putative presence of a signal peptide was determined using SignalIP 3.0 (http://www.cbs.dtu.dk/services/SignalP/) [98]. The theoretical molecular weight (Mw), isoelectric point (pI) and amino acid composition were determined with the tools available from the Expasy website (http://web.expasy.org/protparam/) [99]. Putative phosphorylation and glycosylation sites were checked with the NetPhos 2.0 server (http://www.cbs.dtu.dk/services/NetPhos/) [100] and the YinOYang 1.2 server (http://www.cbs.dtu.dk/services/YinOYang/) [101,102], respectively. In order to identify potential protein domains and functional sites, protein sequences were queried against the InterPro database (http://www.ebi.ac.uk/Tools/pfa/iprscan/) [103] and the SMART tool (http://smart.embl-heidelberg.de/) [104]. Sequence alignments were performed with EMBOSS Needle (http://www.ebi.ac.uk/Tools/pca/emboss_needle/) [105], T-Coffee (http://www.ebi.ac.uk/Tools/msa/tcoffee/) [106] and EsPript 2.2 (http://espript.ibcp.fr/ESPript/ESPript/) [107].

2.3.8 Extraction of the shell organic matrix

Shells were rinsed in tap water and brushed, then immersed 24 h in a solution of sodium hypochlorite (dilution 1:20, 0.13% active chlorine) to remove the organic
contaminants. They were subsequently rinsed in milli-Q water and dried. The external prismatic layer was removed by mechanical polishing. The internal nacreous layer was crushed and sieved (< 200 µm) to obtain a fine calcium carbonate powder, which was decalcified overnight with cold dilute acetic acid (5 % v/v), according to the standard procedure [81]. The resulting solution was centrifuged (30 min/ 3900 G / 4 °C) to separate the acetic acid-soluble fraction (ASM = supernatant) from the acetic acid-insoluble one (AIM = pellet). The ASM was filtered on cellulose membrane (5 µm), then ultrafiltered on a 10 kDa-cutoff membrane (Amicon) to concentrate the macromolecules, which were subsequently dialysed at 4 °C for few days against milli-Q water (with several water changes). The solution was finally lyophilized. The AIM was resuspended in milli-Q water, the suspension, centrifuged, and the supernatant put aside. After several suspension/centrifugation cycles, the final pellet was freeze-dried, while the pooled supernatants were added to the ASM. Both ASM and AIM were quantified, by direct weighing of the lyophilisates on a precision balance.

2.3.9 Protein purification and characterization on mono-dimensional gel and on Western blots

The ASM was tested by conventional mono-dimensional SDS-PAGE (Bio-Rad, mini-Protean III). After a complete denaturation with Laemmli buffer [108], gels were stained with silver nitrate [109]. Alternately, gels were electro-transferred on PVDF membranes, and the membranes were subsequently exposed to a specific antibody elicted against the target protein, according to a standard ‘Western blot’ procedure [110]. For detecting the antigen-antibody complex, we used a secondary
antibody coupled to alkaline phosphatase (Sigma A3687, 1:30000), and the complex was stained with NBT/BCIP (Sigma B5655). The ASM was used to purify the protein of interest, according to the procedure described in [81]. In brief, this implies a blind fractionation of the ASM on a preparative SDS-PAGE (Bio Rad, Prep Cell model 491) equipped downstream with a fraction collector, followed by a test of the eighty collected fractions on dot-blot, with an antibody elicited against a peptide of the protein deduced from the DNA sequence (see below). The purity of the protein fraction was checked on a 16% acrylamide gel, which was stained with silver nitrate.

2.3.10 Proteomic analysis of the purified fraction

A proteomic analysis was performed in order to check the identity of the protein fraction purified by preparative electrophoresis. Briefly, the fraction was enzymatically digested with trypsin [111], then purified on Vivapure C18 micro membrane (Vivascience). The purified samples were analyzed by nano-LC-MS/MS using a nano-Liquid Chromatography system (LC Packings, Dionex) and a nano-ESI-qQ-TOF mass spectrometry system (QSTAR XL, AB Sciex). Mass spectrometry data were acquired using Analyst QS 1.1 software (AB Sciex) operated in IDA mode as previously described [111]. Protein identification was performed with ProteinPilot 3.0 software (Applied Biosystems) using the Paragon database search algorithm with the sequence determined from the DNA sequencing.
Chapter 2 Novel Molluskan Proteins Retrieved from Proteomics: a Case Study with Upsalin

2.3.11 Antibody production and ELISA testing

A peptide, corresponding to the sequence ACPRGSSNRYDDPEGF, was designed, synthesized and coupled with a carrier (KLH) for eliciting polyclonal antibodies in two white rabbits, according to the 28 days Speedy program developed by Eurogentec: immunizations at days 0, 7, 10, and 18, and blood sampling at days 0 (pre-immune serum), 21 (intermediate) and 28 (last bleeding). Polyclonal antibodies were tested by ELISA, according to a standard procedure [112]. This experimental procedure allowed determining the reactivity and the titre of the antibody elicited against its target peptide. In this case, the tested antigen was the nacre ASM of Unio pictorum.

2.3.12 Glycosylation studies

The qualitative and quantitative characterization of post-translational modifications (PTMs), i.e., glycosylation, of the purified protein was investigated by two techniques: FTIR and monosaccharide analysis. FTIR spectra were recorded from dry lyophilized purified protein samples on a Bruker Vector 22 equipped with a Specac Golden Gate™ ATR device (Specac Ltd., Orpington, UK) in the wave number range of 4000–500 cm⁻¹ (ten scans at a spectral resolution of 2 cm⁻¹). The assignment of absorption bands was performed by comparison with previous spectra descriptions available in the bibliography. For quantifying the monosaccharide content of the purified protein, lyophilized samples were hydrolyzed in trifluoroacetic acid (TFA) (100 µL, 2 M) at 105 °C for 4 h. These hydrolytic conditions do not allow the quantification of GalNAc and GlcNac,
which are converted into GalN (galactosamine) and GlcN (glucosamine), respectively. Sialic acids, such as N-acetylneuraminic acid, are destroyed during the hydrolytic procedure. Samples were evaporated to dryness before being dissolved with NaOH (100 µL, 20 mM). The sugar contents of the hydrolysates were determined by high performance anion exchange with pulsed amperometric detection (HPAE-PAD) on a CarboPac PA100 column (Dionex Corp., Sunnyvale, CA, USA). Carbohydrate standard (Sigma, St Louis, MO, USA) was injected at 16, 8 and 4 ppm. Non-hydrolyzed samples were analyzed similarly to detect free monosaccharides that could have contaminated the sample during dialysis. For the reasons indicated above, this technique does not allow quantifying sialic acids.

2.3.13 In vitro interaction of the purified protein with calcium carbonate

CaCO₃ precipitation was performed in vitro by slow diffusion of ammonium carbonate vapor in calcium chloride solution [48]. The test was adapted as follows: solutions of CaCl₂ (200 µL, 10 mM) containing different amounts of the purified protein (0 to 20 µg/mL) were introduced in 16-well culture slides (BD Falcon; Becton Dickinson, Franklin Lakes, NJ, USA). Blank controls were performed without any sample. They were incubated for 48 or 72 h at room temperature or at 4°C in a closed desiccator containing crystals of ammonium bicarbonate. They were dried and directly observed with a tabletop scanning electron microscope Hitachi TM1000.
2.3.14 Immunogold localization of the purified protein on shell fragments

Immunogold labeling was performed on shell fragments as described previously [113], using the antibody raised against the purified protein (diluted 1:5000), and a secondary antibody (goat anti-rabbit, dilution 1:400), coupled with 5 nm gold particles (British Biocell International, Cardiff, UK, ref. EM.GAR5). The size of the gold particles was increased further, by incubating the shell fragments in a silver-enhancing solution (BBI, ref. EKL15) for 15 minutes. The samples were then rinsed, dried and directly observed with a Hitachi TM1000 SEM. Control experiments were performed similarly without the first antibody step (but with the secondary antibody and the silver enhancement incubations) or with the pre-immune serum (with all the incubation steps), or with silver enhancement step alone.

2.4 Results

2.4.1 Characterization of Unio pictorum shell

*Unio pictorum* belongs to the order Unionoida, one of the two orders of the subclass Palaeoheterodonta, in the Bivalvia class. The shell of *U. pictorum* is made of two fully mineralized layers of calcium carbonate (Figure 2.2 B). The outer mineralized layer is composed of prisms that develop perpendicularly to the outer shell surface. The internal layer is based on superimpositions of extremely thin flat tablets, which typifies the nacreous layer (Figure 2.2 C). The tablets are arranged in the “brickwall microstructure” pattern, a feature shared by most bivalvian nacres.
Unlike in pteriomorphid bivalves, both nacre and prismatic layers of *U. pictorum* are fully aragonitic. On the external surface, a leathery olive-greenish thin organic sheet, the periostracum, covers the prismatic layer (Figure 2.1 A).

2.4.2 Identification of a nucleotide sequence coding for a 12 kDa protein

In order to amplify a short sequence from the nacre of *U. pictorum*, a set of 16 degenerate primers was designed based on 8 peptide sequences corresponding to the 12 and 16 kDa protein bands previously analyzed by mass spectrometry [96]. One pair of primers corresponding to the peptides (ECSDPEV-fwd and EFYGVCCPLR-rev) gave a specific amplification product of 250 base pairs (bp) (Figure 2.3).
Chapter 2 Novel Molluskan Proteins Retrieved from Proteomics: a Case Study with Upsalin

Since the translation product of the sequenced fragment was consistent with 4 of the peptides identified for the 12 kDa protein band, the full-length sequence coding for this protein was further investigated. To this end, specific primers were designed from the amplified fragment, to obtain the 5’ and 3’-ends by RACE PCR. The overlapping of the two amplified sides corresponded to a transcript sequence of 524 bp, which was confirmed by amplifying the full transcript with two primers situated at the 5’ and 3’-ends. This sequence comprises a start codon, a coding sequence of 375 bp, a TAA stop codon, the polyadenylation signal AATAAA and the poly-A tail, 127 bp downstream the stop codon (Figure 2.4 A).

Figure 2.3: Amplification of a DNA fragment encoding for a shell protein. (A) Electrophoresis in agarose gel (1.5 %). T: DNA markers, P: 250 bp Upsalin amplified fragment, +C: positive control [amplified cDNA of U. pictorum β-actin (450 bp)], -C: negative control. (B) Predicted amino acid sequence for the 250 bp nucleotide sequence. Amino acids in bold were used as references to design the two degenerate primers: forward (DP fwd) and reverse (DP rev). Amino acids inside boxes correspond to the peptides previously identified from MS/MS spectra [26].
Section 2.4 Results

Figure 2.4: Nucleotide and predicted amino acid sequences of Upsalin precursor. (A) Nucleotide sequence with the forward and reverse primers in bold and the polyadenylation signal boxed. (B) Translated coding sequence region of Upsalin. The putative peptide signal is underlined and PTMs are highlighted with (*) for phosphorylations and (G) for glycosylations.

2.4.3 Primary structure and molecular features of Upsalin

The open reading frame encodes a protein of 125 residues that was named Upsalin (Figure 2.4 B). Upsalin has a predicted signal peptide with a cleavage site between the positions 16 and 17 and a putative transmembrane region (pos. 5 to 23). The mature form of the protein comprises 109 aa and exhibits a theoretical molecular weight of 12.3 kDa. This value is in conformity with the 12 kDa band analyzed by MS-MS spectra analysis and from which the 16 degenerate primers were designed [96]. Five amino acid residues, respectively Gly, Pro, Arg, Cys and Tyr, constitute 59% of Upsalin sequence, with an individual percentage comprised between 10
and 13% (Table 2.1). The frequency of acidic amino acids (Asp, Glu) is low, corresponding in sum to less than 8% (Table 2.1). Consequently, Upsalin has a theoretical pI of 9.07. A computer search for putative post-translational modifications (PTMs) was performed, based on the NetPhos and YinOYang algorithms. Six putative phosphorylation sites were found, among which four on Ser (positions 32, 74, 75, 112), one on Thr (position 109) and one on Tyr (position 78) (Figure 2.4 B). Two putative glycosylation sites, involving Ser residues were also detected at positions 33 and 74 (Figure 2.4 B). All these potential PTMs are located on the five peptides that were previously identified by tandem mass-spectrometry, but which did not exhibit PTMs. This finding implies two things: first, at least one Upsalin isoform does not bear PTMs in the nacre extract that was used for acquiring mass-spectrometry data; secondly, one cannot exclude that other PTMs-bearing Upsalin-peptides exist as well, but that they were not identified by our former proteomic analysis.

Table 2.1: **Computed parameters for the mature Upsalin protein sequence** (after the removal of the signal peptide): amino acid composition (%), isoelectric point (pI) and molecular weight (Mw).

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>%</th>
<th>pI</th>
<th>Mw (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>1.8</td>
<td>9.07</td>
<td>12.3</td>
</tr>
<tr>
<td>Arg</td>
<td>11.9</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>11.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>12.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>1.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>4.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>12.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>3.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>10.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>7.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4.4 Tissue specific gene expression of Upsalin

In order to confirm Upsalin full sequence predicted by the RACE approach and to analyze its expression in other tissues, two specific primers from the 3’ and 5’-ends were synthesized and tested on cDNA produced from the mantle, gills, foot and adductor muscles. Figure 2.5 A shows the results from one of the two specimens tested, after RNA extraction and DNase treatment in reproducible conditions. After PCR reaction, all four products were purified, cloned and sequenced. In the four amplifications, the nucleotide sequence was coincident with a size of 400 bp, confirming the presence of Upsalin in the four tested tissues of the freshwater mussel. However, much stronger bands appeared for the mantle and the adductor muscles while the corresponding bands from gills and foot were fainter, suggesting lower expression levels of Upsalin in these tissues. A quantitative analysis of the transcription levels was carried out on three specimens using the same set of primers, as shown on Figure 2.5 B. Upsalin expression ratios were normalized with β-actin and the expression in the mantle was used as the control condition by having the relative expression of 1. The qPCR results confirmed what was observed on gel with standard PCR: transcript levels were the highest in the mantle of the three tested samples, slightly lower in the adductor muscles but negligible in the gills and foot (Figure 2.5 B). These results suggest that Upsalin is not mantle-specific but might be also involved in other functions outside shell formation. However, one should keep in mind that adductor muscles are firmly attached to the shell internal side, which in other words, means that they are located closely to calcification sites.
Figure 2.5: Analysis of the expression of upsalin transcript in four different tissues: mantle, gills, adductor muscles, and foot. (A) Electrophoresis (agarose, 1.5%) of the PCR products resulting from a standard PCR carried out with specific primers designed from the 5' and 3'-ends. The positive control was performed by amplification of a housekeeping gene and the negative control was prepared without addition of cDNA. (B) Histogram of the real-time PCR. Each column represents the relative expression ratio of upsalin in one specific tissue by a mean expression ratio (n=6). The transcription of upsalin in the mantle is used as reference.

2.4.5 Homology search

In order to check whether the sequence of Upsalin was unique or whether it could be affiliated to other proteins or protein families, homology searches were performed with Upsalin nucleotide sequence, using the tblastx program from NCBI. The most significant result (e-value of 2E-40) was obtained against the ‘est’ database of all organisms with a sequence encoding a putative protein of the mantle of the Chinese pearl mussel *Hyriopsis cumingii* (GenBank: 312832200). Figure 2.6 shows the optimal alignment for the two sequences with 94 identities (in black). Interestingly, both *U. pictorum* and *H. cumingii* belong to the same small family of freshwater mussels Unionidae, within the Palaeoheterodonta subclass. At a lower
Section 2.4 Results

taxonomic level, *U. pictorum* and *H. cumingii* are classified into different subfamilies, Unioninae and Ambleminae, respectively. This clearly suggests that Upsalin or related proteins may be present in the mantle tissues of other members of the Unionidae family, in particular genera that are closely related to *Unio*.

Figure 2.6: Pairwise alignment of the predicted amino acid sequences of Upsalin from *Unio pictorum* and the corresponding orthologous from *Hyriopsis cumingii* (produced by EMBOSS Needle): "|"- identical residues, ":"- conserved substitution, and "·"- semi-conserved substitution.

Blast searches were also performed with the predicted amino acid sequence of Upsalin against the UniprotKB. The best matches with an e-value < 1 were obtained with proteins from the fruit flies and the yellow fever mosquito. The former proteins are predicted to be on the extracellular region (GO:0005576) and to have a serine-type endopeptidase inhibitor activity (GO:0004867), including several annotated domains along their sequence.

We decided to select the non-redundant protein segments of the best blastp hits and align them with Upsalin sequence from *U. pictorum* and its homolog from *H. cumingii*. The optimal alignment was obtained with the tool T-Coffee and the visualization was performed with EsPript 2.2. Figure 2.7 shows several similar groups along the alignment that ends at a highly conserved motif of six residues.
(YGVCCP). This motif is located at the C-terminal side of Upsalin, while positioned more upstream for all the insect sequences. All the selected protein segments that gave high scoring pairs with Upsalin constitute part of an EGF-like (Epidermal Growth Factor) domain (IPR003645) and of one cysteine-rich repeat (IPR006150). However, when performing SMART or InterPro searches with Upsalin, no domains within the query were found aside from the signal peptide. The domains mentioned above were still detected but with a score above the required threshold.

Figure 2.7: Multiple sequence alignment of Upsalin with segments of proteins from nine different species of fruit fly and from one mosquito. UniprotKB accession numbers are listed on the left and species name on the right. A: strict identity, A: similarity, A: similarity across groups.

2.4.6 Purification and characterization of Upsalin by SDS-PAGE from shell extracts

Although peptide fragments of Upsalin were obtained by proteomics in a previous study [96], a thorough demonstration of the effective presence of Upsalin
in the shell of *U. pictorum* was required. To this end, we purified directly the protein from *U. pictorum* shell extracts. We first extracted the acetic acid-soluble matrix (ASM) of the nacre of *U. pictorum*. A quantification of both ASM and AIM (acetic acid-insoluble matrix) indicated that ASM represents about 0.04% of the weight of the nacre shell powder, while AIM, 0.5%. In subsequent analyses, ASM only was used, and tested on 1D gels, as shown in Figure 2.8 A (lane 2). ASM is characterized by the abundance of polydisperse and discrete macromolecules. A polyclonal antibody was prepared from a peptide of Upsalin, and, when tested against the nacre ASM on ELISA (Figure 2.8 E), showed a strong reactivity, suggesting the presence of the protein in the extract. In subsequent experiments using this antibody, the second bleed (and not the last one) was used, because of its highest reactivity. ASM was fractionated by preparative SDS-PAGE coupled to a fraction collector, and the Upsalin fraction was detected by dot-blotting each fraction and incubating the membrane with the antibody, as shown on Figure 2.8 B. A pure fraction, eluted between tubes 17 and 21, was obtained, tested on gel (Figure 2.8 A, lane 3) and on Western blot (Figure 2.8 D, lane 3), together with ASM (Figure 2.8 A, lane 2 and Figure 2.8 D, lane 2). The purified fraction exhibits an apparent molecular weight of 12 kDa, which is consistent with the calculated molecular weight of the mature Upsalin. Furthermore, the Western blot results unambiguously show that the antibody is specific of the purified fraction, and does not cross-react with other macromolecules of the ASM (Figure 2.8 D, lane 2). In parallel, the purified fraction of Figure 2.8 A (lane 3) was analyzed by nano-LC-MS-MS, after trypsic digestion. Three peptides of 31, 25 and 15 aa, covering 65% of the sequence of mature Upsalin, were obtained (Figure 2.8 C). These peptides, flanked by the basic residues R and K upstream and R downstream, completely match with the sequence of Upsalin.
This result proved that the purified fraction is Upsalin, which, in other words, means that Upsalin is a protein from the ASM fraction associated to the shell of *U. pictorum*.
Section 2.4 Results

(Figure 2.8 cont.) The sequences of the three peptides identified by nano-LC-MS/MS of the purified protein fraction are indicated on the right. This analysis confirms that the sequence obtained by PCR and 3’/5’-RACE corresponds to the purified protein fraction (lane 3). (C) Western blot of the ASM (lane 2) and of the purified Upsalin (lane 3); the signal is specific of Upsalin (lane 1: molecular weight markers). (D) ELISA of the polyclonal anti-Upsalin peptide antibody against the nacre ASM of U. pictorum. PPI: preimmune serum, SB: second bleed, LB: last bleed. SB was used in the subsequent experiments.

2.4.7 Glycosylation of Upsalin

We investigated the glycosylation patterns of the purified Upsalin fraction. FTIR (ATR) spectra were acquired from Upsalin lyophilisates, as shown on Figure 2.9 A. The characteristic protein peaks are observed at 1653 cm$^{-1}$ (Amide I), at 1522 and 1472 cm$^{-1}$ (Amide II), and at 1212 cm$^{-1}$ (Amide III) [114]. In the range of 4000-2500 cm$^{-1}$, the broad band absorption at 3271 cm$^{-1}$ and the sharp bands located at 2956, 2918, and 2850 cm$^{-1}$ are assigned to the presence of N-H and aliphatic C-H groups, respectively [115]. Characteristic peaks of sugar moieties are observed in the range of 1200-900 cm$^{-1}$, at 1105, 1060, 1001 and 965 cm$^{-1}$, and can be attributed in particular to C-C-O and C-O-C stretchings [78,116]. Furthermore, the weak absorption at 1243 cm$^{-1}$ may reflect the presence of sulphated groups [52].

The monosaccharide composition of Upsalin was determined after hydrolysis and is shown on Figure 2.9 B. On a list of eleven classical monosaccharides, including acidic (GlcA, GalA), neutral (Fuc, Rha, Ara, Xyl, Gal, Man, Glc) and amino sugars (GlcN, GalN), only seven were detected, representing a total of 5 ng per µg of Upsalin lyophilisate. This indicates that Upsalin is only weakly glycosylated (0.5 wt-% glycosylation). The two amino sugars, GlcN and GalN, represent together more than 60% of the composition, and Glc, Gal and Man, 14, 9 and 8% respectively. Interestingly, the acetylated forms of GlcN and GalN (which convert to GlcN and GalN upon hydrolysis) are mostly the monomers, which link the
saccharidic moieties to Ser or Thr (O-linked sugars), or to Asn (N-linked sugars). No trace of Fuc, Rha, and of the acidic sugars GalA and GlcA were detected. We cannot exclude the possibility that the monosaccharide analysis after hydrolysis underestimates the amount of covalently linked sugars, if these sugars are mainly of the sialic acids-type. If so, the FTIR spectrum does not bring a conclusive evidence on the presence of sialic acids since they are characterized by a FTIR-peak at 1608 cm$^{-1}$ [117], which, in the present case, may be completely masked by the Amide I peak at 1653 cm$^{-1}$.

Figure 2.9: Qualitative and quantitative analysis of saccharidic moieties of purified Upsalin. (A) FTIR (ATR) spectrum of Upsalin showing in particular the amide I and II vibration modes of the protein, located at 1653 and 1522 cm$^{-1}$, respectively, and, in the 1200–1000 cm$^{-1}$ range, the characteristic absorptions of sugar moieties (C–C–O and C–O–C stretchings) resulting from glycosylation. (B) Monosaccharide composition of purified Upsalin, as determined by HPAE-PAD chromatography.

2.4.8 In vitro crystallization assay with purified Upsalin

The purified Upsalin was assayed to check its ability to interact with the formation of calcium carbonate in vitro, owing to the diffusion method. The results are shown on Figure 2.10. Increasing concentrations of Upsalin were tested, from 0 µg/mL (control experiments) to 20 µg/mL.
Section 2.4 Results

For the blank experiment, typical rhomboedrons of calcite were produced. Upsalin, even at high concentrations, did not modify the shape of the crystals. In particular, we did not observe polycrystalline aggregates, as often observed when intermediate concentrations are tested ($5 - 10 \mu g/mL$). We did not record inhibition of the crystal formation, even at higher concentrations.

It seems that Upsalin has no interference effect on the precipitation of calcium carbonate. By opposition, the ASM of *U. pictorum* exhibits a dose-dependent effect, as it has been shown in earlier publication [78].

Figure 2.10: *In vitro* crystallization of calcium carbonate in the presence of Upsalin. (A) Blank experiment (no Upsalin). (B) [Upsalin]=5 $\mu g/mL$. (C) [Upsalin]=10 $\mu g/mL$. (D) [Upsalin]=20 $\mu g/mL$. Almost no effect is recorded. (E–G) Effect of the ASM at similar concentrations [7]. (E) [ASM]=1 $\mu g/mL$. (F) [ASM]=5 $\mu g/mL$. (G) [ASM]=20 $\mu g/mL$. 
2.4.9 *In situ* localization of Upsalin in the shell

The localization of Upsalin directly in the shell was performed, owing to the antibody raised against the Upsalin peptide. Six experiments were conducted, and the results shown in Figure 2.11 are the best obtained from the six experiments. All the experiments were performed on cross-sections of shell fresh fractures, in order to see the distribution of the signal in the shell thickness. We observed that Upsalin is localized in the nacreous shell layer. The signal seems to be concentrated in the interlamellar interface, but it is mostly discontinuous. In addition, a clear signal was also obtained in the outer prismatic layer of the shell. In this case, the signal was localized along the periprismatic sheaths that surround individual prisms. These results suggest that Upsalin is localized in both mineralized layers of the shell of *U. pictorum*.

Figure 2.11: Immunogold staining of *U. pictorum* shell fragments with the polyclonal anti-Upsalin peptide antibody. (A) Negative control. Scale bar: 50 µm. (B) Results for the prismatic layer. Scale bar: 15 µm. (C) Results for the nacreous layer. Upsalin is present in both layers.
2.5 Discussion

In this manuscript we describe the identification of a new protein, associated to the shell biomineralization process of the freshwater mussel *Unio pictorum*. We characterized Upsalin both at the transcriptional and protein levels. To this end, we used the data set of peptides previously published [96] on the same species, to design degenerate primers that were subsequently tested on mantle cDNAs. The full sequence of the transcript was obtained by 5’ and 3’ RACE. This sequence encodes a 125 residue-long basic soluble protein, Upsalin, which possesses a signal-peptide. In addition of being secreted, Upsalin is incorporated to the shell: Upsalin was indeed purified from the ASM of the nacreous layer, and the sequence of the purified protein, confirmed by proteomics. Upsalin is weakly glycosylated. An antibody elicited against an immunogenic peptide of Upsalin showed that it is present in both layers (nacreous and prismatic) of the shell of *U. pictorum*. Thus Upsalin, together with Prismalin-14 [84], Mucoperlin [51], N14/N16/Pearlin [118–120] and BMSP represents one of the few examples of a shell protein, which is characterized both at the transcriptional and protein levels, and for which *in vitro* characterization was performed after purification from the shell.

As opposed to other biomineralization transcripts that are exclusively expressed in the mantle [13, 36], the expression of Upsalin is particularly high in the mantle and adductor muscles and it is also faintly detected in the gills and foot. A similar situation was observed before with some members of the shematrin family and with the methionine-rich protein MRNP-34 [95], which are thought to provide a framework for calcification. Shematrin-1, -2, -5 and -6 showed expression levels not restricted to the mantle but also present in other tissues such as the adductor muscles [94].
What could be the function of Upsalin in biomineralization? Firstly, a blast search at the nucleotide level indicates that the highest similarity found so far is with a hypothetical mantle protein of the Chinese freshwater pearl mussel *Hyriopsis cumingii*, the function of which is totally unknown. Upsalin and this protein may be considered as true homologs. This suggests that Upsalin or Upsalin-like proteins may also be found in other genera of the Unionidae, but further experimental RT-PCR investigations are required, in different representatives of this family, and, beyond, of the whole Palaeoheterodonta subclass. At present day, because of the scarcity of molecular data in NCBI or SwissProt databases for this bivalve group, *in silico* analyses for retrieving Upsalin homologs are irrelevant. Additionally, we were not able to detect any Upsalin homolog even in the large 270,000 EST database from *Pinctada margaritifera* mantle cDNA library [88], suggesting that Upsalin is not present in some other mollusks with a nacreous shell, and could be taxon specific. In addition, Upsalin shows similarity with a group of proteins of insect origin (Diptera). The similarity region concerns about 80 residues located in the first half of the sequence of extracellular proteins with serine-type endopeptidase inhibitor activity (GO: 0004867). The best hits include proteins from nine species of *Drosophila*. The alignment of Upsalin and of its *Hyriopsis* homolog with these proteins revealed a high degree of sequence conservation with 10 conserved cysteine residues and 11 other residues, 8 of which are hydrophobic (G, L, V, P, Y). Moreover a short fully conserved motif YGVCCP is present at the C-terminal side of the alignment. It is known that cysteine residues play a key-role on the 3D-structure of a protein, as they form intra- and intermolecular disulfide bridges, constraining the protein to proper folding and stabilization in the extracellular domain. In the dipterian proteins, the aligned regions are part of one EGF-like domain and of one cysteine-rich-like domain. According to InterPro resources on
the domain organization of the cysteine-rich repeat IPR006150, it is commonly associated with Kunitz domains that may function as serine peptidase inhibitors. This fact is coherent with our former proteomic analysis on the ASM of *U. pictorum* and of its 12/16 kDa SDS-PAGE bands, which also showed peptides with similar putative serine protease inhibitor domains [96].

Finally, the amino-acid sequence of Upsalin does not exhibit any homology with previously discovered shell-matrix proteins, in particular with proteins associated to nacre microstructures, whatever their origin: abalone [7], pearl oysters [17, 20] or blue mussel [90]. This finding is striking, not to say remarkable, and it points out once more the unexpected diversity of molluskan shell proteins, a fact among others that suggests that mollusks have ‘invented’ more than one single ‘biochemical’ pathway for synthesizing nacre [19, 37].

Unlike many other molluskan shell-associated proteins [13, 38] no internal repeats or low complexity regions are observed in Upsalin primary sequence. Furthermore, Upsalin is only weakly glycosylated, and it is very unlikely that its sugar moieties exert a role in concentrating calcium ions at the vicinity of the nucleation sites [123] as it has been proposed for other proteins [11, 32]. Previous calcium-binding studies on the ASM of *Unio pictorum* by staining with Stains All and by autoradiography with $^{45}$Ca suggest weak calcium binding ability at molecular weights below 14 kDa [78]. In conformity, Upsalin does not interfere *in vitro* with the precipitation of calcium carbonate. These considerations suggest that Upsalin does not display a direct interaction with calcium or calcium carbonate crystals. Furthermore, Upsalin does not exhibit any chitin-binding capacity (data not shown). Taken together, these data suggest that Upsalin is involved in other functions, which are indirectly related to biomineralization. Protecting the matrix against enzymatic degradation, organizing the spatial arrangement of other
matrix macromolecules, or signaling activity towards the epithelial cells of the calcifying mantle represent putative roles, among others, but these hypotheses need to be tested experimentally. Also the fact that Upsalin has a considerable amount of prolines (12.8%) in its sequence, confer a solid backbone with good potential for binding other proteins [124].

Eight years ago, Gotliv and co-workers [49] published a paper that pointed out the fact that several proteins retrieved from shell tissues did not have a clear function in mineralization. Since that publication, although our knowledge about shell proteins has immensely progressed, we are compelled to note that this observation is still valid. Upsalin, like several other proteins, enters the growing collection of orphan proteins with no association to known functional traits or protein families. This clearly indicates that the next coming years should be urgently dedicated to set up approaches such as gene knock-down, as it has recently been performed on the pearl oyster *P. fucata* [125] or the use of two-hybrid systems, that will be extremely useful for deciphering the functions of molluskan shell proteins in biomineral deposition.