Molecular Cloning and Characterization of Lustrin A, a Matrix Protein from Shell and Pearl Nacre of *Haliotis rufescens**

Xueyu Shen‡§¶, Angela M. Belcher§¶‖**, Paul K. Hansma¶‡‡, Galen D. Stucky¶‖§§, and Daniel E. Morse‡§¶ ¶¶

From the ‡Molecular, Cellular and Developmental Biology Department, ‡‡Physics Department, ‖Chemistry Department, §§Materials Department, §Marine Biotechnology Center, and the ¶Materials Research Laboratory, University of California, Santa Barbara, California 93106

A specialized extracellular matrix of proteins and polysaccharides controls the morphology and packing of calcium carbonate crystals and becomes occluded within the mineralized composite during formation of the molluscan shell and pearl. We have cloned and characterized the cDNA coding for Lustrin A, a newly described matrix protein from the nacreous layer of the shell and pearl produced by the abalone, *Haliotis rufescens*, a marine gastropod mollusc. The full-length cDNA is 4,439 base pairs (bp) long and contains an open reading frame coding for 1,428 amino acids. The deduced amino acid sequence reveals a highly modular structure with a high proportion of Ser (16%), Pro (14%), Gly (13%), and Cys (9%). The protein contains ten highly conserved cysteine-rich domains interspersed by eight proline-rich domains; a glycine- and serine-rich domain lies between the two cysteine-rich domains nearest the C terminus, and these are followed by a basic domain and a C-terminal domain that is highly similar to known protease inhibitors. The glycine- and serine-rich domain and at least one of the proline-rich domains show sequence similarity to proteins of two extracellular matrix superfamilies (one of which also is involved in the mineralized matrixes of bone, dentin, and avian eggshell). The arrangement of alternating cysteine-rich domains and proline-rich domains is strikingly similar to that found in frustulins, the proteins that are integral to the silicified cell wall of diatoms. Its modular structure suggests that Lustrin A is a multifunctional protein, whereas the occurrence of related sequences suggest it is a member of a multiprotein family.

The molluscan shell and pearl are mineralized structured

composites of $CaCO_3$ crystals and organic polymers exhibiting exceptional nanoscale regularity and strength (1–14). Nacre, the lustrous material of pearl and the inner "mother of pearl" layers of many shells, exhibits a fracture toughness ~3,000 times greater than that of the mineral alone (15, 16). Although the organic components typically constitute only ~1% by weight of the biomineralized composite material (17), they are responsible for its organization and the resulting enhancement of fracture toughness (3–7, 18–21). Proteins represent the majority of extracellular organic polymers controlling biomineralization of the shell (8–10); they comprise at least four functional classes, including (i) a nucleating sheet that participates in control of nucleation of the first layer of oriented calcite in deposition of the abalone shell and flat pearl (13, 14, 22), (ii) a family of polyanionic proteins that can be extracted by demineralization of the shell (8–10, 17, 23–25) and have been shown *in vitro* to control the polymorph and atomic lattice orientation by cooperative interaction with the growing crystals (22, 26, 27), (iii) proteins of an insoluble, highly cross-linked matrix, forming an organizing network of interconnected compartments and fenestrated sheets that control the morphology and higher order packing of the $CaCO_3$ crystals (8, 22, 28–32), and (iv) functional enzymes, such as carbonic anhydrase (33), that apparently contribute to the control of biomineralization.

Nacre consists of layers of interlocking thin tablets of aragonite (one of the polymorphs of $CaCO_3$) typically 400 nm thick and 5–10 μm wide surrounded on all sides by thin sheets of organic matrix approximately 30 nm thick (7, 29, 30, 34). The thickness of the mineral tablets apparently is determined by the matrix sheets (22), while their interdigitation is controlled, in part, by the stochastic location of nanopores in the stencil-like sheets through which the crystals grow from one layer to the next (32, 35–38). Electron microscopy, x-ray, and electron diffraction have shown that each sheet consists of a chitin-like core sandwiched between two layers of protein (12, 28–30). Diffraction data suggest that the protein sheets exhibit a β conformation similar to that of silk fibroin (28), a finding supported by the fact that glycine and alanine are the most abundant amino acids (8, 9, 29, 30). A glycine- and alanine-rich matrix protein recently was cloned from the nacre-producing tissue of the bivalve, *Pinctada fucata* (39), and a matrix protein rich in glycine, lysine, and valine has been extracted and characterized from the abalone, *Haliotis rufescens* (40). To understand the molecular mechanisms governing self-assembly of the matrix and biomineralization of nacre, one of us (A. M. B) has developed a procedure for extraction of other classes of matrix proteins from the nacre of *H. rufescens* shell and flat pearl with urea and SDS (31). Here we report the complete amino acid sequence of one of these proteins which we have named Lustrin A. The results reveal a new class of shell matrix

protein, with a repeating modular structure containing sequences similar to those of two extracellular matrix protein superfamilies known from other animal tissues.

MATERIALS AND METHODS

RNA Purification—Rapidly growing specimens (75–100 mm) of red abalone (*H. rufescens*), a gastropod mollusc, were obtained from The Cultured Abalone Inc. (Goleta, CA). Total RNA was isolated from the mantle, muscle, gill, and stomach tissues using TriZOL reagent (Life Technologies, Inc.) and the modified TRI Reagent procedure for RNA isolation (41). The mantle pallial cell layer (responsible for secretion of the organic and inorganic precursors of the shell) was extracted by vigorous shaking with TriZOL reagent; the underlying muscle was discarded. Other tissues were homogenized in TriZOL reagent (stomach contents were removed before the tissue was homogenized). Poly(A)$^+$ RNA from mantle tissue was isolated using the Poly(A)Tract mRNA isolation system IV (Promega Corp.) with the following modifications: annealed oligo(dT)-mRNA hybrids were washed with $0.2 \times$ SSC (0.03 M sodium chloride, 0.003 M sodium citrate) instead of $0.1 \times$ SSC; the elution of mRNA was performed at 40 °C instead of at room temperature.

RT-PCR[1] Amplification—Degenerate oligonucleotide primers for PCR amplification were synthesized to correspond to portions of the N-terminal sequence and an internal sequence determined for one of the purified matrix proteins (Table I) (31). Degenerate primer D1 (GARCCNGGNYTRAAYGT) encoding EPGLNV, a part of the N-terminal peptide, and D2 (CARCANACNCCNGGYTT) corresponding to the antisense strand of the sequence encoding KPGVCC, a part of the internal peptide, were obtained from Cruachem Inc. Mantle mRNA (200 ng) was used for RT primed with a poly(dT)$_{12–18}$ primer; the reaction was catalyzed by Superscript II reverse transcriptase (Life Technologies, Inc.) in a 20-μl reaction. The resulting cDNA then was diluted to 200 μl with H$_2$O, and 2 μl were used in the PCR reaction which also included 2 μM of each primer (D1 and D2), $1 \times$ AmpliTaq DNA polymerase buffer (Perkin Elmer), 2 mM MgCl$_2$, 200 μM dNTP, 5% (v/v) formamide, and 1 unit of AmpliTaq DNA polymerase (Perkin Elmer). PCR amplification (performed in a Cetus DNA Thermal Cycler from Perkin Elmer) conditions included an initial step at 94 °C for 5 min followed by 40 cycles at 94 °C for 1 min, 47 °C for 1.5 min, and 72 °C for 1.5 min. A final extension step of 72 °C for 7 min was performed after the cycles. PCR reactions with only one degenerate primer (D1 or D2) were incubated in parallel as negative controls.

Construction and Screening of Mantle cDNA Library—Approximately 5 μg of mantle poly(A)$^+$ RNA, isolated from four animals, was used to construct the cDNA using the Stratagene ZAP Express cDNA synthesis protocol. Small cDNA molecules were removed by gel filtration over Sephacryl S-400. The cDNA then was inserted into the lambda ZAP Express vector and packaged with the Gigapack II Gold extract from Stratagene. The primary library contained 1.2×10^6 plaque-forming units (PFU) and was subsequently amplified once to yield a titer of 2×10^6 PFU/μl with a frequency of nonrecombinant lambda vectors below 0.1%.

The library was first screened by PCR amplification. Two gene-specific primers, GTCGTTGTGGAGTGCGT (G1) corresponding to nucleotides (nt) 65–81 of the 242-bp RT-PCR product (Fig. 1), and ACCTCGAACACACACCCAG (G2) matching the antisense sequence of nt 200–216 of the same RT-PCR product, were obtained from Cruachem Inc. PCR screening reactions were carried out under the same conditions as described above with the following changes: no formamide was added to the reactions, the denaturation step was shortened to 30 s, annealing was done at 62 °C for 1 min, and extension was performed for 45 s. The appearance of a 142-bp PCR product indicated the existence of a positive clone. Initially, approximately 5×10^5 PFU from the mantle cDNA library were plated onto ten 150-mm NZY plates at a density of 50,000 PFU per plate. Phage particles from each plate were eluted into 8 ml of SM buffer (0.1 M NaCl, 10 mM MgSO$_4$, 50 mM Tris-HCl, pH 7.5, and 0.01% w/v gelatin), and 1 μl of each separate phage suspension was used in the PCR reaction. Ten PCR reactions were performed (one for each of the ten plates); nine generated the 142-bp product. Phage suspension from one of the nine positive plates was chosen randomly and diluted for secondary screening; the phage were plated onto twenty 100-mm NZY plates at a density of 2,500 PFU per plate. PCR reactions were performed for each of the 20 plates; four

yielded the 142-bp product. Phage suspension from one of these four positive plates was again chosen randomly; further dilutions were made and PCR screenings were conducted sequentially as above. After four rounds of dilution and PCR screening, a single positive clone (designated clone 1) was isolated (Fig. 1). The 993-bp insert of clone 1 was labeled with digoxigenin-11-dUTP (DIG) using a Genius DNA random labeling kit (Boehringer Mannheim) and used to screen the library again. Hybridization was carried out at 42 °C in 50% formamide, 2% blocking reagent (Boehringer Mannheim), $5 \times$ SSC, 0.1% sarkosyl, and 0.02% SDS. Filters were washed in $0.5 \times$ SSC and 0.1% SDS at 65 °C for 1 h. Probe binding was detected by an anti-DIG Fab conjugated to alkaline phosphatase, and colorimetric detection was done with nitro blue tetrazolium and 5-bromo-4-chloro-3-indolyl phosphate according to the protocol recommended by Boehringer Mannheim.

Isolation of 5′-End of cDNA by PCR—Reverse transcription was carried out as described above using gene-specific primer G2. The sequence of G2 matches that of the antisense strand of nt 274–290 of the longest clone (clone 5) isolated from the library (Fig. 1). A second gene-specific primer TCGAACACACCCAGTGGTTG (G3), corresponding to the antisense sequence of nt 268–287 of clone 5, was used in the PCR reaction. The conditions for tailing and PCR reaction were as described previously (42).

DNA Sequencing and Sequence Analyses—Inserts of the λ clones were excised *in vivo* into the pBK-CMV phagemid vector. PCR products were subcloned into the pBluescript vector. They were sequenced by the dideoxynucleotide chain termination method (43) using Sequenase Version 2.0 (U. S. Biochemical Corp.). Oligonucleotide primers were synthesized (Operon Technologies, Inc.) to permit sequencing of both strands of the entire clone. Compressions resulting from secondary structures in the DNA were eliminated by replacing dGTP with dITP. Sequence data were analyzed with the Wisconsin Sequence Analysis Package by Genetics Computer Group, Inc.

Northern Blot Analysis—Samples of 25 μg of total RNA from various tissues were electrophoresed on a 1.4% agarose formaldehyde gel, transferred to a Hybond-N$^+$ nylon membrane (Amersham), and UV-cross-linked to the membrane. The membrane was hybridized with ^{33}P-labeled cDNA probe at 42 °C in 50% formamide, $5 \times$ SSPE (0.75 M NaCl, 0.05 M NaH$_2$PO$_4$, 0.005 M EDTA), $5 \times$ Denhardt's solution (0.1% Ficoll, 0.1% polyvinylpyrrolidone, and 0.1% bovine serum albumin), 0.5% SDS, 200 μg/ml salmon sperm DNA for 12 h, and washed in $1 \times$ SSC, 0.5% SDS at 55 °C for 30 min before being exposed to Kodak X-Omat film.

RESULTS

Isolation of Lustrin A cDNA Clones—The first Lustrin A cDNA clone was isolated using the RT-PCR strategy. Two degenerate primers, D1 and D2, were designed corresponding to the N terminus and an internal peptide sequence of Lustrin A (see "Materials and Methods"). A prominent 242-bp product was generated from mantle mRNA using these primers (Fig. 1). Nucleotide sequences at the two ends of this fragment matched those of D1 and D2. The deduced amino acid sequences at the N and C termini of the amplified fragment were (EPGLN-V)NCTT and PPA(KPGVCC), respectively. EPGLNV and KPGVCC were used to design D1 and D2 and therefore were present as expected. NCTT immediately following EPGLNV and PPA immediately preceding KPGVCC were perfect matches to the N-terminal and the internal peptide sequences of Lustrin A (Table I) (31), thus confirming the correspondence of the amplified 242-bp fragment with part of the sequence encoding Lustrin A. Two gene-specific primers, G1 and G2, were designed based on the sequence obtained from this fragment (see "Materials and Methods").

To obtain a full-length cDNA clone of Lustrin A, a cDNA library was constructed starting from mRNA extracted from the mantle pallial cells that secrete the shell precursors. This library was screened by PCR using the gene-specific primers G1 and G2 as described under "Materials and Methods." One positive clone (clone 1) was isolated; its insert of 993 bp included the original 242-bp RT-PCR fragment, and the deduced polypeptide sequence contained a perfect match for a second internal peptide sequence of Lustrin A (Fig. 1). Although this result verified that clone 1 encodes Lustrin A, it apparently

---

[1] The abbreviations used are: RT-PCR, reverse transcription-polymerase chain reaction; bp, base pair(s); PFU, plague-forming unit(s); kb, kilobase(s); nt, nucleotide(s); DIG, digoxigenin-11-dUTP.

A

```
RT-PCR
(242bp)          [==]      (107–348)
                            E
clone #1         [/]==[////]|       (33–1025)
(993bp)                     E
clone #5         [/]===[////]|[·················]   (33–4439)
(4407bp)
5' RACE          [/]==       (1–319)
(319bp)
Lustrin A        [/]==[////]| E    B  E  B  E   [·················]  (1–4439)
cDNA                        E     |  |  |  |
(4439bp)
```

B

```
   1  GGTGGCGGATTTTACTCTGACAGTCATGGAGCGTTTTCTCTGGGTCCTGTGTATAGCTGCCGGGTTCTCTGTCAACTATGGGCTGAGAAGGGCGCCTTAT  100
   1                      [M  E  R  F  L  W  W  V  L  C  I  A  A  G  F  S  V  N  Y  G] L  R  R  A  P  Y   25
 101  CCGTGTGAACCGGGGTTAAATGTGAACTGTACAACTGGAGAATGTCGGTTAGTGTTCAGTTGCAGTCTTCGTCGTTGTGGAGTGCGTCCTGAATGTGTAG  200
  26   P  C  E  P  G  L  N  V  N  C  T  T  G  E  C  R  L  V  F  S  C  S  L  R  R  C  G  V  R  P  E  C  V  D   59
 201  ATCGTTCACCGGTACCATCAATTAACTGTACCATTGGAAAACCTACCATTGACAACTCCAGGAAAATCTCTTGTGCGCCAGACGGGTCATGTCCAGC     300
  60   R  S  P  V  P  S  I  N  C  T  I  G  K  P  T  I  D  T  N  L  Q  E  I  S  C  A  P  D  G  S  C  P  A   92
 301  AACCACTGGGTGTGTTCGAGGTCCCCCAGCCAAGCCAGGAGTGTGTTGTTTAATCCTTCTTCAGGACCTCCGGGACCTCCGAGACCTCCGGGACCTCCG  400
  93   T  T  G  C  V  R  G  P  P  A  K  P  G  V  C  C  F  N  P  S  S  G  P  P  G  P  P  R  P  P  G  P  P   125
 401  AGACCTCCAGGACCACCACAACCAGATCCAAATCTTTTAGACCCTTGCTTCCCTGGCAAAAATGTGAACTGCACAAGTGGAGAAGTGCGCTTGATGGCAG  500
 126   R  P  P  G  P  P  Q  P  D  P  N  L  L  D  P  C  F  P  G  K  N  V  N  C  T  S  G  E  C  R  L  M  A  D  159
 501  ATTGCCAACACCAGAGTTGCCCGGCACTTCCGTATTGTGTTGCTCCTTCGCCGAACGTGACAGTCCCATGTCCTATCGGTAAATCTGCTATCGACCGCAA  600
 160   C  Q  H  Q  S  C  P  A  L  P  Y  C  V  A  P  S  P  N  V  T  V  P  C  P  I  G  K  S  A  I  D  R  N   192
 601  CCTCCGCGAGTTCTCGTGTTTACGAAACAGAGACGCTTGCCCACGATCCACTGGCTGTGTGGTCGGCGCTCAGGGATCAGCAGCCGTTTGTTGTTACAGA  700
 193   L  R  E  F  S  C  L  R  N  R  D  A  C  P  R  S  T  G  C  V  V  G  A  Q  G  S  A  A  V  C  C  Y  R   225
 701  CCTCCACTGGTACCAGGACCACCACCGACCGATCCAAATCCTTTAGATCCTTGCTTCCCTGGTAAAAATGTGAACTGCACAGCAGGTGAATGTCGGCTGG  800
 226   P  P  L  V  P  G  P  T  P  T  D  P  N  P  L  D  P  C  F  P  G  K  N  V  N  C  T  A  G  E  C  R  L  V  259
 801  TAGCAGACTGCTCTCGAAAAGGATGTCCTGCCGGCCCCACGTGTGTTGACCCCTCACCCGTACCATCACTGAACTGTGACATTGGCAAGCCTGCTCTGAA  900
 260   A  D  C  S  R  K  G  C  P  A  G  P  T  C  V  D  P  S  P  V  P  S  L  N  C  D  I  G  K  P  A  L  N   292
 901  TTCCTACGGCAATGAAATCTCTTGTGCAGGAGGGGGAGCATGTCCAGTCAACACTGTGTGTGGTGGCACATCCATCAGGGGCACCCGCCGTCTGCTGTTTC 1000
 293   S  Y  G  N  E  I  S  C  A  G  G  G  A  C  P  V  N  T  V  C  V  A  H  P  S  G  A  P  A  V  C  C  F   325
1001  AAACCAGCTGGCCCAACAACACCCCAACCACCAACAATACCCCAACCACCAACAACGCCATCATCCCCTACAGGCGATCCTTGCGAACCAGGGGTGAATG 1100
 326   K  P  A  G  P  T  P  Q  P  P  T  P  Q  P  P  P  T  T  P  S  S  P  T  G  D  P  C  E  P  G  V  N  V   359
1101  TCAACTGCACAGCGGGTACATGCGAGACTGGTGGTGGGGACTGTCGATTCCCTGGGTGCCCCGCTGTTCCTAAGTGTGTCGATCCCTCCTCAAAACCGTCACT 1200
 360    N  C  T  A  G  T  C  R  L  V  V  D  C  R  F  P  G  K  N  V  C  V  D  P  S  S  K  P  S  L   392
1201  GAACTGCTCAATAGGTGACCCCGCTCTGAACCCAAACTTGCAGGAAATCTCCTGTGTGGGGGGAGCTGCATGTCCCCGAAACACAGCGTGTTTGCCGCC 1300
 393    N  C  S  I  G  D  P  A  L  N  P  N  L  Q  E  I  S  C  V  G  G  G  A  A  C  P  R  N  T  A  C  F  A  A  425
1301  CCGTCAGGATCACCAGCAGTCTGTTGCTACACCTCAGGGCCACCAAGACCCGAACCACCAAGCCCATCCCCCCTACAGGCGATCCTTGTGAACCCGGGG 1400
 426   P  S  G  S  P  A  V  C  C  Y  T  S  G  P  P  R  P  E  P  P  S  P  S  P  P  T  G  D  P  C  E  P  G  V  459
1401  TGAATGTCAACTGCACAGCGGGTACATGCAGACTGGTGGAGGACTGTCGAATCCGTGGGTGCCCCGCTGTCCCTAAGTGTATCGACCGAGACCCACCACT 1500
 460   N  V  N  C  T  A  G  T  C  R  L  V  E  D  C  R  I  R  G  C  P  A  V  P  K  C  I  D  R  D  P  P  L   492
1501  CCCGCCTCCACCTGACGTCTGCCCCGTGGGTACACCTGTACTTGGCGCAGACTTAAAGCAGCTGTATTGTGGCCGCGGAGGGGAGAAGATGTCCCTGGAAT 1600
 493   P  P  P  D  V  C  P  V  T  P  V  L  G  A  D  L  K  Q  L  Y  C  G  R  G  G  R  R  C  P  W  N   525
1601  ACTTACTGCGTTATTCATCCCGCGGATAGATATGCTGTTTGCGTGTTTTGGTTCTGGTCCAGCTTCAGCAATCGCACCAACATCAGCACCAGGACCTGTGG 1700
 526   T  Y  C  V  I  H  P  A  D  R  Y  A  V  C  C  F  G  S  G  P  A  S  A  I  A  P  T  S  A  P  G  P  V  D  559
1701  ATCCCTGTGAACCTGGCGTCAATGTGAACTGTACGATAGGCGTTTGTCGACTGGTGGCAAATTGTGATTACTGGCCCTGTCCTGCGCGCCCAACATGTGT 1800
 560   P  C  E  P  G  V  N  V  N  C  T  I  G  V  C  R  L  V  A  N  C  D  Y  W  P  C  P  A  R  P  T  C  V   592
1801  TGATCACTCCCCAGAACCGTCACTGAACTGCACAATAGGTGACCCCGCACTGAACGGGAAAATTAGAGGAATTCTCCTGTGTTGGGGGACGCCTTTGTCCC 1900
 593   D  H  S  P  E  P  S  L  N  C  T  I  G  D  P  A  L  N  G  K  L  E  E  F  S  C  V  G  G  R  L  C  P   625
1901  CTAAACACAGCGTGTCTTGCCGCCCCTTCAGGGTCACCAGCAGTCTGTTGCTACAGGCCACCTGTCGCCATCACCCCAGCTCCAACAACAGTTCCAAATAC 2000
 626   L  N  T  A  C  L  A  A  P  S  G  S  P  A  V  C  C  Y  R  P  P  V  A  I  T  P  A  P  T  T  V  P  I  P  659
2001  CAGTGTCAACAGCAGCACCAACATCAGCACCAGGACCTGTGGATCCCTGTCAACCTGGTGTCAATGTGAACTGTACGAAAGGCGAATGTCGACTGGTGGC 2100
 660   V  S  T  A  A  P  T  S  A  P  G  P  V  D  P  C  G  P  G  V  N  V  N  C  T  K  G  E  C  R  L  V  A   692
2101  CATTTGTAAATACTGGCCGTGTTTTGCGCTCCCAACGTGTGTTGATCCCTCCCCGCCACCGTCAGTGGAGTGTCCAGTGGGAAAACCAGCGTTGGATGAA 2200
 693   I  C  K  Y  W  P  C  F  A  L  P  T  C  V  D  P  S  P  P  P  S  V  E  C  P  V  G  K  P  A  L  D  E   725
2201  AAGCTCGAAGAATTCTCTTGCAAAGATTGCCCCTTCAATACCGTGTGTTATAAAGGCGCTGTCTGTTGTGTCCCATGGTCAGGAAACAGACCATCAGGTC 2300
 726   K  L  E  E  F  S  C  K  D  C  P  F  N  T  V  C  Y  K  G  A  V  C  C  V  P  W  S  G  N  R  P  S  G  P  759
2301  CCGCAGGTCCCGCAGGTCCCGCAGGTCCCGAACGTCCCGCAACCTCCGTCCCACTGGACCCTTGCACGCCTGGTTTGAATGTCAACTGCACAAGCGGTGT 2400
 760   A  G  P  A  G  P  A  G  P  E  R  P  A  T  S  V  P  L  D  P  C  T  P  G  L  N  V  N  C  T  S  G  V   792
2401  CTGTCGACTTGTCGAGGACTGTAGACGACCTGGCTGCCCCGCCGTCCCACATGCATTGACCGAGACCCACCACTTCCACCTCCACCTGACGTCTGTCCT 2500
 793   C  R  L  V  E  D  C  R  R  P  G  C  P  A  V  P  T  C  I  D  R  D  P  L  P  P  P  P  D  V  C  P   825
2501  GTGGGTACACCTGTACTTGGCAGAGACTTAAAGCAGCTGTATTGTGGCCGCGGAGGGAAAAGATGTCCCGGGAATACTTACTGCGTTATTCATCCCGCGG 2600
 826   V  G  T  P  V  L  G  R  D  L  K  Q  L  Y  C  G  R  G  G  K  R  C  P  G  N  T  Y  C  V  I  H  P  A  D  859
2601  ATAGATATGCTGTATGCTGTTTTGGTTCTGGTCCAGGTCCGCAGCCACCTATACCTACCACCCCCACCTCCAACAACGTACCCGTGCACACCTGCCAATAT 2700
 860   R  Y  A  V  C  C  F  G  S  G  P  G  P  Q  P  P  I  P  T  P  P  P  P  T  T  Y  P  C  T  P  A  N  I   892
2701  TAACTGCACAGCCGGCGAATGTCGTCTCGTGGCGTATTGCAACGCGGTGCCTTGTGGGAGGACAACGCCAACATGTGTTGACCCCTCGCCGCCCCAACC 2800
 893    N  C  T  A  G  E  C  R  L  V  A  Y  C  N  A  V  P  C  G  R  T  T  P  T  C  V  D  P  S  P  P  P  T  925
2801  CGAAAATGCCCAGTGGGTAAGCCTGTTCTCACCCCTAGGCTGACGGAGTTCAGATGTTACCCCAGGGTCCGTCTGTGTCCAGGTGACTCATTCTGTCTCA 2900
 926   R  K  C  P  V  G  K  P  V  L  T  P  R  L  T  E  F  R  C  Y  P  R  V  R  L  C  P  G  D  S  F  C  L  R  959
2901  GGGGTCCAGGTGACGAACCAGGAGTATGTTGTTGGGATAACAGATTACGGCCAACACAAGGTTCAGGCTCAGGCTCAGGCTCAGGCTCAGGCTCAGGTTC 3000
 960   G  P  G  D  E  P  G  V  C  C  W  D  N  R  L  R  P  T  Q  G  S  G  S  G  S  G  S  A  S  G   992
3001  CAGCTCTAGTTCCAGCTCAGGGTCAACTTCGGGTTCTGGCTCAGGGTCTGGCTCAGGATCAAGTTCAGGATCGGGCTCAGGATCAAGTTCAGCGTCTGGC 3100
 993   S  S  S  S  S  G  S  T  S  G  S  G  S  G  S  G  S  G  S  S  S  A  S  G   1025
3101  TCAGGATCAAGTTCAGGGTCTAGCTCAGGATCAAGCTCAGGATCAAGTCAGGATCAAGTTCAGGGTCTGGCTCGGGGTCTGGCTCAGGATCAAGTTCA  3200
1026   S  G  S  S  S  G  S  S  S  G  S  S  S  G  S  S  S  G  S  S  S  G  S  S  S  G  S  S  S  A  1059
3201  CGTCTGGCTCAGGGTCAAGCTCAGGGTCTAGCTCAGGGTCTAGCTCAGGATCAGGCTCAGGATCAAGTTCAGCGTCTGGCTCAGGATCAAGTTCAGGGTC 3300
1060   S  G  S  S  S  G  S  S  S  G  S  S  S  G  S  S  S  A  S  G  S  G  S  S  S  G  S  S  S  G  S   1092
3301  TGGCTCAGGGTCTAGTTCGGGGTCTGGCTCAGGATCTGGCTCAGGATCAAGTTCGGGGTCTGGCTCAGGGTCTAGTTCAGGATCAAGCTCAGTATCAAAC 3400
1093   G  S  G  S  S  S  G  S  G  S  G  S  G  S  S  S  G  S  G  S  G  S  S  S  S  V  S  N   1125
3401  TCATGGACTGGCTCAGGATCAAGTTCAGGGTCTGGCTCAGGATCAAGCTCATGGTCTGGCTCAGGGTCTAGTTCAGGGACTGGCTCGGGATCAAGCTCAT 3500
1126   S  W  T  G  S  S  S  G  S  S  S  G  S  S  S  W  S  G  S  S  S  G  T  G  S  S  S  S  W  1159
3501  GGTTTGGCTCAGGATCAAGTTCAGGGTCTGGCTCAGATTCTAGTTCAGGATCAAGTTCAGCGTCTGGCTCAGGGTCAAGCTCAGGGTCAAGCTCAGGATC 3600
1160   F  G  S  G  S  S  S  G  S  D  S  S  G  S  S  S  A  S  G  S  S  S  G  S  S  S  G  S   1192
3601  AGGCTCAGGATCAAGCTTATGGTTTGGCTCAGGATCAAGTTCAGGGACTGGCTCGGGGTCTAGCTCAGGATCAAGTTCAGGGTCTGGCTCAGGATTCTAGT 3700
1193   G  S  G  S  S  L  W  F  G  S  G  S  S  S  G  T  G  S  G  S  S  S  G  S  G  S  S  S  S  G  T   1225
3701  TCAGGATCAAGCTCAGGGTCTACATCTGGGTCAAGTTCAGGCTCTGGTTCAGCGTCAGGCTCAGGCACAGGCTCAGGAAAGGGCGCAAGTTATGACACTG 3800
1226   S  G  S  S  S  G  S  T  S  G  S  S  S  G  S  A  S  G  S  G  T  G  S  G  K  G  A  S  Y  D  T  D  1259
3801  ATGCAGACTCGGGCTCAGATAACGAGATCCCCTGGCTATCTGCCGCAAGACCCATGCACTCCCGGACTCTACATCAACTGCACAGCGGGCACCTGTCGACT 3900
1260   A  D  S  G  S  D  N  R  S  P  G  Y  L  P  Q  D  P  C  T  P  G  L  Y  I  N  C  T  A  G  T  C  R  L  1292
3901  CACGGCATGGTGCTTATACAACTTCTGCCCGGCAGTGCCGACCTGTGTTGACTCATCTCCTGATGCCAGTGGGGAATGTCCAGTTGGTCTACCGGCTTTG 4000
1293   T  A  W  C  L  Y  N  F  C  P  A  V  P  T  C  V  D  S  S  P  D  A  S  G  E  C  P  V  G  L  P  A  L  1325
4001  AACTACTTTAATAAAGAAGTCTCCTGCAGAACCAACTTACAGTGTCCCAGTAATACCTACTGTAAAAGCCCTGGTATATGTTGCTACCGTGGACCAAATAG 4100
1326   N  Y  F  N  K  E  V  S  C  R  T  N  L  Q  C  P  S  N  T  Y  C  K  S  P  G  I  C  C  Y  R  G  P  I  A  1359
4101  CTAGACCTCGAAGTTCAAGATATTTGGCAAAATATCTTAAACAAGGACGTCAGGCAAAAGATTGCAGAAAGCGTCTGCAGAAGCCAGGGTCAGGCTCCAGT 4200
1360   R  P  R  S  S  R  Y  L  A  K  Y  L  K  Q  G  R  S  G  K  R  L  Q  K  P  G  S  C  P  A  V  R  P  D   1392
4201  CTGGGCGGGAATGTGTCGTCCGGTGCTTCTGTGATAACGATTGCCGGGGCAATCTGAAATGCTGCAGCAATGGGTGGTGCGACATGCCAGAAGCCC     4300
1393   W  A  G  I  C  V  V  R  C  F  C  D  N  D  C  R  G  N  L  K  C  C  S  N  G  C  R  T  C  Q  K  P   1425
4301  TGCTTTGTCTAGATACTGTATTATAAATAATTTGCAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA 4400
1426   C  F  V  *   1428
4401  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  4439
```

Peptide	Sequence	Position in predicted amino acid sequence
N-terminal	LRRAPYPX(EPGLNV)XCTT	20–37
Internal 1	PPA(KPGVCC)FNP	100–111
Internal 2	TGXVVGAQGSA	209–219

does not contain the full-length cDNA as it lacks both the start and stop codons.

Rescreening the mantle cDNA library with a DIG-labeled insert from clone 1 resulted in the identification of 15 positive clones from approximately one million clones. DNA was extracted from all 15 clones, and restriction mapping analyses were performed. Based on the resulting restriction patterns, the 15 clones can be classified into two groups. One group contained four positive clones for Lustrin A. The longest clone (designated clone 5) had an insert of 4,407 base pairs (Fig. 1). The other group contained 11 clones that exhibit restriction patterns different from that of Lustrin A. The partial nucleotide sequence of clone 7 in the second group was found to encode a polypeptide almost identical to Lustrin A with the exception of a few single amino acid substitutions (Fig. 2) and deletions (data not shown) in the region sequenced. 5′-Rapid amplification of cDNA ends was used to obtain the start codon and its flanking sequence (see "Materials and Methods"). This yielded a 319-bp fragment that provided the first 32 nt of the Lustrin A cDNA (Fig. 1).

Northern Blot Analysis—Northern analysis revealed that Lustrin A mRNA is expressed specifically by cells of the mantle epithelium (Fig. 3). Three probes, representing the 5′-, middle, and 3′-regions of the full-length Lustrin A cDNA, were used: nt 33–1699, nt 3333–3824, and nt 3824–4439. The first of these probes hybridized to two RNA species, 4.7 kilobases (kb) and 5.5 kb in size, respectively, in the RNA isolated from the mantle pallial tissue but not in the RNA from muscle, gill, or stomach tissues. Identical results were obtained when the other two probes were used (data not shown).

Nucleotide Sequence and Deduced Amino Acid Sequence— Sequence analysis of the 4,439-bp cDNA of Lustrin A revealed an open reading frame encoding 1,428 amino acids with the translation initiation codon ATG at nucleotide position 26 (Fig. 1*B*). At position −3 from this initiation codon there exists an adenine nucleotide, and at position +4 there is a guanine nucleotide, representative of a Kozak initiation sequence (44). The first 19 amino acids apparently comprise a signal peptide. It has a hydrophobic core of 11 residues flanked by relatively hydrophilic residues, including a basic arginine residue near the N terminus, similar to known signal peptides found in proteins that are processed in the endoplasmic reticulum and subsequently secreted from the cell. A glycine residue found before the expected cleavage site is consistent with the rule of von Heijne (45). An in-frame stop codon is located at nucleotide position 4309, with a putative polyadenylation signal AATAAT located 13 nucleotides downstream from the stop codon and 10 nucleotides upstream from the poly(A) tail. It thus is very likely

that this cDNA represents the full-length copy of the Lustrin A mRNA coding region.

Modular Structure of Lustrin A—Based on the predicted amino acid sequence, Lustrin A is rich in Ser (16%), Pro (14%), Gly (13%), and Cys (9%) residues (Table II) and contains 15 potential *N*-glycosylation sites. The calculated mass for Lustrin A before any post-translational modifications is 116 kDa.

The deduced amino acid sequence of Lustrin A reveals that it has a modular structure (Fig. 4). Ten cysteine-rich repeats (C1–C10) were identified in Lustrin A. Each repeat consists of 75 to 88 amino acid residues, of which 12 are cysteine. The last two cysteine residues in each repeat are arranged as a Cys-Cys segment, while other cysteine residues are spaced by stretches of 4 to 15 residues. The 10 cysteine-rich repeats share a high degree of sequence identity (45–90%) with each other (Fig. 5). Cys-*X*-Cys and Cys-*X*-*X*-Cys segments, which are frequently found in heavy metal-binding proteins, are not present in these repeats. These repeats also show no sequence similarity to any of the known calcium-binding motifs in the PROSITE data bank. Near the N terminus of each repeat there is a conserved NCT sequence containing a potential site for *N*-glycosylation. The regularity and high conservation of its position in each of the cysteine-rich domains suggest a possible functional role. A second potential *N*-glycosylation site also is present near the center of C1, C2, C4, and C6 (Fig. 5).

The first nine cysteine-rich repeats are connected by eight proline-rich domains (P1–P8). The amino acid sequences of these proline-rich domains are shown in Fig. 6*A*. These domains are 17 to 30 amino acids in length and contain a high proportion of proline residues, most frequently arranged in Pro-Pro, Pro-*X*-Pro, Pro-*X*-*X*-Pro segments. No sequence homology has been observed among the proline-rich domains. However, the sequence of P1 shows some similarity to that of collagen I from various organisms including mammals and insects (46, 47), particularly in the PPGPP that is repeated 3 times in the P1 domain (Fig. 6*B*). Amino acid analysis of Lustrin A did not detect any hydroxyproline, although it was specifically assayed (31); similarly, the absence of regular third position glycines in P1 further show that this is not a triple-helical collagen domain.

The two cysteine-rich domains nearest to the C terminus, C9 and C10, are connected by a large glycine- and serine-rich domain. Of the 272 residues in this domain, 250 are either glycine or serine. Large domains rich in glycine and serine residues have been found in human and mouse cornified envelope protein loricrin (48, 49) and the extracellular matrix protein keratin (50). The GS domain in Lustrin A shares 48% identity with human loricrin, 46% with mouse loricrin, and 47% with a corresponding domain (163 amino acids in size) of human type I cytokeratin 9. Lower correspondence with the silk proteins, sericin (44%) (51) and fibroin heavy chain (40% in 175-amino acid overlap) (52), also was observed.

The cysteine-rich domain C10 is followed by a stretch of 30 amino acids rich in basic residues. Six arginine and four lysine residues are located in this domain. The side chains of these residues are very likely to be positively charged at the pH (~7.4) of the extracellular (extrapallial) space in which matrix assembly and shell mineralization occur (53).

The C-terminal domain of Lustrin A is 45 residues long and exhibits strong sequence similarity to a protein family that

Fig. 1. *A*, cDNA clones of Lustrin A. Clone names and sizes are listed on the *left. Numbers in parentheses* on the *right* indicate the nucleotide range of each clone in the full-length cDNA of Lustrin A. Restriction sites in the cDNA clones are: *E*, *Eco*RI; *B*, *Bam*HI. *B*, cDNA sequence of Lustrin A and the deduced amino acid sequence. The 242-bp RT-PCR fragment sequence is *highlighted*. The putative polyadenylation signal AATAAT at position 4325 is *underlined. Brackets* enclose the 19-amino acid signal peptide sequence. *Underlined* amino acid sequences are identical with those from protein sequencing data (31). Potential *N*-glycosylation sites are shown in *bold. Asterisk* represents the termination codon.
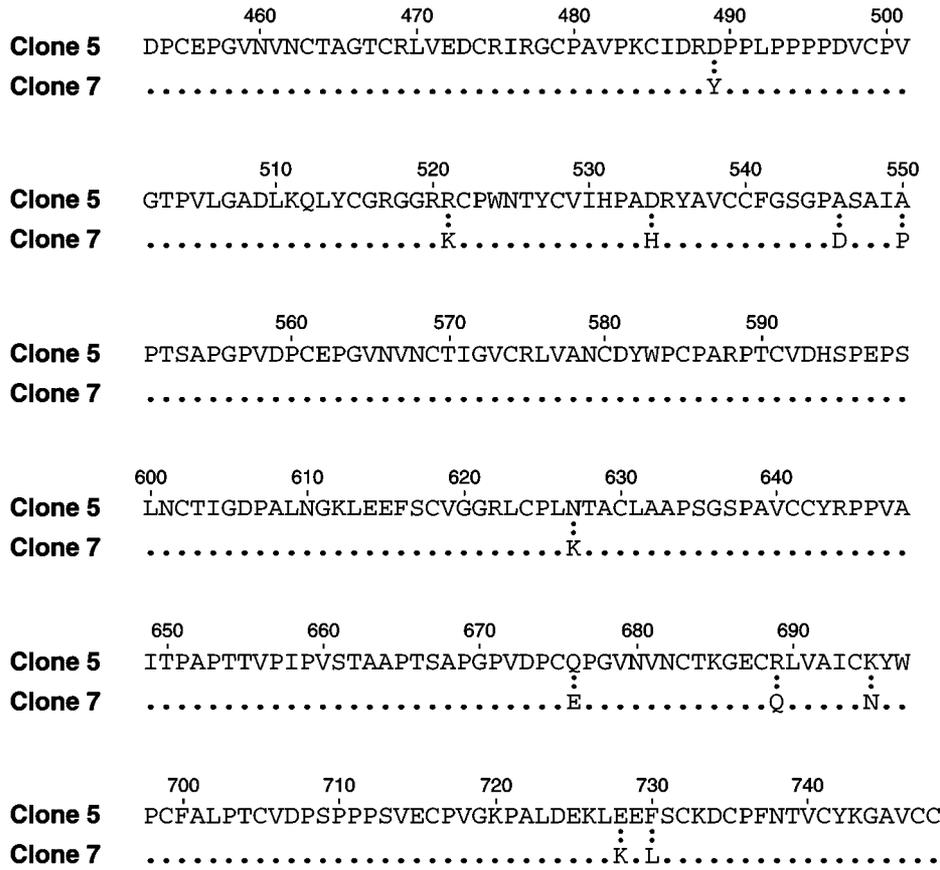
```
                460       470       480       490       500
Clone 5    DPCEPGVNVCTAGTCRLVEDCRIRGCPAVPKCIDRDPPLPPPPDVCPV
                                                  .
Clone 7    .................................Y.............


                510       520       530       540       550
Clone 5    GTPVLGADLKQLYCGRGGRRCPWNTYCVIHPADRYAVCCFGSGPASAIA
                                  .           .            .   .
Clone 7    ..................K.............H............D...P


                560       570       580       590
Clone 5    PTSAPGPVDPCEPGVNVNCTIGVCRLVANCDYWPCPARPTCVDHSPEPS
Clone 7    .................................................


                600       610       620       630       640
Clone 5    LNCTIGDPALNGKLEEFSCVGGRLCPLNTACLAAPSGSPAVCCYRPPVA
                                            .
Clone 7    ............................K....................


                650       660       670       680       690
Clone 5    ITPAPTTVPIPVSTAAPTSAPGPVDPCQPGVNVNCTKGECRLVAICKYW
                                              .           .     .
Clone 7    ..............................E............Q.....N..


                700       710       720       730       740
Clone 5    PCFALPTCVDPSPPPSVECPVGKPALDEKLEEFSCKDCPFNTVCYKGAVCC
                                             . .
Clone 7    ..............................K.L..................
```

FIG. 2. **Comparison of partial amino acid sequences deduced from nucleotide sequences of clones 5 and 7.** Both clones were identified by library screening using DIG-labeled clone 1 insert. Clone 5 codes for Lustrin A, and clone 7 codes for a polypeptide closely related to Lustrin A (see "Results"). *Dots* in the clone 7 sequence represent amino acid residues identical to those in clone 5. *Numbers* on *top* of the clone 5 sequence indicate positions of the amino acid residues in the Lustrin A sequence.

includes several protease inhibitors: red sea turtle basic protease inhibitor (also known as chelonianin) (54), human and pig antileukoproteinase (55, 56), the human elastase-specific inhibitor elafin (57), and several other extracellular proteins: rat, mouse, and camel whey acidic protein (58–60), rat WDNM1 protein (61), and human and chicken Kallmann syndrome protein (62–64). Proteins in this family are distinguished by an array of eight cysteine residues in their sequence known as the "four-disulfide core" motif (65). The partial alignment of some of these proteins with the C-terminal domain of Lustrin A is shown in Fig. 7. Although the function of the whey acidic protein is not known, it has been hypothesized that the WDNM1 protein and the Kallmann syndrome protein have protease inhibiting activity (61–64).

DISCUSSION

We have isolated and characterized the cDNA coding for Lustrin A, a nacre matrix protein with a unique modular structure. Northern blot analyses verify that Lustrin A mRNA is synthesized specifically in the mantle pallial cells that are responsible for secreting shell proteins (Fig. 3). The Lustrin A cDNA clone we isolated is 4,439 bp long, only ~300 bp shorter than the 4.7-kb transcript observed in the Northern hybridization. This difference in size is readily attributable to the absence of some of the 5′-untranslated region and/or the poly(A) tail from the cDNA clone. The detection of two mRNA species with Lustrin A-specific probes indicates either that another gene closely related to Lustrin A also is expressed in the mantle or that the Lustrin A gene can be alternatively spliced to yield two transcripts. In support of this suggestion, we found that
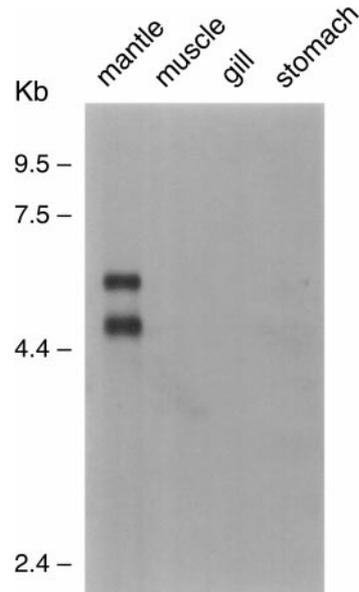

FIG. 3. **Tissue-specific expression of Lustrin A.** Total RNA (25 μg) isolated from abalone mantle, muscle, gill, and stomach were subjected to Northern blot analysis (see "Materials and Methods"). Nt 33–1699 of Lustrin A cDNA was used as a probe. Two transcripts (4.7 kb and 5.5 kb in size) were detected in RNA from the mantle tissue. No signal was detected in RNA from other tissues. Identical results were obtained when two other probes, nt 3333–3824 and nt 3824–4439, were used (data not shown). RNA gel was stained with ethidium bromide before blotting to verify that equal amounts of RNA from different tissues were loaded and no significant degradation of RNA had occurred.

Amino acid	Residues	Mol %
Ser	234	16.39
Pro	198	13.87
Gly	191	13.38
Cys	131	9.17
Val	89	6.23
Ala	85	5.95
Thr	76	5.32
Arg	70	4.90
Leu	63	4.41
Gln	56	3.92
Asp	55	3.85
Lys	32	2.24
Glu	30	2.10
Ile	29	2.03
Tyr	28	1.96
Phe	24	1.68
Asn	18	1.26
Trp	12	0.84
His	5	0.35
Met	2	0.14
Total	1428	

clone 7 encodes a polypeptide almost identical to Lustrin A (Fig. 2); this may be the product of another closely related gene or a splicing or allelic variant of Lustrin A.

The calculated mass for Lustrin A before any post-translational modifications is 116 kDa. However, the polypeptide chain extracted from the nacre matrix and used for amino acid sequence analysis is only 65 kDa (31). This discrepancy in size could result from one or more of the following possibilities: (i) Lustrin A may have been fragmented during extraction from its covalently cross-linked network in the matrix (31), (ii) post-transcriptional and/or post-translational modification may reduce the size of the protein, and (iii) Lustrin A and the 65-kDa polypeptide may be encoded by two different but closely related genes. We are producing antibodies against Lustrin A (the protein defined by the sequence of the full-length cDNA cloned and characterized in the work presented here) to investigate these possibilities.

The characteristic feature of Lustrin A is its modular structure, consisting of: (i) ten highly conserved cysteine-rich domains, (ii) eight proline-rich domains, (iii) a glycine- and serine-rich domain, (iv) a basic domain, and (v) a C-terminal domain with marked sequence similarity to known proteinase inhibitors.

There are 10 cysteine-rich repeats in Lustrin A. The high degree of sequence identity among them suggests that they are likely to undergo similar folding. The high frequency and complete positional conservation of the cysteine residues within these repeats suggest that these domains may have globular structures stabilized by intradomain disulfide bonds, although the possibility of some interchain disulfide bonds cannot be excluded. Proline residues also are abundant in the cysteine-rich repeats; they are scattered throughout the entire repeat, and their positions are often conserved. Because of the known α-helix- and β-sheet-disrupting effect of proline residues (66), it is not surprising that Chou-Fasman calculations (67) predict that the cysteine-rich repeats contain no extended α-helices or β-sheets. A search for heavy metal- or calcium-binding motifs in these repeats was unfruitful. We hypothesize that the main function of the cysteine-rich repeats is in protein-protein interactions governing the complex self-assembly of the multicomponent matrix that in turn helps control the mineralization of nacre. Our finding that the yield of Lustrins is increased by the

inclusion of thiol reagents in the extraction medium[2] supports our hypothesis that Lustrins may be covalently linked via disulfide bonds with one another or with other proteins of the nacre matrix.

The proline-rich domains, although sharing no sequence homology, are similar in that they all have high proline content. Prolines in these domains generally are located in tandem or at every other one or two positions (Fig. 6*A*). This density of proline residues is predicted to force these domains to adopt extended structures, most likely in the form of "polyproline II helixes" (extended structures of three residues per turn) (66). Such proline-rich regions often are found to be involved in binding, serving as "sticking arms" (66) by virtue of the exposure of the side chains of other residues for interaction with other molecules, but the lack of homology among the non-proline portions of the eight different domains may make this role in Lustrin A less likely. Because the proline-rich domains all are located between the cysteine-rich domains and will have extended, rod-like structures, their primary function may be to serve as spacers separating the cysteine-rich domains so these can fold independently. One of the proline-rich domains, P1, shows some similarity to collagen I (Fig. 6*B*). Collagen is the major matrix protein in several other biomineralized extracellular materials such as dentin, bone (68, 69), and avian eggshell (70, 71). The data shown here demonstrate for the first time the existence of a sequence with some collagen-like domains in the abalone shell nacre matrix, but the absence of hydroxyproline and the regular third-position glycines characteristic of collagen limit this similarity.

The cysteine-rich module and the proline-rich module are arranged in tandem and repeated nine and eight times, respectively, in the N-terminal two-thirds of Lustrin A (Fig. 4). A similar arrangement of cysteine-rich modules and proline-rich modules (with very different sequences) has been observed in frustulins (72, 73), a family of glycoproteins intimately associated with diatom cell walls. Interestingly, this arrangement is the only similarity shared by these two families of proteins, which differ in almost all other aspects of their structures. The sizes and numbers of their cysteine-rich repeats, the number of cysteine residues in each repeat, the spacing between the cysteine residues and the hydroxylation of prolines in the proline-rich domains all are dissimilar, as are their sequences. However, it is tempting to speculate that the similar modular arrangement of the biomineralization matrix-like proteins from the calcium carbonate shell of an animal and the silica "shell" of a unicellular microalga may reflect the convergent evolution of diblock copolymer-like protein domains that might serve related functions.

The glycine-, serine-rich region is the largest discrete structural domain in Lustrin A. It contains 85 glycine and 165 serine residues (for a total of 250 residues out of 272), most in $(GS)_x$ and $(GSSS)_y$ repeats. Analyses of secondary structure using the Chou and Fasman method (67) indicate that this domain has no α-helix or β-sheet content but is rich in turns. Flexibility indices indicate a high degree of flexibility attributable to the large number of glycine residues. Six aromatic residues, including four tryptophans and two phenylalanines, are located in this domain, with the phenylalanine in tandem with the tryptophan. It is well known that the aromatic side chains have a tendency to stack through alignment of their phenyl rings at a preferred distance of 4.5–7 Å parallel (74) or perpendicular (75) to one another. This association of the tryptophan and phenylalanine side chains thus will force the $(GS)_x$ and $(GSSS)_y$

---

[2] A. M. Belcher, J. Hagopian, and D. E. Morse, unpublished observations.

A

signal	MERFLWVLCIAAGFSVNYG	19
N-terminus	LRRAP	24
C1	YPCEPGLNVNCTTGECRLVFSCSLRRCGVRPECVDRSPVPS INCTIGKPTIDTNLQEISCAPDGSCPATTGCVRGPPAKPGVCC	108
P1	FNPSSGPPGPPRPPGPPRPPGPPQPDPNLL	138
C2	DPCFPGKNVNCTSGECRLMADCQHQSCPALPYCVAPSPNVT VPCPIGKSAIDRNLREFSCLRNRDACPRSTGCVVGAQGSAAVCC	223
P2	YRPPLVPGPTPTDPNPL	240
C3	DPCFPGKNVNCTAGECRLVADCSRKGCPAGPTCVDPSPVPS LNCDIGKPALNSYGNEISCAGGGACPVNTVCVAHPSGAPAVCC	324
P3	FKPAGPTTPQPPTIPQPPTTPSSPTG	350
C4	DPCEPGVNVNCTAGTCRLVVDCRFPGCPAVPKCVDPSSKPS LNCSIGDPALNPNLQEISCVGGAACPRNTACFAAPSGSPAVCC	434
P4	YTSGPPRPEPPSPSPPTG	452
C5	DPCEPGVNVNCTAGTCRLVEDCRIRGCPAVPKCIDRDPPLPPPP DVCPVGTPVLGADLKQLYCGRGGRRCPWNTYCVIHPADRYAVCC	540
P5	FGSGPASAIAPTSAPGPV	558
C6	DPCEPGVNVNCTIGVCRLVANCDYWPCPARPTCVDHSPEPS LNCTIGDPALNGKLEEFSCVGGRLCPLNTACLAAPSGSPAVCC	642
P6	YRPPVAITPAPTTVPIPVSTAAPTSAPGPV	672
C7	DPCQPGVNVNCTKGECRLVAICKYWPCFALPTCVDPSPPPS VECPVGKPALDEKLEEFSCKDCPFNTVCYKGAVCC	748
P7	VPWSGNRPSGPAGPAGPAGPERPATSVPL	777
C8	DPCTPGLNVNCTSGVCRLVEDCRRPGCPAVPTCIDRDPPLPPPP DVCPVGTPVLGRDLKQLYCGRGGKRCPGNTYCVIHPADRYAVCC	865
P8	FGSGPGPQPPIPTPPPPTT	884
C9	YPCTPANINCTAGECRLVAYCNAVPCGRTTPTCVDPSPPPT RKCPVGKPVLTPRLTEFRCYPRVRLCPGDSFCLRGPGDEPGVCC	969
	WDNRLRPTQ	978
"GS" domain	GSGSGSGSGSGSGSGSSSSSSSGSTSGSGSGSGSGSSSGSGSGSSSASGS GSGSGSSSASGSGSSSGSGSGSSSGSGSGSGSGSSSGSGSGSSSGSSS VSNSWTGSGSSSGSGSGSSSWSGSGSSSGTGSGSSSWFGSGSSSGSGS DSSSGSSSASGSGSSSGSSSGSGSGSSLWFGSGSSSGTGSGSSSGSSS GSGSDSSSGSSSGSTSGSSSGSGSASGSGTGS	1250
	GKGASYDTDADSGSDNRSPGYLPQ	1274
C10	DPCTPGLYINCTAGTCRLTAWCLYNFCPAVPTCVDSSPDAS GECPVGLPALNYFNKEVSCRTNLQCPSNTYCKSPGICC	1353
Basic	YRGPIARPRSSRYLAKYLKQGRSGKRLQKP	1383
C-terminus	GSCPAVRPDWAGICVVRCFCDNDCRGNLKCCSNGCGRTCQKPCFV	1428

B
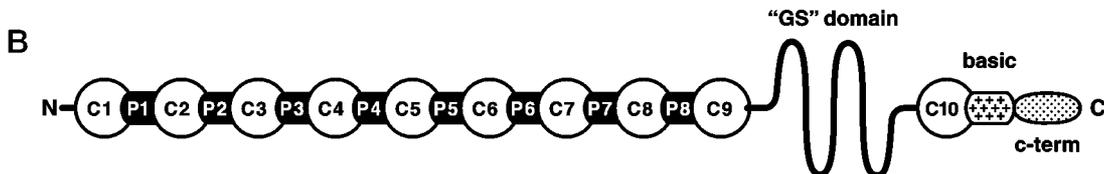
Fig. 4. *A*, modular structure of Lustrin A. *Boxed sequence* (*P1*) shows a limited sequence similarity to collagen I (see Fig. 6*B* and discussion of limitations in text). The glycine- and serine-rich domain (*GS domain*) also is *boxed*. Aromatic residues in the GS domain are *shaded*. Basic residues in the basic domain are in *bold*. Lysine, asparagine, and tyrosine residues in the basic domain are *underlined*. *Highlighted* C-terminal domain shares homology with proteins in the four-disulfide core family (see Fig. 7). *B*, schematic representation of the modular structure of Lustrin A.

repeats to form "glycine loops" (76), although in Lustrin A serine rather than glycine will be the predominant residue in the loops. Because the interactions between the phenyl rings are weak, the glycine loops can be reversibly opened when a stretching force is applied. Such rubber-like glycine loops pre-

viously had been found in three classes of proteins (one of which is an extracellular matrix superfamily): keratins and other intermediate filament proteins, loricrins, and single-stranded RNA-binding proteins (76). The presence of glycine loops in Lustrin A suggests that it may have an elastic property

```
      1                                                  45
C1   YPCEPGLNVNCTTGECRLVFSCSLRRCG.VRPECVD...RSPVPS
C2   DPCFPGKNVNCTSGECRLMADCQHQSCP.ALPYCVA...PSPNVT
C3   DPCFPGKNVNCTAGECRLVADCSRKGCP.AGPTCVD...PSPVPS
C4   DPCEPGVNVNCTAGTCRLVVDCRFPGCP.AVPKCVD...PSSKPS
C5   DPCEPGVNVNCTAGTCRLVEDCRIRGCP.AVPKCIDRDPPLPPPP
C6   DPCEPGVNVNCTIGVCRLVANCDYWPCP.ARPTCVD...HSPEPS
C7   DPCQPGVNVNCTKGECRLVAICKYWPCF.ALPTCVD...PSPPPS
C8   DPCTPGLNVNCTSGVCRLVEDCRRPGCP.AVPTCIDRDPPLPPPP
C9   YPCTPA.NINCTAGECRLVAYCNAVPCGRTTPTCVD...PSPPPT
C10  DPCTPGLYINCTAGTCRLTAWCLYNFCP.AVPTCVDS...SPDAS

      46                                               89
C1   INCTIGKPTIDTNLQEISCAPDG.SCPATTGCVRGPPAKPGVCC
C2   VPCPIGKSAIDRNLREFSCLRNRDACPRSTGCVVGAQGSAAVCC
C3   LNCDIGKPALNSYGNEISCAGGG.ACPVNTVCVAHPSGAPAVCC
C4   LNCSIGDPALNPNLQEISCVGGA.ACPRNTACFAAPSGSPAVCC
C5   DVCPVGTPVLGADLKQLYCGRGGRRCPWNTYCVIHPADRYAVCC
C6   LNCTIGDPALNGKLEEFSCVGGR.LCPLNTACLAAPSGSPAVCC
C7   VECPVGKPALDEKLEEFSCKD....CPFNTVCYKG.....AVCC
C8   DVCPVGTPVLGRDLKQLYCGRGGKRCPGNTYCVIHPADRYAVCC
C9   RKCPVGKPVLTPRLTEFRCYPRVRLCPGDSFCLRGPGDEPGVCC
C10  GECPVGLPALNYFNKEVSC.RTNLQCPSNTYCKS.....PGICC
```

FIG. 5. **Sequence alignment of the 10 cysteine-rich domains (*C1–C10*) of Lustrin A.** Sequences of all 10 repeats are aligned according to positions of the conserved cysteine residues. Consensus amino acid residues (>50% conservation) are *shaded*. Conserved cysteine residues are *highlighted*. Potential *N*-glycosylation sites are in *bold*. Positions of the domains are shown in Fig. 4*A*.

A
```
P1  FNPSSGPPGPPRPPGPPRPPGPPQPDPNLL    15/30
P2  YRPPLVPGPTPDPNPL                  7/16
P3  FKPAGPTTPQPPTIPQPPTTPSSPTG        10/26
P4  YTSGPPRPEPPSPSPPTG                8/18
P5  FGSGPASAIAPTSAPGPV                4/18
P6  YRPPVAITPAPTTVPIPVSTAAPTSAPGPV    9/30
P7  VPWSGNRPSGPAGPAGPAGPERPATSVPL     8/29
P8  FGSGPGPQPPIPTPPPPTT               9/19
```

B
```
P1    FNPSSGPPGPPRPPGPPRPPGPPQPDPNLL
Col I RTGDAGPAGPPGPPGPPGPPGPPSGGYDLS
```

FIG. 6. *A*, sequences of the proline-rich domains (*P1–P8*). Proline residues are *highlighted. Numbers* on the *right* correspond to the number of proline residues relative to the number of total residues in each domain. Positions of domains are shown in Fig. 4*A*. *B*, sequence alignment of P1 with collagen I α-1 chain from calf. Residues 732–761 of Col I are shown; those showing consensus are *shaded*.

enabling it to function as an extensor molecule.

The strikingly arginine- and lysine-rich basic domain in Lustrin A suggests that this region would have the capacity to interact with the anionic molecules during formation of the shell. Three classes of anions are involved (or potentially involved) in this mineralization: (i) the CO_3^{2-} that participates directly in formation of the $CaCO_3$ crystal lattices, (ii) the polyanionic aragonitic proteins shown to be responsible for cooperative control of crystal polymorph selection and atomic lattice orientation (22, 26, 27), and (iii) the anionic groups of sugars in the matrix polysaccharides and glycoproteins, potentially including carboxylates, uronates, sulfates, etc. Additionally, of course, this domain also has the capacity to participate in protein-protein cross-linking by virtue of its high lysine, tyrosine, and glutamine content (77, 78). There are four lysine, three tyrosine, and two glutamine residues in this domain. Lysine and glutamine residues can form N^ϵ-(γ-glutamyl)lysine isopeptide bonds, tyrosine residues can form diphenyl linkages, and the lysine side chains can be modified to participate in intermolecular cross-linking via Schiff's base conjugates. While only speculative at this point, it is interesting to find potential anchoring sites for the polyanionic aragonite-determining proteins in this domain.

The presence of a C-terminal protease inhibitor-like domain

in Lustrin A suggests that if this domain is functional the protein also may have the capacity to bind proteases or other proteins with protease-like structures. Covalent linkage of the N terminus of this domain to the polypeptide chain preceding it is unlikely to affect this binding capacity, based on the following observations. First, proteins of the four-disulfide core family analyzed thus far all have similar structures due to their four intramolecular disulfide bonds. Crystal structures of the elastase-specific inhibitor elafin (79) and human seminal plasma inhibitor (also called human antileukoproteinase) (80), and the solution structure of Na^+,K^+-ATPase inhibitor SPAI-1 revealed by NMR (81) all exhibit a stretched planar spiral shape. Because the C-terminal domain of Lustrin A contains the eight cysteine residues at the canonical positions, it may exhibit a similar structure. Second, the N-terminal region of elafin is flexible, located on the opposite side of the active site loop and not essential for elastase inhibitory activity (79). Third, two protease inhibitor domains are present in human antileukoproteinase. The N-terminal domain binds trypsin, and the C-terminal domain binds chymotrypsin. The covalent linking of the second domain to the first does not affect its structure or function (80).

It is possible that the C-terminal protease inhibitor-like domain in Lustrin A functions as an active protease inhibitor, protecting the secreted proteins of the shell-forming matrix and machinery from degradation. Incorporation of protease inhibitors into an outer protective layer seems to be a common scheme used by a variety of organisms. Thus, for example, elafin (a serine protease inhibitor) and cystatin α (a cysteine protease inhibitor) are found cross-linked to loricrin and other cornified cell envelop proteins in the keratinocytes during terminal differentiation in mammalian epidermis (82, 83).

It is not surprising to find that Lustrin A shows partial homology to several extracellular matrix proteins, as the molluscan shell is a specialized extracellular material produced from epithelial secretions. It is interesting to note, however, that Lustrin A combines several structural motifs into one molecule: the alternating cysteine-rich module and proline-rich module design seen also (although to a lesser extent) in frustulins, the glycine loop motif of loricrins and keratins, the four-disulfide core motif of several protease inhibitors, and a

FIG. 7. **Sequence alignment of the C-terminal domain of Lustrin A with several proteins in the four-disulfide core family.** Canonical cysteine residues are *highlighted*. Other conserved amino acid residues are *shaded*. Proteins used in the figure are the WDNM1 protein from rat (residues 1384–1428), the Kallmann syndrome protein (*KALM*) from human (residues 17–60), the basic protease inhibitor chelonianin (*CHE*) from red sea turtle (residues 132–178), the whey acidic protein (*WAP*) from rat (residues 81–129), the antileukoproteinase (*ALK1*) from pig (residues 70–115), and the elastase-specific inhibitor elafin (*ELAF*) from human (residues 74–117).

basic domain. Such a combination is unique and suggests that Lustrin A is a multifunctional protein. In addition to its structural role in the insoluble nacre matrix framework, Lustrin A also may play important roles interacting with the polyanionic aragonite-determining proteins, protecting the protein components of the matrix from degradation, and conferring elastic resiliency to the high performance microlaminate composite of the molluscan shell.

REFERENCES

1. Wada, K. (1961) *Bull. Natl. Pearl Res. Lab.* **7,** 703–785
2. Towe, K. M., and Hamilton, G. H. (1968) *Calcif. Tissue Res.* **1,** 306–318
3. Bevelander, G., and Nakahara, H. (1969) *Calcif. Tissue Res.* **3,** 84–92
4. Wise, S. W. (1970) *Ecolgae Geol. Helv.* **63,** 775–797
5. Wise, S. W. (1970) *Science* **167,** 1486–1488
6. Nakahara, H. (1979) *Jpn J. Malacology* **38,** 205–211
7. Nakahara, H. (1983) *Biomineralization and Biological Metal Accumulation* (Westbroek, P., and De Jong, E. W., eds) pp. 225–230, Reidel, Dordrecht
8. Meenakshi, V. R., Hare, P. E., and Wilbur, K. M. (1971) *Comp. Biochem. Physiol.* **40B,** 1037–1043
9. Gregoire, C. (1972) in *Chemistry and Zoology* (Florkin, M., and Scheer, B. eds) pp. 45–102, Academic Press, New York
10. Weiner, S. (1979) *Calcif. Tissue Int.* **29,** 163–167
11. Nakahara, H., Bevelander, G., and Kakei, M. (1982) *Japan J. Malac.* **41,** 33–46
12. Weiner, S., and Traub, W. (1984) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **304,** 425–434
13. Fritz, M., Belcher, A. M., Radmacher, M., Walters, D. A., Hansma, P. K., Stucky, G. D., Morse, D. E., and Mann, S. (1994) *Nature* **371,** 49–51
14. Zaremba, C., Belcher, A. M., Fritz, M., Li, Y., Mann, S., Hansma, P. K., Morse, D. E., Speck, J. S., and Stucky, G. D. (1996) *Chem. Mater.* **8,** 676–690
15. Currey, J. D. (1977) *Proc. R. Soc. Lond. B Biol. Sci.* **196,** 443–463
16. Jackson, A. P., Vincent, J. F. V., and Turner, R. M. (1988) *Proc. R. Soc. Lond. B Biol. Sci.* **234,** 415–440
17. Cariolou, M. A., and Morse, D. E. (1988) *J. Comp. Physiol.* **157B,** 717–729
18. Weiner, S., and Addadi, L. (1997) *J. Mater. Chem.* **7,** 689–702
19. Sarikaya, M. (1994) *Microsc. Res. Tech.* **27,** 360–375
20. Morse, D. E., Cariolou, M. A., Stucky, G. D., Zaremba, C. A., and Hansma, P. K. (1993) *Mater. Res. Soc. Symp. Proc.* **292,** 59–67
21. Belcher, A. M., Hansma, P. K., Stucky, G. D., and Morse, D. E. (1997) *Acta Metal. Mater.*, in press
22. Belcher, A. M., Wu, X. H., Christensen, R. J., Hansma, P. K., Stucky, G. D., and Morse, D. E. (1996) *Nature* **381,** 56–58
23. Crenshaw, M. A. (1972) *Biomineralisation* **6,** 6–11
24. Weiner, S., and Hood, L. (1975) *Science* **190,** 987–989
25. Worms, D., and Weiner, S. (1986) *J. Exp. Zool.* **237,** 11–20
26. Falini, G., Albeck, S., Weiner, S., and Addadi, L. (1996) *Science* **271,** 67–69
27. Walters, D. A., Smith, B. L., Belcher, A. M., Paloczi, G. T., Stucky, D. E., and Hansma, P. K. (1997) *Biophys. J.* **72,** 1425–1433
28. Weiner, S., and Traub, W. (1980) *FEBS Lett.* **111,** 311–316
29. Weiner, S. (1986) *Crit. Rev. Biochem.* **20,** 365–408
30. Nakahara, H. (1991) *Mechanisms and Phylogeny of Mineralization in Biological Systems* (Suga, S., and Nakahara, H. eds) pp. 343–350, Springer-Verlag, New York
31. Belcher, A. M. (1996) *Spatial and Temporal Resolution of Interfaces, Phase Transitions and Isolation of Three Families of Proteins in Calcium Carbonate-based Biocomposite Materials.* Ph.D. thesis. University of California, Santa Barbara, CA.
32. Schäffer, T. E., Ionescu-Zanetti, C., Proksch, R., Fritz, M., Walters, D. A., Almqvist, N., Zaremba, C. M., Belcher, A. M., Smith, B. L., Stucky, G. D., Morse, D. E., and Hansma, P. K. (1997) *Chemistry of Materials* **9,** 1731–1740
33. Miyamoto, H., Miyashita, T., Okushima, M., Nakano, S., Morita, T., and Matsushiro, A. (1996) *Proc. Natl. Acad. Sci. U. S. A.* **93,** 9657–9660
34. Erben, H. K., and Watabe, N. (1974) *Nature* **284,** 128–130
35. Manne, S., Zaremba, C. M., Giles, R., Huggins, L., Walters, D. A., Belcher, A. M., Morse, D. E., Stucky, G. D., Didymus, J. M., Mann, S., and Hansma, P. K. (1994) *Proc. R. Soc. Lond. B Biol. Sci.* **256,** 17–23
36. Watabe, N. (1981) *Prog. Crystal Growth Characteristics* **4,** 99–147
37. Mutvei, H. (1979) *Scanning Electron Microsc.* **2,** 457–462
38. Wada, K. (1972) *Biomineralisation* **6,** 141–159
39. Sudo, S., Fujikawa, T., Nagakura, T., Ohkubo, T., Sakaguchi, K., Tanaka, M., Nakashima, K., and Takahashi, T. (1997) *Nature* **387,** 583–584
40. Bowen, C. E., and Tang, H. (1996) *Comp. Biochem. Physiol.* **115A,** 269–275
41. Chomzynski, P., and Mackey, K. (1995) *Biotechniques* **19,** 942–945
42. Frohman, M. A. (1990) *PCR Protocols: a Guide to Methods and Applications* (Innis, M. A., Gelfand, D. H., Sninsky, J. J., and White, T. J. eds) pp. 28–38, Academic Press, New York
43. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
44. Kozak, M. (1986) *Cell* **44,** 283–292
45. von Heijne, G. (1985) *J. Mol. Biol.* **184,** 99–105
46. Fietzek, P. P., Rexrodt, F. W., Wendt, P., Stark, M., and Kuehn, K. (1972) *Eur. J. Biochem.* **30,** 163–168
47. Kramer, J. M., Cox, G. N., and Hirsh, D. (1982) *Cell* **30,** 599–606
48. Hohl, D., Mehrel, T., Lichti, U., Turner, M. L. Roop, D. R., and Steinert, P. M. (1991) *J. Biol. Chem.* **266,** 6626–6636
49. Mehrel, T., Hohl, D., Rothnagel, J. A., Longley, M. A., Bundman, D., Cheng, C., Lichti, U., Bisher, M. E., Steven, A. C., Steinert, P. M., Yuspa, S. H., and Roop, D. R. (1990) *Cell* **61,** 1103–1112
50. Langbein, L., Heid, H. W., Moll, I., and Franke, W. W. (1993) *Differentiation* **55,** 57–71
51. Okamoto, H., Ishikawa, E., and Suzuki, Y. (1982) *J. Biol. Chem.* **257,** 15192–15199
52. Tsujimoto, Y., and Suzuki, Y. (1979) *Cell* **18,** 591–600
53. Crenshaw, M. A. (1972) *Biol. Bull.* **143,** 506–512
54. Kato, I., and Tominaga, N. (1979) *Fed. Proc.* **38,** 832
55. Heinzel, R., Appelhans, H., Gassen, G., Seemuller, U., Machleidt, W., Fritz, H., and Steffens, G. (1986) *Eur. J. Biochem.* **160,** 61–67
56. Farmer, S. J., Fliss, A. E., and Simmer, R. C. M. (1990) *Mol. Endocrinol.* **4,** 1095–1104
57. Saheki, T., Ito, F., Hagiwara, H., Saito, Y., Kuroki, J., Tachibana, S., and Hirose, S. (1992) *Biochem. Biophys. Res. Commun.* **185,** 240–245
58. Campbell, S. M., Rosen, J. M., Hennighausen, L. G., Strech-jurk, U., and Sippel, A. E. (1984) *Nucleic Acids Res.* **12,** 8685–8697
59. Hennighausen, L. G., and Sippel, A. E. (1982) *Nucleic Acids Res.* **10,** 3733–3744
60. Beg, O. U., Von Bahr-Lindstrom, H., Zaidi, Z. H., and Joernvall, H. (1986) *Eur. J. Biochem.* **159,** 195–201
61. Dear, T. N., and Kefford, R. F. (1991) *Biochem. Biophys. Res. Commun.* **176,** 247–254
62. Legouis, R., Hardelin, J. P., Levilliers, J., Claverie, J. M., Compain, S., Wunderle, V., Millasseau, P., Le Paslier, D., Cohen, D., Caterina, D., Bougueleret, L., Delemarre, van der Wall, H., Lutfalla, G., Weissenbach, J., and Petit, C. (1991) *Cell* **67,** 423–435
63. Franco, B., Guioli, S., Pragliola, A., Incerti, B., Bardoni, B., Tonlorenzi, R., Carrozzo, R., Maestrini, E., Pieretti, M., Taillonmiler, P., Brown, C. J., Willard, H. F., Lawrence, C., Persico, M. G., Camerino, G., and Ballabio, A. (1991) *Nature* **353,** 529–536
64. Legouis, R., Cohen-Salmon, M., Del Castillo, I., Levilliers, J., Capy, L., Mornon, J. P., and Petit, C. (1993) *Genomics* **17,** 516–518
65. Hennighausen, L. G., and Sippel, A. E. (1982) *Nucleic Acids Res.* **10,** 2677–2684
66. Williamson, M. P. (1994) *Biochem. J.* **297,** 249–260
67. Chou, P. Y., and Fasman, G. D. (1978) *Adv. Enzymol.* **47,** 45–148
68. Schlueter, R. J., and Veis, A. (1964) *Biochemistry* **3,** 1650–1665
69. Weiner, S., Arad, T., Ziv, V., and Traub, W. (1992) *Chemistry and Biology of Mineralized Tissues* (Slavkin, H., and Price, P. A., eds) Excerpta Medica, Amsterdam
70. Wong, M., Hendrix, M. J. C., Mark, K. V. D., Little, C., and Stern, R. (1984) *Dev. Biol.* **104,** 28–36
71. Arias, J. L., Fink, D. J., Xiao, S-Q., Heuer, A. H., and Caplan, A. I. (1993) *Int. Rev. Cytol.* **145,** 217–251
72. Kröger, N., Bergsdorf, C., and Sumper, M. (1996) *Eur. J. Biochem.* **239,**

259–264

73. Kröger, N., Bergsdorf, C., and Sumper, M. (1994) *EMBO J.* **13,** 4676–4683
74. Burley, S. K., and Petsko, G. A. (1985) *Science* **229,** 23–28
75. Burley, S. K., and Petsko, G. A. (1989) *Trends Biotech.* **7,** 354–359
76. Steinert, P. M., Mack, J. W., Korge, B. P., Gan, S-Q., Haynes, S. R., and Steven, A. C. (1991) *Int. J. Biol. Macromol.* **13,** 130–139
77. Brunet, P. C. J. (1967) *Endeavor* **26,** 68–74
78. Gordon, J., and Carriker, M. R. (1980) *Marine Biol.* **57,** 251–260

79. Tsunemi, M., Matsuura, Y., Sakakibara, S., and Katsube, Y. (1996) *Biochemistry* **35,** 11570–11576
80. Grütter, M. G., Fendrich, G., Huber, R., and Bode, W. (1988) *EMBO J.* **7,** 345–351
81. Kozaki, T., Kawakami, Y., Tachibana, S., Hatanaka, H., and Inagaki, F. (1994) *Pept. Chem.* 405–408
82. Steinert, P. M., and Marekov, L. N. (1995) *J. Biol. Chem.* **270,** 17702–17711
83. Takahashi, M., Tezuka, T., and Katunuma, N. (1992) *FEBS Lett.* **308,** 79–82