学位論文

Primary structure of the major soluble protein in the shell matrix of the scallop *Patinopecten yessoensis*

ホタテガイにおける殻内気質中の
可溶性タンパク質の一次構造

平成11年11月博士（理学）申請

東京大学大学院理学系研究科
地質学専攻

更科　功

①

Primary structure of the major soluble protein in the shell matrix of the scallop *Patinopecten yessoensis*

1999

Geological Institute,
University of Tokyo

Isao Sarashina

Abstract

Soluble proteins in the foliated calcite shell layer of the scallop *Patinopecten yessoensis* and related pectinid species (*Pecten albicans* and *Chlamys farreri*) were characterized using biochemical and molecular biological techniques. SDS-PAGE of these molecules revealed eight protein bands (150, 125, 104, 77, 49, 42, 40, and 15 kDa) for *P. yessoensis*, five bands (77, 49, 42, 36, and 18 kDa) for *P. albicans*, and three bands (65, 30, and 17 kDa) for *C. farreri*. At least three proteins (104, 77, and 49 kDa) of *P. yessoensis* share the N-terminal amino acid sequence (LDTDK DLEFH LDSLL NAA), and at least three (77, 49, and 42 kDa) of *P. albicans* also share the N-terminal sequence (SDTDA DTDED EEN). Periodic Acid/Schiff staining indicated that the all the proteins of *P. yessoensis* and three of the proteins (77, 49, and 42 kDa) appeared in SDS-PAGE gels are glycosylated. Stains-All staining indicated that at least four proteins (150, 125, 104, and 77 kDa) of *P. yessoensis* may have cation-binding potential.

A cDNA encoding one of these proteins in *P. yessoensis* (MSP-1) was amplified by polymerase chain reaction (PCR) and was characterized through molecular cloning of the amplified products. The full-length cDNA of 2978 base pairs in length contained an open reading frame coding for a hydrophobic putative signal peptide of 20 amino acid residues followed by a polypeptide of 820 amino acid residues. The deduced sequence of MSP-1 revealed a highly modular structure, consisting of a high proporton of Ser (31%), Gly (25%), and Asp (19.5%); MSP-1 typifies an acidic glycoprotein of mineralized tissues.

The characteristic feature of MSP-1 is its repeated modular structure, being comprised of an N-terminal domain, a basic domain, an SGD domain, four highly conserved units, and a C-terminal domain. The

N-terminal, SGD, and C-terminal domains are highly acidic, and could interact with calcium ions. The basic domain can potentially interact with other acidic proteins exploiting the positively charged character and/or the capability of making protein-protein cross-linkage by virtue of its high lysine content.

The four highly conserved units are arranged in tandem. The repeating structure would enable MSP-1 to cover the crystal surface with only a small number of the molecules, resulting in the high functional effect on crystal growth, such as nuleator and/or inhibitor. Each unit contains the SG domain, the D domain, and the G domain. The SG domain showed the highest sequence similarity with the "GS" domain of Lustrin A, a gastropod shell matrix protein, and may have a high degree of flexibility. One function of the SG domains may be to serve as spacers separating the D and G domains so that each of them can fold independently. The G domain has a core of basic residues. The core may interact with other acidic matrix proteins.

It has often been postulated that Asp-rich domains of shell proteins play a role as a template on which epitaxial growth of mineral phase takes place, but definite amino acid sequences of the Asp-rich domains have not been reported. The D domain sequence of MSP-1 described here is the first to be reported for the typical Asp-rich domains of shell proteins. Each D domain contains a total of 9 to 10 potential phosphorylation sites and a total of 3 to 5 potential *N*-glycosylation sites. The D domains are more acidic than the other three acidic domains in MSP-1 and have consensus sequence motifs of (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp). In contrast with prevalent expectations, (Asp-Gly)$_n$-, (Asp-Ser)$_n$-, and (Asp-Gly-X-Gly-X-Gly)$_n$- type sequence motifs do not exist in the D domain, demanding revision of previous theories of protein-mineral interactions. All the four acidic domains (including the D domain) of MSP-1 may interact with calcium ions on crystal surfaces and act as non-specific inhibitors, while only the D domain, perhaps, can have other functions, such as specific inhibition/nucleation of particular crystal planes, and nucleation of calcite. The conservativeness of amino acid sequences among the four D domains, the existence of some regular motifs, and the exclusive use of Asp residues out of the two acidic amino acid residues, Asp and Glu, suggest that COO$^-$ groups in the side chains of Asp residues are taking a rigid arrangement in the D domains, and this regularity of COO$^-$ groups may be important for specific control of crystal growth. The D domains showed a high sequence similarity with phosphophoryn, a major protein in dentin, being involved in mineralization. But the regular motifs such as (Asp-Ser-Ser) and (Asp-Ser) of phosphophoryn do not exist in the D domains of MSP-1, and the overall arrangement of Asp-residues are different between them. Because Asp residues have been thought to be important in protein-mineral interactions, they may be functionally related, but the evolutionary relation between them is obscure.

The following two possible models are proposed as to the calcite layer formation controlled by MSP-1:

(1) The D domains in MSP-1 in solution interact with one kind of crystal surfaces of foliated calcite laths, the rhombohedral forms (110), resulting from a precise stereochemical fit between the distance of the Asp residues of MSP-1 and the spacing between calcium ions in the crystal lattices, while they do not interact with other surfaces, such as the rhombohedral forms (101). Consequently, inhibition of crystal growth occurs only on the (110) surfaces. The (101) surfaces of the laths are allowed to grow and the characteristic shape of the laths is formed.

(2) The D domains in MSP-1 in solution interact with particular growing surfaces of foliated calcite laths, the rhombohedral forms (101), and inhibit further crystal growth on these crystal planes. Then, the G

domains (and the basic domain) in the MSP-1 are attached by some substrates, such as insolble matrix proteins, to which another layer of MSP-1 is attached, with the D domains being arranged on the opposite side, exposed to the solution. Consequently, COO^- groups in the Asp residues in the D domains adsorb Ca^{2+} in solution, inducing oriented nucleation of calcite. Based on the results obtained in this study, suggestions for further studies are also made.

Contents

List of Figures

List of Tables

Abbreviations

bp	base pairs
cDNA	complementary DNA
DNA	deoxyribonucleic acids
EDTA	ethylenediamine tetra-acetate
kb	kilobase pairs
kDa	kilodaltons
LB broth	Luria-Bertani broth
mRNA	messenger RNA
PAGE	polyacrylamide gel electrophoresis
PAS	periodic acid and Schiff's reagent
PCR	polymerase chain reaction
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acids
SDS	sodium dodecylsulphate
SSC	saline sodium citrate
SSPE	sodium chloride, sodium phosphate, EDTA
Tris	tris(hydoroxymethyl)amino ethane

Three-letter abbreviation of the common amino acids

One-letter code in parenthesis

Gly (G)	glycine	Ala (A)	alanine
Val (V)	valine	Leu (L)	leucine
Ile (I)	isoleucine	Met (M)	methionine
Phe (F)	phenylalanine	Pro (P)	proline
Ser (S)	serine	Thr (T)	threonine
Cys (C)	cysteine	Asn (N)	asparagine

Glu (Q)	glutamine	Tyr (Y)	tyrosine
Trp (W)	tryptophan	Asp (D)	aspartate
Glu (E)	glutamate	His (H)	histidine
Lys (K)	lysine	Arg (R)	arginine

1. Introduction

The study of evolution of organisms in the Earth's history necessarily depends upon fossils as primary data. A greater part of the fossil record is are derived from hard tissues of organisms that lived in the past. Palaeontologists have investigated the morphology and mineralogy of fossilized hard tissues, however, the evolutionary processes of those hard parts have not yet satisfactorily been explained. To solve this problem, it is needed to clarify the mechanisms of biomineralization in organisms living in the present.

Organisms are known to form some 60 different kinds of minerals in over 55 different phyla (Lowenstam and Weiner 1989). More than half of these minerals contain calcium, and among the most common are various calcium carbonates, constituting the exoskeletons and endoskeletons of invertebrates and calcium phosphates found in vertebrate bones and teeth (Lowenstam 1981).

Two fundamentally different processes of mineral formation can be distinguished. One basic process of mineral formation, exemplified by some bacterial species and various green and brown algae, is known as "biologically induced mineral-forming" process (Lowenstam 1981). This process is characterized by bulk extracellular and/or intercellular mineral formation without the elaboration of organic matrices.

The other process of mineral formation, observed, for example, in mollusks, vertebrates, and in many other animal phyla, is "organic matrix-mediated" biomineralization. Organisms construct an organic framework or mold composed mainly of proteins and/or polysaccharides, into which appropriate ions are introduced and then crystals are nucleated to grow. This "organic matrix-mediated" process is much more rigorously controlled than "biologically induced mineral-forming" process

(Lowenstam 1981, Weiner and Traub 1984).

Minerals produced by organisms which adopt the "organic matrix-mediated" biomineralization often have crystal shapes clearly different from those formed inorganically. These biominerals are a composite of inorganic crystals and organic matrix molecules. It is generally postulated that the elaborate fabrication of biominerals arises from specific molecular interactions at inorganic-organic interfaces (Mann et al. 1993; Mann 1996), and that the organic matrix represents many of the important molecules involved in the interactions controlling crystal growth (e.g., Watabe and Wilbur 1960; Loweam and Weiner 1989, Mann 1996).

Molluscan shells are composed of two or more shell layers, each of which has a distinctive microstructure, such as the nacreous, prismatic, and foliated structure. The shells are almost exclusively composed of calcium carbonate of either of the two polymorphs, namely calcite or aragonite. Considering the patterns in ontogeny and phylogeny, it is evident that formation of these microstructures and polymorphs are under genetic control.

Matrix molecules in calcium carbonate, especially of molluscan shells, have been studied to a considerable extent to unveil their roles in the mineralization processes. The matrix molecules have been classified conventionally into two classes based on the solubility in aqueous solutions. The soluble matrix macromolecules are acidic and usually represent a relatively minor portion of all the macromolecules that make up the organic matrix (Samata and Krampitz 1981, Samata 1988). The major components are the insoluble macromolecules such as chitin and silk fibroin-like proteins. The insoluble ones are relatively hydrophobic when compared to the acidic soluble ones, and those two classes of organic molecules can almost always be distinguished clearly from each other (Lowenstam and Weiner 1989).

The insoluble matrix is thought to be largely intercrystalline (Krampitz 1982) and make up the mold or framework in which mineralization occurs, while the soluble matrix is known as intracrystalline or located on the intercrystalline matrix surfaces, and play more versatile roles, but those functions are still poorly understood (Addadi and Weiner 1997). Therefore, knowledge of the functions of the matrix molecules, especially of the soluble ones, in mediating mineral deposition seems critically important to understand biomineralization.

Following functions have been advocated for the soluble matrix molecules of molluscan shells to date:

(1) Induction of oriented nucleation

It is often postulated that the organic matrix plays a role as a template on which epitaxial growth of mineral phase takes place. In this process, repeating amino acid sequences in soluble matrix proteins such as $(\text{Asp-X})_n$ or $(\text{Asp-Gly-X-Gly-X-Gly})_n$ have been thought as calcium-binding motifs (Weiner and Hood 1975; Weiner 1983; Runnegar 1984; Weiner and Traub 1984; Addadi and Weiner 1985, Addadi et al. 1987; Weiner and Addadi 1991).

(2) Inhibition of crystal growth

Acidic soluble proteins are thought to attach to the calcium ions on the surface of crystal lattices, thereby inhibiting further additions of ions and halting crystal growth (Weiner and Traub 1980; Wheeler et al. 1981; Wilbur and Bernhardt 1984; Addadi and Weiner 1985; Wheeler 1992).

(3) Control of aragonite-calcite polymorphism

Many organisms selectively form either calcite or aragonite, and a striking example of biological control is the ability to determine which polymorph of calcium carbonate will be precipitated at a given location (Lowenstam and Weiner 1989). What factor has the greatest influence on the polymorph formation of calcium carbonate is an old problem, and many explanations have hitherto proposed (Jamieson 1953; Kitano 1962a; Kitano 1962b; Kitano and Hood 1962; Kitano and Hood 1965; Kitano et al. 1969; Watabe and Wilbur 1960). Recently, convincing pieces of evidence have been demonstrated by *in vitro* experiments that soluble matrix proteins play an essential role in controlling polymorph occurrence (Falini et al. 1996, Belcher et al. 1996).

(4) Enhancement of mechanical properties of crystals

Sea urchin skeletal elements, protein-crystal composites, do not cleave like an inorganic calcite crystal, but break with chonchoidal fractures, which are usually accompanied by unusual strength of the materials (Raup 1962, Nissen 1969; Emlet 1982). It is postulated that the sea urchin skeletal elements are mechanically reinforced by proteins located within crystals (Berman et al. 1988; Berman et al. 1990). Also in molluscan shells, the organic matrix is hypothesized to reinforce shells mechanically (Curry 1990, Wheeler 1992, Heuer 1998).

(5) Catalysis of HCO_3^- formation

The mantles of shell-forming mollusks are known to contain carbonic anhydrase activity (Freeman and Wilbur 1948). Miyamoto et al. (1996) revealed that nacrein, a major soluble protein of the nacreous layer of a pearl oyster, has a domain similar to carbonic anhydrase and carbonic anhydrase activity. These results suggest that the carbonic anhydrase activities in molluscan mantles result from the soluble proteins themselves, which may also participate in the calcium carbonate formation.

Most of the evidence to support the above hypotheses on the functions of soluble proteins has been obtained through *in vitro* experiments. A weakness of these studies is that unpurified proteins or protein fractions of dubious homogeneity have been applied in the biochemical analyses and in the *in vitro* mineralization experiments, and that the stereochemical relationships between the organic and inorganic phases have been presumed without precise information of the fine structures of the proteins.

To have insight into the underlying mechanisms of the protein-mineral interactions, it seems essential to know the primary structures of the proteins involved; however, only a limited number of amino acid sequences have been determined so far for the calcium carbonate matrix proteins, and most of them are partial sequences.

Repeating sequence motifs such as the (Asp-X)$_n$ type repeating sequence (Weiner and Hood 1975; Weiner 1983; Runnegar 1984; Weiner and Traub 1984; Addadi and Weiner 1985, Addadi et al. 1987; Weiner and Addadi 1991), (Asp-Gly-X-Gly-X-Gly)$_n$ type (Runnegar 1984), (Asp)$_n$ type (Wheeler 1992), (Asp-Pro-Thr-Asp) type (Weiner 1983), and (Asp-Gly-Ser-Asp) and (Asp-Ser-Gly-Asp) types (Weiner 1981; Weiner 1983) have been predicted for molluscan soluble matrix proteins in connection with their functions, but all these hypotheses have been based on indirect observations.

Concerning insoluble matrix proteins in molluscan shells, amino acid sequences have been determined for three proteins, i. e., MSI60 in nacreous-aragonitic shell layer, MSI31 in prismatic-calcitic shell layer (Sudo et al. 1997), and Lustrin A in nacreous-aragonitic shell layer (Shen et al. 1997). But as to soluble matrix proteins, complete primary structure has been determined for only one protein, i. e., nacrein in nacreous-aragonitic shell layer (Miyamoto et al. 1996), and full amino acid sequences for proteins in calcitic layers have not been reported so far.

Here I present the full amino acid sequence for the molluscan shell protein MSP-1, a major soluble protein of the foliated calcite shell layer of the scallop *Patinopecten yessoensis*, and discuss its bearing on the functions of matrix proteins in calcium carbonate biomineralization. I also describe biochemical characters of the soluble shell proteins of some pectinid species and make comparisons among these shell proteins.

2. Materials and Methods

(1) Materials

Living specimens of a cold-water pectinid species *Patinopecten yessoensis* were purchased from a local dealer in Tsukiji, Tokyo. Air-dried shells of warm-water pectinids *Pecten albicans* and *Chlamys farreri* were originally collected alive near Mikawaissiki, Aichi, central Japan, and had been stored in University Museum, University of Tokyo.

(2) Isolation of soluble proteins from scallop shells

Both right and left shell valves of the specimens of the three species were thoroughly cleaned mechanically and incubated for 48 hr at room temperature in a 10 %(v/v) solution of sodium hypochlorite to destroy surface contaminants. After thorough washing with ultrapure water, the marginal portion of the shell, consisting only of the outer shell layer of foliated calcite, was crushed to fine fragments. The matrix proteins were extracted by dissolution of the shell flakes (100 g) in 3 liters of 0.5 M EDTA, pH 8. The extraction was performed at 4 °C with continuous stirring for 72 hr. The preparation was then filtered through cheesecloth to remove viscous insoluble materials, and desalted by ultrafiltration using the Minitan tangential flow system (Millipore). In this procedure, the concentration of EDTA was reduced to less than 10^{-6} molar, at which point the sample was concentrated to about 50 ml, then lyophilized.

(3) SDS-PAGE analyses

i) SDS-PAGE

The extracted macromolecules (1 mg per each well) were separated by SDS-PAGE (Laemmli 1970), using slab gels (15 cm x 15 cm) of 2 mm thickness containing 10 %(w/v) polyacrylamide, at 25 mA, 4 °C, for 6 hours.

ii) Coomassie Brilliant Blue staining

After electrophoresis, the gel was stained in a solution of 0.25 %(w/v) Coomassie Brilliant Blue R, 25 %(v/v) ethanol, and 10 %(v/v) acetic acid for at least 3 hours followed by continuous washing with a solution of 25 %(v/v) ethanol and 10 %(v/v) acetic acid until protein bands become visualized.

iii) Staining with Periodic acid and Schiff's reagent

After electrophoresis, the gel was washed with 7.5 %(v/v) acetic acid seven times with each interval separated by a 30 minutes soaking to fix the glycoproteins and remove sucrose. After oxidation for 1 hour with 0.2 %(w/v) periodic acid in the same solvent, the gel was washed for seven times as before and then stained with Schiff's reagent for 8 hours (Holden et al. 1971). The gel was soaked for 15 minutes in a solution of 0.5 %(w/v) sodium pyrosulfite, with the solution refreshed twice after five and ten minutes, to visualize protein bands, followed by washing with deionized water.

iv) Silver staining

After electrophoresis, the gel was fixed for 1 hour in a solution containing 40 %(v/v) methanol and 10 %(v/v) acetic acid, then fixed for 2 hours in a solution of 10 %(v/v) ethanol and 5 %(v/v) acetic acid, with the solution refreshed once after 1 hour. After oxidation for 10 minutes with 10 %(v/v) Oxidizer (Silver Stain kit; Bio-Rad), the gel was washed for 10 minutes with deionized water more than 7 times until the yellow color is completely removed from the gel. After staining for 30 minutes with 10 %(v/v) Silver reagent (Silver Stain kit; Bio-Rad), the gel was washed for two minutes with deionized water. The gel was developed for 15 minutes in a solution containing 3.2 %(w/v) Developer (Silver Stain kit; Bio-Rad), with the solution refreshed twice after five and ten minutes, followed by soaking for 5 minutes with 5 %(v/v) acetic acid to stop the reaction.

v) Stains-All staining

After electrophoresis, the gel was washed and fixed by soaking the gel for 30 minutes in a solution of 25 %(v/v) methanol six times with the solution refreshed in each time. After washing and fixation, the gel was stained with a solution containing 0.0025 %(w/v) Stains-All (SIGMA), 25 %(v/v) methanol, 7.5 %(v/v) formamide, and 30 mM Tris-HCl (pH 8.8) in dark for 3 hours (Campbell et al. 1983).

(4) N-terminal sequence determination

Following separation by SDS-PAGE, the proteins were electroblotted onto polyvinylidene difluoride membrane (ProBlott; Applied Biosystems) in Caps buffer (10 mM, pH 11) containing 10 %(v/v) methanol, prior to staining with Coomassie Brilliant Blue R. N-terminal amino acid sequence analysis of the immobilized protein samples was by Edman degradation using an automated protein sequencer (Perkin-Elmer

Applied Biosystems). Sequences were determined at least twice for each protein band reproduced by different SDS-PAGE gels.

(5) RNA purification and cDNA synthesis

Total RNA from the mantle tissue of a single specimen of *P. yessoensis* was extracted using ISOGEN (Nippon Gene) and the single-step method for RNA isolation (Chomzynski 1993). The RNA (5 μg) was applied as template for reverse transcription to prepare single-strand cDNA in a 12 μl reaction, primed with a "hybrid" primer, TCGAATTCGGATCCGAGCTC(T)$_{17}$, using the SuperScript preamplification system (Life Technologies).

Alternatively, to obtain double-strand cDNA, the Marathon cDNA amplification kit (Clontech) was used. The RNA (4 μg) was applied as template for reverse transcription to prepare first-strand cDNA in a 10 μl reaction, primed with a "Marathon cDNA synthesis primer", TTCTAGAATTCAGCGGCCGC(T)$_{30}$XN (V = G, A, or C; N = G, A, C, or T). The resulting first-strand cDNA (10 μl) was used in a 80 μl second-strand cDNA synthetic reaction catalyzed by DNA polymerase I. The preparation (10 μl) was then subjected to blunt-ending of the double-strand cDNA catalyzed by T4 DNA Polymerase in a 82 ul reaction, using the Marathon cDNA Amplification Kit (CLONTECH). Finally, the cDNA (5 μl) was ligated, with T4 DNA Ligase, to the Marathon cDNA Adaptor in a 10 μl reaction using the same kit.

(6) PCR amplifications

i) 3' RACE using degenerate primers

Two sense primers, P1 and P2 (Table 1) corresponded, respectively, to the LDTDKD and LEFHLD (for one-letter abbrevations of amino acids, see p. 8), parts of the N-terminal sequence determined by Edman degradation. Both primers (P1 and P2) are degenerate, containing all the possible oligonucleotide sequences for each amino acid sequence. The antisense "adaptor" primer (PA), which comprises of the same sequence as the 5' half of the "hybrid" primer, TCGAATTCGGATCCGAGCTC, was also synthesized for the PCR amplification of the region between the point corresponding to the N-terminal end of the mature protein (MSP-1) and the 3'-end of the transcript (3' RACE: rapid amplification of cDNA ends protocol; Frohman 1990). The 10 μl of single-strand cDNA was diluted to 100 μl with H_2O, and 1 μl was used in the PCR reaction which also included 10-100 pM of each primer (P1 and PA), 1 x Ex Taq DNA polymerase buffer (TAKARA), 100 mM dNTP, and 1 unit of Ex Taq DNA polymerase (TAKARA). A Cetus DNA Thermal Cycler (Perkin Elmer) was employed with an initial step at 94 °C for 3 min, then 30 cycles at 94 °C for 30 sec, 46-52°C for 30 sec, 72 °C for 2 min, followed by a final extension step of 72 °C for 5 min. A second round of PCR reaction was performed with the P2-PA primer pair using the PCR products (1 μl) of the previous round of reactions as template, in order to verify specific amplification of the target cDNA fragment. PCR reactions with only one degenerate primer (P1 or P2) were performed in parallel as negative controls.

ii) 3' RACE using specific primers

In order to design unique primers to facilitate more reliable amplification of the DNA sequence encoding the entire protein, attempts were made to design specific, rather than degenerate, primers for the 3'

RACE.

The target cDNA sequence, encoding the N-terminal sequence of 20 amino acid residues determined as mentioned above, was amplified by a method known as reverse transcription-polymerase chain reaction (RT-PCR). The degenerate antisense primer (P3; Table 1) was designed to correspond to the sequence encoding NAAED, the last part of the N-terminal peptide sequence of 20 amino acids. The sense primer (P1) and the antisense primer (P3) were used (50 pM of each) in the PCR reaction which also included 1 μl of the diluted single-strand cDNA, 1 x Taq DNA polymerase buffer (Life Technologies), 3 mM $MgCl_2$, 100 mM dNTP, and 1 unit of Taq DNA polymerase (TOYOBO). A Cetus DNA Thermal Cycler was employed with an initial step at 94 °C for 3 min, then 30 cycles at 94 °C for 30 sec, 52 °C for 30 sec, 72 °C for 1 min, followed by a final extension step of 72 °C for 5 min. PCR reactions with only one degenerate primer (P1 or P3) were incubated in parallel as negative controls. The resulting PCR products of 60 base pairs (bp) in length (corresponding to 20 amino acids), were sequenced directly by the chain termination method using BigDye Terminator Cycle Sequencing Kit and an automated DNA sequencer (Perkin-Elmer Applied Biosystems) as described later.

Three gene-specific sense primers, P4, P5, and P6 (Table 1), designed based on the sequence determined by the above method for the PCR amplification of the region between the point corresponding to the N-terminal end of MSP-1 and the 3'-end of the transcript. The PCR amplification using the primer pair of P4-PA (10 pM of each) and the diluted single-strand cDNA (1 μl) as template, catalyzed by Ex Taq, was performed first, under the condition including an initial step at 94 °C for 2 min, then 35 cycles at 94 °C for 30 sec, 62 °C for 30 sec, 72 °C for 2 min, followed by a final extension step of 72 °C for 5 min. The first PCR amplifications using the primer pairs of P5-PA and P6-PA were also performed under the same conditions as the PCR amplification using the primer pair of P4-PA. A second and a third round of PCR reactions were performed with the P5-PA and P6-PA primer pairs, respectively, using the PCR products (1 μl) of the previous round of reactions as template. PCR reactions with only one primer (P4, P5, P6, or PA) were included as negative controls.

The PCR amplifications using double-strand cDNA ligated to the Marathon cDNA Adaptor, rather than single-strand cDNA, as template were also performed. The primer P4, P5, or P6 (Table 1) was utilized as the sense primer, and Marathon cDNA Adaptor Primer 1 (AP1, Table 1) or Marathon cDNA Adaptor Primer 2 (AP2, Table 1) was utilized as the anti-sense primer. The 2 μl of double-strand cDNA was diluted to 100 μl with Tricine-EDTA buffer (10 mM Tricine-KOH, pH 8.5, 0.1 mM EDTA), and 5 μl was used in the PCR reactions. The PCR amplification, catalyzed by Ex Taq, was performed under the condition including an initial step at 94 °C for 1 min, then 5 cycles at 94 °C for 30 sec, 72 °C for 4 min, 5 cycles at 94 °C for 30 sec, 70 °C for 4 min, and 25 cycles at 94 °C for 30 sec, 68 °C for 4 min. A second round of nested PCR reaction was performed with the sense primer of P5 or P6 and anti-sense primer of AP2 using the PCR products of the previous round as template. Reactions were performed under the same condition as the first round of PCR reactions. Reactions with only one primer (P4, P5, P6, AP1, or AP2) were included as negative controls.

In order to perform even more reliable amplification of the DNA sequence encoding the entire protein, another antisense primer, "adaptor-T" primer (PT), TCGAATTCGGATCCGAGCTCTT, which has two extra Ts following the "adaptor" primer sequence, was synthesized for the PCR amplification of the same region. The PCR amplification using the primer pair of P4-PT (10 pM of each) and the single-strand cDNA as template, catalyzed by Ex Taq, was performed first, under the condition including an initial step at 94 °C for 2 min, then 35 cycles at 94 °C for 30 sec, 62 °C for 30 sec, 72 °C for 2 min, followed by a final extension step of 72 °C for 5 min. A second and a third round of PCR reactions were performed with the P5-PT and P6-PT primer pairs, respectively, using the PCR products (1 μl) of the previous round of reactions as template, in order to verify specific amplification of the target cDNA fragment. Reactions with only one primer (P4, P5, P6, or PT) were included as negative controls.

iii) 5' RACE

The diluted double-strand cDNA (10 μl), ligated to the Marathon cDNA Adaptor, and the primer pair of P3 and AP1 (Table 1) was utilized for the PCR amplification of the region between the point corresponding to the second codon position of the 20th amino acid residue from the N-terminal end and the 5'-end of the transcript (5' RACE; Frohman 1990). A second round of PCR reaction was performed with the primer pair of P3 and AP2 (Table 1) using the PCR products of the previous round as template. PCR reactions were performed under the condition including an initial step at 94 °C for 2 min, then 30 cycles at 94 °C for 30 sec, 52 °C for 30 sec, 72 °C for 1 min, followed by a final extension step of 72 °C for 5 min. PCR reactions with only one primer (P3, AP1, or AP2) were included as negative controls.

The amplified 5' RACE products were subcloned, screened, and sequenced as described later.

iv) Amplification of full-length cDNA

The sense primer, P7, and the antisense primer, P8, were designed based on the nucleotide sequences of the 5'- and 3'-termini of the cDNA, respectively. Using this primer pair, PCR reactions were performed for the full-length cDNA amplification using the diluted first-strand cDNA reaction as template under the condition including an initial step at 94 °C for 2 min, then 35 cycles at 94 °C for 30 sec, 62 °C for 30 sec, 72 °C for 2 min, followed by a final extension step of 72 °C for 5 min. PCR reactions with only one primer (P7, or P8) were included as negative controls.

(7) Subcloning of cDNA and screening

PCR products of cDNA 5'-end (2 μl) or the unique 2.9 kb total length cDNA (2 μl) were used in the ligation reaction (10 μl) which also included 50 ng of pGEM-T vector (Promega) , 1 x T4 DNA ligase buffer (Promega), and 3 units of T4 DNA ligase (Promega). The reaction was performed overnight at 4 °C.

The resulting ligation products were used for transformation of the *Escherichia. coli* NM522 cells, which were made competent by the calcium chloride method as described by Sambrook et al. (1989). A mixture of 10 μl of the ligation products and 100 μl of the competent cells was placed on ice for 30 minutes followed by heat shocking of the cells exactly for 2 minutes at 42 °C. LB broth (1 ml) containing the cells transformed with the ligation products was incubated for 1 hour at 37 °C. The transformed cells pelleted by centrifugation were diluted to 100 μl with LB broth, and 50 μl of the preparation was plated onto antibiotic (ampicillin) plates. The plated cells (NM522) were compatible with blue/white color screening and standard ampicillin selection.

Verification of positive clones was carried out by PCR amplifications. Each white colony was picked up and suspended in 10 μl of autoclaved distilled water. The suspension was used for a PCR reaction primed with M13 forward and reverse primers, catalyzed by Taq DNA polymerase, under the condition including an initial step at 94 °C for 2 min, then 30 cycles at 94 °C for 30 sec, 50 °C for 30 sec, 72 °C for 2 min. PCR products were electrophoresed using 1.5 %(w/v) agarose gels, and visualized by ethidium bromide staining to select clones with inserts of expected sizes.

(8) Small-scale preparation of plasmid DNA

Single bacterial colonies were transferred into 5 ml of LB medium and incubated overnight at 37 °C in a rotary shaking incubator. The bacterial cells collected by centrifugation at 15000 rpm for 5 minutes at 4 °C were resuspended in 0.5 ml of STE buffer (0.1 M NaCl, 10 mM Tris-HCl [pH 8.0], and 1 mM EDTA [pH 8.0]). The cells were pelleted once again by centrifugation, and were resuspended in 100 μl of ice-cold Solution I (50 mM glucose, 25 mM Tris-HCl [pH 8.0], and 10 mM EDTA [pH 8.0]). The preparations were vigorously vortexed, and 200 μl of freshly prepared Solution II (0.2 N NaOH and 1%(w/v) SDS) were added. After mixing and incubation on ice for 5 minutes, 150 μl of ice-cold Solution III (3 M potassium, and 5 M acetate) were added, mixed gently and stored on ice for 5 minutes. After centrifugation at 15000 rpm for 5 minutes at 4 °C, the supernatant was transferred to a new tube, and plasmid DNA was purified by phenol/chroloform extraction and ethanol precipitation.

(9) DNA sequencing and analysis

The PCR products or the inserts in plasmids were sequenced by Sanger method using dye-labeled terminators (Perkin-Elmer Applied

Biosystems Inc.). Sequencing of the PCR products were primed with 50 pmol of one of the same primers as used for the PCR reactions, and sequencing of the inserts were primed with 3.2 pmol of the M13 forward, M13 reverse, or an internal primer (P9 - P19, Table 1). The PCR products (60 ng) or the inserts (300 ng) were applied as template in 20 μl of sequencing reactions, each of which also included the primer and 8 μl of BigDye Terminator RR Mix (Perkin-Elmer Applied Biosystems Inc.). A Cetus DNA Thermal Cycler was employed with 30 cycles at 96 °C for 30 sec, 50 °C for 15 sec, 60 °C for 4 min. The sequencing reaction products pelleted by ethanol precipitation were diluted to 4 μl with loading buffer (83 %(v/v) formamide, 4.2 mM EDTA, and 0.83 %(w/v) blue dextran (SIGMA)), heat-shocked at 90 °C for 2 minutes, and 1.8 μl of the preparation were analyzed by a Model 377 DNA sequencer (Perkin-Elmer Applied Biosystems Inc.).

Resulting sequence data were analyzed with the software GENETYX-MAC ver. 9 (Software Development). The analyses of secondary structure were carried out using the method of Chou and Fasman (1978). Homology searches of both nucleotides and proteins were carried out with FASTA and BLAST using DDBJ data bank (National Institute of Genetics, Mishima, Japan).

(10) Northern blot analysis

Poly $(A)^+$ RNA was prepared from the total RNA extracted from mantle tissue using the Oligotex-dT30-Super (Japan Synthesis Rubber). A sample of 1.8 μg of the isolated poly(A^+) RNA was electrophoresed on a 1%(w/v) formaldehyde-agarose gel, electrically transferred to NYTRAN membrane (Schleicher & Schuell), and UV-cross-linked to the membrane. The membrane was prehybridized at 42 °C for one hour in a solution containing 0.2%(w/v) blocking reagent (Böhringer Manheim), 50%(v/v) formamide, 5 x SSPE, 1%(w/v) SDS, and 100 μg/ml denatured salmon sperm DNA, and hybridized at 42 °C overnight in the same buffer, to which a probe was added. The probe was a cDNA fragment (corresponding to nucleotide position, 1-845 in Figure 6) that was radio-labeled with [α-^{32}P] dATP by random prime labeling method using Megaprime DNA labelling system (Amersham), and the concentration of the probe in hybridization buffer was 5 x 10^6 cpm/ml. The membrane was washed at 60 °C with three changes of 0.1 x SSPE and 0.1%(w/v) SDS for 10 minutes each. Hybridyzation signals were detected by autoradiography.

3. Results

(1) Biochemical properties of organic matrix molecules

Organic molecules were extracted from shells of three pectinid species, *P. yessoensis*, *P. albicans*, and *C. farreri*. Each extraction procedure yielded about 1 mg of soluble total organic molecules per three grams of shell flakes. SDS-PAGE of the organic molecules of *P. yessoensis* revealed eight protein bands of 150, 125, 104, 77, 49, 42, 40, and 15 kilodaltons (kDa) in apparent molecular mass, when stained with Coomassie Brilliant Blue (Figure 1, Lane A). This pattern was replicated when stained with silver except for the band of 15 kDa which was not stained with silver (Figure 1, Lane C). All these protein bands (including the 15 kDa one) were stained red by the PAS staining (Figure 1, Lane B), indicating that all these components are glycosylated. Stains-All staining stained major four protein bands of 150, 125, 104, and 77 kDa blue (Figure 1, Lane D), indicating that the four components may have cation-binding potential. PAGE under non-denaturing conditions revealed a similar gel pattern as in SDS-PAGE, confirming that these proteins are highly acidic (data not shown).

SDS-PAGE of the organic molecules of *P. albicans* revealed four protein bands of 77, 49, 42, and 18 kDa in apparent molecular mass, when stained with Coomassie Brilliant Blue (Figure 2). This pattern was replicated when stained with silver, except that a protein band of 36 kDa was newly obserbed when stained with silver (Figure 2, Lane C). The three major protein bands of 77, 49, and 42 kDa were stained red by the PAS staining (Figure 2, Lane B), indicating that these components are glycosylated.

SDS-PAGE of the organic molecules of *C. farreri* revealed three protein bands of 65, 30, and 17 kDa in apparent molecular weight, when stained with Coomassie Brilliant Blue (Figure 3, Lane A) and no band was observed when stained with the PAS method (data not shown).

(2) N-terminal sequences of organic matrix molecules

Three components of *P. yessoensis* of 104, 77, and 49 kDa were subjected to N-terminal sequencing analysis. The 104 and 77 kDa components had the same amino acid sequence of LDTDK DLEFH LDSLL NAA (in one-letter abbrevation of amino acids). The sequence for the 49 kDa component was LDTDK DLEFH LDSLL NAAED. Thus N-terminal sequencing of the three components revealed that all three share the same amino acid sequence at least for the first eighteen residues.

Three major components of *P. albicans* of 77, 49, and 42 kDa were subjected to N-terminal sequencing analysis, which also revealed that all three components share the same amino acid sequence, SDTDA DTDED EEN, at least for the first thirteen residues sequenced.

Analysis for the 17 kDa component of *C. farreri* indicated the following provisional N-terminal sequence for this component: L(orD)DP(orD)DP(orD) DDD.

(3) Nucleotide sequence and deduced amino acid sequence of MSP-1

i) 5' RACE

Five prime RACE using the primer pair of P3 and AP1 resulted in amplification of 0.15 kb DNA fragments, which were also observed as a single band in the second PCR using the primer pair of P3 and AP2.

Direct sequencing of these products revealed that they represent a sequence containing a region for the N-terminal sequence determined by Edman degradation, verifying that the target cDNA was successfully amplified.

ii) 3' RACE

Table 2 summerizes the results of 3' RACE experiments attempted using various primer pairs and two different methods of cDNA preparations.

a) 3' RACE using degenerate primers

Three-prime RACE was performed using the single-strand cDNA and primer pairs P1-PA and P2-PA several times under different PCR conditions (amount of templates and primers, annealing temperatures, and numbers of PCR cycles) but target cDNA was not amplified (Table 2).

b) 3' RACE using specific primers

Three-prime RACE was performed several times using the single-strand cDNA and primer pairs, P4-PA, P5-PA, and P6-PA. Though a part of target cDNA of 2.1 kb was amplified, the amplified cDNA was not the authentic full length cDNA because sequence determinations indicated that it lacks a poly(A) tail (Table 2).

Three-prime RACE was also performed several times using the double-strand cDNA as the template with the Marathon cDNA Amplification kit (CLONTECH), primed with P4-AP1, P4-AP2, P5-AP1, P5-AP2, P6-AP1, and P6-AP2, but the target cDNA was not amplified (Table 2).

However, 3' RACE using the single-strand cDNA and the primer pairs P4-PT, P5-Pt, and P6-PT yielded two bands of amplified cDNA (2.8 kb and 3.3 kb, Table 2).

iii) Amplification and characterization of the full length cDNA for MSP-1

To obtain a full length cDNA, the primer P7 was designed based on the determined 5'-end sequence of the cDNA (Table 1). Using the primer pair of P7-PT, two components of cDNA of *P. yessoensis* (2.9 kb and 3.4 kb) were amplified (Figure 4).

The 2.9 kb product was subcloned and screened by blue/white color selection. From a total of 150 colonies, 3 positive clones were obtained and the cDNA insert of 2.9 kb was sequenced for each of these clones.

The antisense primer, P8, was designed based on the determined 3'-end sequence of cDNA of 2.9 kb. Using the primer pair of P7 and P8, only one component (2.9 kb) was amplified (Figure 4). Northern blot analysis using a 5' portion of the amplified and subcloned 2.9 kb cDNA as the probe revealed that the 2.9 kb mRNA is expressed in the mantle tissues of *P. yessoensis* (Figure 5).

Figure 6 shows the complete nucleotide sequence for the 2978 bp cDNA encoding MSP-1 (molluscan shell protein 1; Sarashina and Endo 1998). The nucleotide sequence revealed an open reading frame of 820 amino acids with the translation initiation codon ATG at the nucleotide position 35. The 20 amino acids upstream of the cleavage site apparently comprise a signal peptide, which has a hydrophobic core of 13 residues, and the first 20 residues downstream of the cleavage site agreed with the chemically determined N-terminal sequence of the protein extracted from the shells. An in-frame stop codon is located at the nucleotide position

2555, with a putative polyadenylation signal ATTAAA located 370 nucleotides downstream from the stop codon and 20 nucleotides upstream from the poly(A) tail. It is thus very likely that this cDNA represents the full length copy of the MSP-1 mRNA coding region.

The deduced amino acid sequence contains a high proportion of serine, glycine, and aspartate residues (Table 3), in agreement with the bulk composition of the soluble matrix of the same species (Kasai and Ohta 1981), supporting the fact that MSP-1 represents the major component of the soluble matrix. The amino acid composition is also comparable with that of a purified soluble fraction of oyster shell (Wheeler 1992; Table 3), and is in fact typical for a mollusc soluble matrix fraction (Lowenstam and Weiner 1989). The isoelectric point of the deduced amino acid sequence of MSP-1 is 3.2, as expected for this Asp-rich protein. The calculated mass for MSP-1 assuming no post-translational modifications is 74.5 kDa.

(4) Modular structure of MSP-1

The deduced amino acid sequence of MSP-1 revealed that it has a modular structure, and highly conserved units being repeated in tandem four times, in the central part (Figures 7, 8).

The N-terminal domain of MSP-1 is 42 residues long, and an α-helix was predicted for this domain by the analyses of secondary structure using the method of Chou and Fasman (1978). The N-terminal domain is followed by the basic domain, a basic and hydrophilic domain (Figure 9), containing five Lys-Gly-X-Y (X = Gly or Ser, Y = Ser or Asn) segments in tandem, intercalated by a Thr-Arg-Ser-Ser segment in the middle. Following the basic domain is a 27 residue serine and glycine-rich region ("SGD" domain) which is acidic due to the presence of seven aspartate and two glutamate residues.

Following the SGD domain, four highly conserved units are arranged in tandem. Each unit contains three domains, designated here as the "SG", "D", and "G" domains.

The SG domains are 29 - 43 residues long, almost solely composed of serine and glycine, dominated by $(\text{Ser-Gly})_n$ and $(\text{Ser})_n$ repeats.

Each SG domain is followed by the D domain, the largest and seemingly the most important of the revealed domains of MSP-1. All four D domains are hydrophilic (Figure 9) and comprise 95 amino acids, containing 34-35 aspartate residues, and share a high degree of sequence similarity with each other (Figure 10). The D domains contain some regular motifs such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp). Soluble proteins in foliated calcite shell layer of mollusks are phosphorylated (Borbas et al. 1991). Thus it is likely that MSP-1 is also phosphorylated, since it has been extracted from the foliate calcite layer of *P. yessoensis*. A total of 9 to 10 potential phosphorylation sites was found in each D domain (Figure 10), and a total of 3 to 5 potential *N*-glycosylation sites was also found in each D domain (Figure 10).

A glycine-rich domain (G domain), which is 34 - 39 residues long, follows the D domain. Each G domain has a basic core in the central part, such as Lys-X-Arg-Lys-X-X-Lys-Arg, Lys-X-Arg-Lys-X-X-Lys-X-X-Arg, or Lys-X-Arg-Lys, where X is Gly. Lys and Arg are basic amino acid residues (Figure 11). The four G domains also exhibit a high degree of sequence similarity with each other (Figure 11).

Following the four highly conserved units is the C-terminal domain, being 50 residues long, and acidic due to the presence of eight aspartate and five glutamate residues.

4. Discussion

(1) Identity of the protein MSP-1

SDS-PAGE of the organic matrix molecules for *P. yessoensis* revealed eight species of proteins. The three proteins of 104, 77, and 49 kDa of *P. yessoensis* the same N-terminal amino acid sequence with each other. Two other proteins of 150 and 125 kDa in size share similar biochemical characters with the 104 and 77 kDa proteins, such as glycosylation and cation-binding potential. The 49 kDa protein is glycosylated, but has no evidence for cation-binding potential. However, the band for the 49 kDa component is faint to some extent when stained with Coomassie Brilliant Blue, and, therefore, the lack of signals for the 49 kDa band in Stains-all staining, may not necessarily mean that it does not have the cation-binding potential. In any case, although the N-terminal amino acid sequences of the larger two proteins (150 and 125 kDa) are unknown, it appears possible that these two proteins share the same N-terminal amino acid sequence with the 104, 77, and 49 kDa proteins.

Since MSP-1 is glycosylated, and quite possibly phosphorylated, the molecular weight of MSP-1 may be greater than that calculated from the amino acid sequence deduced from the cDNA. Furthermore, MSP-1 is highly acidic, and in such a case the molecular weight may be overestimated by SDS-PAGE.

The molecular mass of MSP-1 calculated from cDNA is 74.5 kDa, and there are four proteins in the SDS-PAGE gel (150, 125, 104, and 77 kDa) that have a greater size than 74.5 kDa. Therefore, these four proteins are considered as candidates for MSP-1.

In PCR reactions for the full length cDNA amplification, two bands (2.9 kb and 3.4 kb) were amplified for *P. yessoensis*, and northern blot analysis detected a band for the 2.9 kb mRNA. The detected signal in the northern blot analysis was broad to some extent, and it is possible that it also included the signal for the 3.4 kb mRNA. Further experiments are required to determine whether there exist more than one transcript encoding for MSP-1-related sequences.

Assuming that there are at most two species of transcripts encoding for MSP-1-related sequences, at least two of the three proteins that share the same N-terminal sequence (104, 77, and 49 kDa proteins) should have been derived from the same transcript, but were subjected to different degrees of post-translational processing (such as glycosylation, phosphorylation, and C-terminal cleavage).

The protein derived from the full length mRNA of 2.9 kb is defined as MSP-1, but it still remains unclear which mature protein appeared in the SDS-PAGE gel represents MSP-1.

(2) Structural characteristics of MSP-1

The characteristic feature of MSP-1 is its modular structure, consisting of the N-terminal domain, the basic domain, the SGD domain, four highly conserved units, and the C-terminal domain.

The N-terminal domain of MSP-1 is very acidic and contains an (Asp-X)$_3$ motif, which is conserved in the three pectinid species studied (Figure 12). This repeating sequence is reminiscent of the (Asp-Y)$_n$ type motif, that has been considered to bind calcium ions, functioning as a template for mineralization (Weiner and Hood 1975). Therefore this regular arrangement of acidic amino acid residues in the N-terminal domain could be functionally important.

The basic domain of MSP-1 contains five Lys-Gly-X-Y segments, producing a more or less regular arrangement of basic residues in the basic domain. It is reasonable to assume that the positively charged basic amino acid residues in the matrix proteins interact with the carbonate ions, which participate directly in the formation of calcium carbonate crystal lattices, but no evidence has been reported for involvement of basic residues in this processes.

Another possibility is that the basic domain may interact with negatively charged amino acid residues in other acidic proteins. The basic domain also has potential to participate in protein-protein cross-linking by virtue of its high lysine content. The lysine side chains can be modified to participate in intermolecular cross-linking via Schiff's base conjugate (Gordon and Carriker 1980). Lustrin A, an insoluble matrix protein of gastropod nacre (Shen et al. 1997) also has a basic domain. While only speculative, the basic domain of Lustrin A was expected to have anchoring sites for acidic proteins, because of its high lysine, tyrosine, and glutamine content (Shen et al. 1997). Those three amino acid residues can participate in protein-protein cross-linking (Gordon and Carriker 1980). Lustrin A has a modular structure, and nine Cys-rich domains and eight Pro-rich domains are arranged alternately in tandem in the central part of Lustrin A. The basic domain is located near the end (C-terminus) of Lustrin A. Concerning MSP-1, four highly conserved units are arranged in tandem in the central part of MSP-1, and the basic domain is located near the end (N-terminus) of MSP-1. Though the amino acid sequences of the two basic domains are not alike, the locations of the two basic domains in the overall domain arrangements are somewhat similar between MSP-1 and Lustrin A, suggesting a functional similarity.

The SGD domain of MSP-1 is as acidic as the N-terminal domain and the D domains. The D domains contain some regular motifs such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp), but the N-terminal domain and the SGD domain do not have these motifs. Therefore, the functions of the latter two acidic domains could well be different from those of the D domain.

Following the SGD domain, four highly conserved units are arranged in tandem. The high degree of sequence similarity among them suggests that they are likely to take a similar folding structure with each other. The four times repeating structure enables MSP-1 molecules to extend on the crystal surface to cover the surface with only a small number of the molecules, resulting in high functional effects of this protein on crystal growth, such as nuleator and/or inhibitor. Each unit contains the SG domain, the D domain, and the G domain.

The method of Chou and Fasman (1978) indicate that the SG domains have no α-helix nor ß-sheet structures but are rich in turns. The domains showed the highest sequence similarity with the "GS" domain of Lustrin A (Shen et al. 1997). The SG domains in MSP-1 may have a high degree of flexibility, as suggested for the GS domain of Lustrin A. Because the SG domain is located between the D and G domains in each unit, one function of SG domains may be to serve as spacers separating the D and G domains so that they can fold independently.

The D domain is highly acidic and the largest domain in MSP-1. It has often postulated that Asp-rich domains in shell proteins plays a role as a template on which epitaxial growth of mineral phase takes place, but the definite amino acid sequence of the Asp-rich domains has not been reported. The D domain of MSP-1 is the first to be characterized as a typical Asp-rich domain in shell proteins.

It is reasonable to assume that the negatively charged acidic amino acid residues in the matrix proteins interact with the calcium ions, somehow providing specific templates for the nucleation of biominerals (Hare, 1963). In addition, the charged amino acids in these proteins can also interact with crystal planes and thus control growth.

Weiner and Hood (1975) proposed a hypothesis that the $(\text{Asp-Y})_n$ type sequence in the soluble proteins from molluscan shells binds calcium ions, and that this repeating sequence may function as a template for mineralization. Most of the Ca^{2+} - Ca^{2+} distances in the crystal lattices of aragonite and calcite range from about 3.0 to 6.5 Å. Infrared spectroscopy indicated that a part of the proteins in molluscan shells is in the ß-sheet conformation (Hotta 1969). When the $(\text{Asp-Y})_n$ type sequence adopts a ß-sheet conformation, the distance from one aspartic acid residue to the next is 6.95 Å. Therefore, this repeating sequence may match the crystalline lattice and function as a template for mineralization (Weiner and Hood 1975; Weiner 1983; Weiner and Traub 1984; Weiner and Addadi 1991).

This model was further extended to explain the crystallography of the three types of foliated calcite by a precise stereochemical fit between the amino acid residues of the matrix protein and the crystal lattices in the surface of a specific crystal face (Runnegar 1984). The spacing of calcium ions along the length of the laths is 19.3 Å in one type of foliated calcite, a distance which is matched by every sixth residue of a parallel ß-sheet (19.44 Å). Taking account of the amino acid composition, the protein was predicted to have the repetitive sequence of $(\text{Asp-Gly-X-Gly-X-Gly})_n$. The other two types of foliated calcite were specified by the repetitive sequence of $(\text{Asp-Y})_n$ of either antiparallel or parallel ß-sheet conformation (Runnegar 1984).

However, these hypotheses of the primary and secondary structures of the acidic matrix proteins have been based on indirect observations. The $(\text{Asp-Y})_n$ domains are absent in a phosphoprotein of oyster shell (Wheeler 1992). Nacrein, a major soluble protein of pearl nacre (Miyamoto et al. 1996), does not contain a typical acidic domain. The results of this study indicate that $(\text{Asp-Y})_n$ type and $(\text{Asp-Gly-X-Gly-X-Gly})_n$ type domains do not exist in the primary sequence of MSP-1 except for the single (Asp-Y)3 sequence in the N-terminal domain. MSP-1 has Asp-rich domains (D domains), which contain some regular motifs such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp). These motifs could represent calcium-binding sites. But the overall arrangement of the acidic residues in the domains suggests that one-dimensional simple models may be insufficient to explain the interactive relationships at the protein-mineral interfaces. If these motifs interact with Ca^{2+} on an already-formed calcite crystal, inhibition of crystal growth may occur, whereas if these motifs interact with Ca^{2+} in solution, nucleation may occur.

The G domains in MSP-1 have no apparent calcium binding motifs. Each G domain has a basic core which could interact with aother acidic protein. Addadi and Weiner (1985) proposed a hypothesis that acidic soluble proteins are nucleators of calcite when adsorbed on a rigid substrate. If the hypothesis is accepted, it is reasonable to assume that positively charged basic cores in the G domains in MSP-1 interact with acidic insoluble proteins, and the negatively charged D domains in MSP-1 act as nucleators of calcite. On the other hand, it is also possible to expect that positively charged basic cores in the G domains of one MSP-1 molecule interact with the negatively charged D domains of another MSP-1 molecule. The arrangements of basic residues in the G domains are similar to those of acidic residues in the D domains. For example, a partial amino acid sequence (580-587) of the third G domain is **K**G**RK**GG**KR** (Basic residues are in bold), and a reversed sequence (483-476) of the third D domain is **D**G**DD**SG**DD** (acidic residues are in bold). The number of the G domains (4) corresponds to that of the D domains. The similarity of the arrangements of basic and acidic residues in the G and D domains, and the correspondence of the number of the G and D domains in MSP-1 somewhat support this interpretation.

The C-terminal domain of MSP-1 is very acidic with both Asp and

Glu residues distributed somewhat irregularly, and does not contain the regular motifs found in the D domain, such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp). In this regard, the C-terminal domain is similar to the N-terminal and SGD domains, and they could be functionally related.

(3) Comparison with other proteins

SDS-PAGE of the organic matrix molecules revealed eight species of proteins for *P. yessoensis*, four for *P. albicans*, and three for *C. farreri*. The three proteins of 77, 49, and 42 kDa in apparent molecular weight are common in size to *P. yessoensis* and *P. albicans*. *C. farreri* does not have any proteins of which molecular weights correspond to those of *P. yessoensis* and *P. albicans*. Although the molecular weight does not match, the 17 kDa protein of *C. farreri* has a similar N-terminal sequence, with an (Asp-X)3 repeat, as those of 104, 77, 49, and 42 kDa proteins of *P. yessoensis* and *P. albicans* (Figure 12). The three proteins of 104, 77, and 49 kDa of *P. yessoensis* share the same N-terminal amino acid sequence with each other, and all these proteins are glycosylated. At least three proteins (77, 49, and 42 kDa) of *P. albicans* also share the same N-terminal amino acid sequence with each other, and all these proteins are glycosylated. Those proteins of 104, 77, and 49 kDa of *P. yessoensis*, 77, 49, and 42 kDa of *P. albicans*, and 17 kDa of *C. farreri*, thus have similar characters, and may be homologous with each other.

In order to find known proteins that have similar sequences to MSP-1, homology search was carried out against all the protein sequences stored in the DDBJ data bank. The D domain of MSP-1 showed the highest sequence similarity with phosphophoryn, a major component of non-collagenous proteins in dentin, being involved in mineralization (Gu et al. 1998, Figure 13). Phosphophoryn has some regular motifs such as (Asp-

Ser-Ser) and (Asp-Ser), which are different from the motifs in the D domains of MSP-1, such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp), and the overall arrangement of Asp-residues are also different between them. The high sequence similarity may be due to the richness in Asp and Ser residues in both phosphoryn and the D domains of MSP-1. Because Asp residues have been thought as very important in protein-mineral interactions, they may be functionally related, but the evolutionary relation between them is obscure.

The SG domain of MSP-1 showed the highest sequence similarity with the "GS" domain of Lustrin A, a matrix protein of gastropod nacre (Shen et al. 1997, Figure 14), as mentioned above. Both sequences are rich in Gly and Ser residues, and the z-score calculated in FASTA analysis, which can evaluate the similality between two sequences with biased amino acid compositions (Lipman and Pearson 1985), is 466.8. This high value of z-score for the two amino acid sequences suggest that the similarity may not have resulted from the biased amino acid compositions nor have occurred by chance, and that the two amino acid sequences may be evolutionarily and/or functionally related.

The N-terminal and basic domains of MSP-1 did not show significant sequence similarity with any proteins in the DDBJ data bank.

The G domain of MSP-1 showed high sequence similarities with many proteins containing $(\mathrm{Ser})_n$ type sequences. The SGD and C-terminal domains showed high sequence similarity with many proteins containing $(\mathrm{Gly})_n$ type sequences. These similarities may be only due to having the Ser-rich or Gly-rich amino acid composition in common among these domains and proteins, and may not be reflecting their evolutionary relationships.

Homology search was also carried out using the BLAST program to search against the complete genome sequence of *Caenorhabditis elegans*

(The *C. elegans* Sequencing Consortium 1998). The T05C12 of hedgehog genes showed the highest sequence similarity with MSP-1 (Figure 15). Hedgehog proteins form a family of secreted molecules, involved in many patterning processes during development of the nervous system, limbs, bones, skin, and germ cells. The T05C12 protein of *C. elegans* is different from hedgehog proteins of other animals, such as *Drosophila* and mouse, with an extra sequence more or less similar to MSP-1 being added to a sequence common with other animals. Apparent functions of MSP-1 and hedgehog proteins appear different, but they could be evolutionarily related.

(4) Possible functions of MSP-1

i) Inhibition and induction of crystal growth

The soluble proteins in molluscan shells are thought as an essential mediator of shell formation, and may play versatile roles in biomineralization. One of the adovocated functions is inhibition of crystal growth.

Weiner and Traub (1980) hypothesized that the calcium-binding Asp residues of soluble proteins may inhibit or initiate crystal growth, depending upon the spacing between the Asp residues relative to that between the calcium ions in the crystal lattice of calcium carbonate. Acidic soluble proteins are thought to attach to the calcium ions on the surfaces of crystal lattice, thereby inhibiting further additions of ions and halting crystal growth.

Wheeler et al. (1981) demonstrated an effect of soluble proteins on the rate of calcium carbonate precipitation *in vitro*. Effects of soluble proteins extracted from oyster shells on the rate of precipitation of calcium carbonate were monitored by the change of pH of a solution of $NaHCO_3$ with or without (negative control) the soluble proteins when $CaCl_2$ was added to the solution. After addition of $CaCl_2$, the pH remained stable until nucleation occurred, and then declined. The length of the stable period cosiderably increased by the addition of soluble proteins. Furthermore, after the nucleation, the rate of crystal growth, as measured by the rate of pH decrease, was lower for the reactions with the soluble proteins than for those in the negative control. These results suggest that the soluble proteins extracted from oyster shells may have inhibitory roles in regulating crystal nucleation and growth.

Wilbur and Bernhardt (1984) found that polyaspartic and polyglutamic acids show an inhibitory effect on crystal formation *in vitro*, supporting the results demonstrated by Wheeler et al. (1981). Inhibition of crystal formation by homopolymers of acidic amino acids, but not by free amino acids, indicated that multiple COO^- groups in a polypeptide chain are required for the inhibition.

Based on the same acidic character of matrix proteins, another hypothesis that oriented nucleation is induced by matrix proteins has been proposed. Weiner and co-workers proposed an hypothesis predicting that soluble matrix proteins contain $(Asp\text{-}Y)_n$ domains, where Y represents serine or glycine (Weiner and Hood 1975; Weiner 1983; Weiner and Traub 1984; Weiner and Addadi 1991). Runnegar (1984) predicted that the matrix protein would have the repeating sequence of $(Asp\text{-}Gly\text{-}X\text{-}Gly\text{-}X\text{-}Gly)_n$. These repeating sequences may match the crystalline lattice and function as templates for mineralization.

The hypothesis that oriented nucleation is induced by acidic proteins is supported by *in vitro* experiments (Addadi et al. 1987, Addadi and Weiner 1985, Weiner and Addadi 1991). In these experiments, acidic soluble proteins extracted from either the calcitic, or the aragonitic shell layer of *Mytilus californianus* were first adsorbed onto a substrate, such as glass or plastic, and were then overlaid by a saturated solution of calcium malonate, calcium fumarate, calcium malate, or calcium carbonate. The calcite crystals nucleated under the presence of the acidic soluble proteins had a preferential orientation of their crystallographic axes, with the c-axis being perpendicular to the nucleation surface. This evidence suggests that the acidic soluble proteins tie down a layer of calcium ions, which are then associated with a layer of carbonate ions.

Addadi and Weiner (1985) interpreted that the acidic soluble proteins play two different roles in biomineralization. Acidic soluble proteins extracted from *M. californianus* interact specifically with certain crystal faces in solution, resulting in a drastic decrease in its growth rate relative to that of unaffected faces. The same proteins, once adsorbed on glass or plastic, act as nucleators of calcite specifically from (001) hexagonal plane. These results supported the interpretation that acidic soluble proteins are nucleators of calcite when adsorbed on a rigid substrate, but are inhibitors when in solution.

MSP-1 contains four kinds of acidic domains; N-terminal, SGD, D, and C-terminal domains. The D domain is different from the other three acidic domains in the following five points: (1) MSP-1 contains four D domains, while it contains one in each of the other tree domains; (2) The D domain is much larger than the other three domains. The D domain is 95 residues long, while the N-terminal, SGD, and C-terminal domains are 42, 27, and 50 residues long, respectively; (3) The D domain has some regular motifs such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp), but the other three domains do not have such regular motifs; (4) the proportion of acidic residues to all residues of D domain (39-40%) is higher than those of N-terminal (24%), SGD (33%), and C-terminal (26%) domains; (5) D domain uses Asp residues almost exclusively as the acidic residues. The proportion of Asp residues to acidic residues (Asp and Glu) of D domain is 92%, while those of the N-terminal, SGD, and C-terminal residues are 70%, 78%, and 62%, respectively.

Therefore, the functions of the D domain could well be different from that of the other three acidic domains. The conservative nature of amino acid sequences among the four D domains, the existence of some regular motifs, and the exclusive use of Asp residues between two acidic residues suggest that COO^- groups in side chains of acidic residues may take a rigid arrangement in the D domains, and that this regular arrangement of COO^- groups may be important for specific inhibition and/or oriented nucleation of calcite growth. Because of the relatively high content of Glu residues, the arrangements of COO^- groups of the other three acidic domains may not be so regular as those of D domains.

The foliated calcite consists of approximately 250 nm thick sheets (folia) formed of subparallel laths (Runnegar 1984). TEM studies have shown that the organic matrix exists between the sheets of foliated calcite of *Crassostrea virginica* (Watabe 1965). Runnegar (1984) classified foliated calcite found in bivalves into three types based on interfacial angles of the growing edges of foliated calcite laths. An X-ray diffraction trace was obtained from an inner surface of a disc cut from near the ventral margin of a left valve of *P. yessoensis*. The dominant peak was the rhombohedral form (101), supporting that the outer shell layer of *P. yessoensis* is composed of the Type 1 foliated calcite in the classification of Runnegar (1984) (Figure 16). The surfaces of the laths of Type 1 foliated calcite are the rhombohedral forms (101) and (110) (Figure 17). In order to form such a crystal shape of the laths, it would be necessary that the crystal growth on the Planes B and C are inhibited, and only the Plane A is allowed to grow (Figure 17).

The two-dimensional arrangements of Asp residues in MSP-1 are unknown. But in view of the regularity of the Asp residues in the D domains, it would be possible to assume that the D domains may attach alternatively either to the Plane B or C, because the Plane B and C have different rhombohedral forms, that is, different two-dimensional arrangements of calcium ions with each other.

In view of the above discussion, two hypothetical models are proposed here on the formation of Type 1 foliated calcite laths controlled by MSP-1.

(1) The laths of Type 1 foliated calcite grow between already-formed insoluble protein sheets. The Plane C (rhombohedral forms (101)) of the laths is in contact with the surfaces of insoluble protein sheets, thus the crystal growth on the Plane C does not occur. The Planes A (rhombohedral forms (101)) and B (rhombohedral forms (110)) are exposed to solution cotaining MSP-1. The D domains in MSP-1 attach only to the Plane B resulting from a precise stereochemical fit between the distance of the Asp residues of MSP-1 and the spacing of calcium ions in the surface of the Plane B, inhibiting further additions of ions and halting crystal growth. Consequently, growth on only the Plane A proceeds, producing the shape of the Type 1 laths.

(2) The laths of Type 1 foliated calcite grow between already-formed insoluble protein sheets. The Planes A (rhombohedral forms (101)) and B (rhombohedral forms (110)) are exposed to solution. The crystal growth of Plane B is inhibited by some soluble protein other than MSP-1. When MSP-1 is secreted to the solution, the D domains in MSP-1 attach only to the Plane A, inhibiting further crystal growth. Then the G domains (and the basic domain) in the MSP-1 are attached by some substrates, such as insolble matrix proteins, to which another layer of MSP-1 is attached, with the D domains being arranged on the opposite side, exposed to the solution. Consequently, COO$^-$ groups in the Asp residues in the D domains adsorb

Ca^{2+} in solution, inducing oriented nucleation and crystal growth.

ii) Implications for the polymorph formation

Of the three common calcium carbonate polymorphs, calcite, aragonite, and vaterite, the former two are commonly formed by organisms. Vaterite, a less stable polymorph, in contrast rarely occurs in organisms. Watabe and Wilbur (1960) reported that insoluble organic matrix, extracted from mollusc shells can control calcite-aragonite polymorphism both *in vitro* and *in vivo*, but the evidence was not conclusive. Recent evidence has demonstrated more definitively that soluble matrix proteins, rather than insoluble ones, can control the polymorphism *in vitro* (Belcher et al. 1996; Falini et al. 1996).

Falini et al. (1996) demonstrated that soluble matrix proteins, rather than insoluble ones, had a primary role in the control of polymorphism *in vitro*, and that in this process the entire matrix assembly, which included the insoluble matrix, was required. Soluble matrix proteins extracted from the aragonitic shell layers of some mollusks induced aragonite formation *in vitro* when preadsorbed on a substrate of ß-chitin and silk fibroin, which were used as an analogue of the insoluble matrix. When preadsobed on a substrate only of ß-chitin, aragonite was not generally induced. Soluble matrix proteins from the calcitic shell layers induced mainly calcite formation under the same conditions.

Belcher et al. (1996) also demonstrated that soluble matrix proteins had a primary role in the control of polymorphism *in vitro*. But they reported that in this process a substrate of ß-chitin and silk fibroin was not required. Therefore, the role of a substrate of ß-chitin and silk fibroin still remains obscure.

Miyamoto et al. (1996) carried out molecular cloning of cDNA encoding nacrein, a major soluble organic matrix protein in the nacreous layer of the pearl oyster *Pinctada fucata*. Analysis of the deduced amino acid sequence revealed that the protein contained two domains: the carbonic anhydrase domain and the Gly-X-Asn (X=Asp, Asn, or Glu) repeat domain. The carbonic anhydrase domain was split into two subdomains with insertion of the Gly-X-Asn repeat domain between them. Nacrein indicated carbonic anhydrase activity, suggesting that nacrein actually functions as a carbonic anhydrase that catalyzes the HCO_3^- formation in the site of nacreous layer formation.

These results that soluble proteins can switch the polymorph of growing crystals, and that a carbonic anhydrase has been isolated as a major soluble matrix component of the aragonitic shell (Miyamoto et al. 1996). This fact suggests that the carbonic anhydrase activity to produce local accumulation of carbonate ions at the site of crystal growth could be important in the formation of aragonite. The absence of a carbonic anhydrase domain in MSP-1, the major soluble proteins of the calcitic shell, tends to support this hypothesis.

MSP-1 has aspartate-rich domains, which contain some regular motifs such as (Asp-Gly-Ser-Asp) and (Asp-Ser-Asp). One of these motifs, (Asp-Ser-Asp), coincide with the expected motif, in which Asp residues are separated by a serine residue, abundant in both the prismatic-calcite and nacreous-aragonite shell layer of *M. californianus* (Weiner 1981). This type of sequence is presumably not responsible for the mineralogical difference. But the other motif, (Asp-Gly-Ser-Asp), coincide with the sequence that was considered as abundant in the prismatic calcite shell layer of *Pinctada margaratifera* (Weiner 1981). The proteins containing the motif, that is, Asp residues are separated by a dipeptide, may be responsible for determining the particular type of microstructure or polymorph of calcium carbonate (Weiner 1983). The (Asp-Gly-Ser-Asp) motif existed in MSP-1 located in the foliated calcite shell layer, suggesting that it could be responsible for determining the type of polymorph, calcite. This hypothesis does not contradict with the absence of this motif, (Asp-Gly-Ser-Asp), in the full amino acid sequence of nacrein, a soluble protein in the nacreous layer of the pearl oyster.

5. Conclusions

The MSP-1 sequence reported here represents the first full length amino acid sequence for the soluble matrix protein in calcitic molluscan shells, having the typical Asp-rich domain, which has widely been believed to play an important role for matrix-mineral interactions.

In contrast with prevalent expectation, $(Asp\text{-}X)_n$- or $(Asp\text{-}Gly\text{-}X\text{-}Gly\text{-}X\text{-}Gly)_n$-type sequence motifs, the "template" motifs, do not exist in the Asp-rich domains of MSP-1, demanding revision of previous theories of protein-mineral interactions.

It is interesting to note that MSP-1 combines several structural motifs into single molecule: N-terminus, basic domain, SGD domain, C-terminus, and the most conspicuous structural motif of highly conserved units consisting of an acidic domain (D domain), basic-core containg domain (G domain), and a possibly flexible domain (SG domain). Those integrated units may play an important role in shell formation. MSP-1 may be involved in both induction of oriented nucleation and inhibition of crystal growth. It is possible that MSP-1 is also involved in control of aragonite-calcite polymorphism.

To understand the functions of matrix proteins, and biomineralization processes in general, it is necessary to determine not only primary structures, but also secondary and tertiary structures, to precisely localize the proteins in the biominerals, and to perform refined *in vitro* experiments using pure proteins. The recombinant clones for the MSP-1 gene provide a basis for those experiments. Furthermore, in order to understand the functions of matrix proteins *in vivo*, it would be useful to produce "transgenic shellfish" with a certain matrix protein gene disrupted. The kind of information obtained in this study is considered as a prerequisite for such a venture.

It now became possible to investigate the phylogenetic distribution of MSP-1 using southern hybridization, in mollusks and related phyla, to explore possible link between the presence of MSP-1 and some characters of shells, including the presence or absence of shells. The taxonomical distribution data would also give important insight into the origin of MSP-1 and the phylogeny of mollusks and related phyla.

Based on the deduced amino acid sequence of MSP-1, polycloned antibodies have been prepared against a systhetic polypeptide, of which sequence corresponds to a part of the D domain of MSP-1. Preliminary experiments indicate that these antibodies react with native proteins extracted from *P. yessoensis* and *P. albicans* (T. Ogawa and K. Endo, pers. communication). Using these antibodies, investigations are now underway to localize MSP-1 in the shells of *P. yessoensis*. Combined with detailed examinations of mineralogy and microstructure of the shells, the *in situ* localization of MSP-1 may help to understand the functions of MSP-1. Furthermore, the antibodies can be used to purify the native MSP-1 through affinity chromatography, and the use of the proteins thus prepared may be fruitful in the *in vitro* mineralization experiments to reveal elemental processes in biomineralization.

Acknowledgements

The author wishes to thank Dr. Kazuyoshi Endo (University of Tokyo) for his supervision of this study, critical reading of the first draft, and encouragement during research. Thanks are due to Prof. Kazushige Tanabe and Associate Prof. Tatsuo Oji (University of Tokyo) for the fruitful discussion and for critical reading of the first draft. I thank Prof. Hiromichi Nagasawa (University of Tokyo) for technical advice in RT-PCR and N-terminal sequence determinations and for the use of the protein sequencer. I also wish to thank Associate Prof. Toshiki Watanabe (University of Tokyo) for important suggestions and the help in performing northern blot analysis.

References cited

Addadi, L., Moradian, J., Shay, E., Maroudas, N. G. and Weiner, S. (1987) A chemical model for the cooperation of sulfates and carboxylates in calcite crystal nucleation: Relevance to biomineralization. Proceedings of National Academy of Science, USA, 84, 2732-2736.

Addadi, L. and Weiner, S. (1985) Interactions between acidic proteins and crystals: stereochemical requirements in biomineralization. Proceedings of National Academy of Science, USA, 82, 4110-4114.

Addadi, L. and Weiner, S. (1997) A pavement of pearl. Nature, 389, 912-915.

Belcher, A. M., Wu, X. H., Christensen, R. J., Hansma, P. K., Stucky, G. D., and Morse, D. E. (1996) Control of crystal phase switching and orientation by soluble mollusc-shell proteins. Nature, 381, 56-58.

Berman, A., Addadi, L., and Weiner, S. (1988) Interactions of sea-urchin skeleton macromolecules with growing calcite crystals - a study of intracrystalline proteins. Nature, 331, 546-548.

Berman, A., Addadi, L., Kvick, A., Leiserowitz, L., Nelson, M., and Weiner, S. (1990) Intercalation of sea urchin proteins in calcite: study of a crystalline composite material. Science, 250, 664-667.

Borbas, J. E., Wheeler, A. P., and Sikes, C. S. (1991) Molluscan shell matrix phosphoproteins: Correlation of degree of phosphorylation to shell mineral microstructure and to in vitro regulation of mineralization. The Journal of Experimental Zoology, 258, 1-13.

Campbell, K. P., MacLennan, D. H., and Jorgensen, A. O. (1983) Staining of the Ca^{2+}-binding proteins, calsequestrin, calmodulin, troponin C, and S-100, with the cationic carbocyanine dye "Stains-all". The Journal of Biological Chemistry, 258, 11267-11273.

Chomzynski, P. (1993) A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples. Biotechniques, 15, 532-537.

Chou, P. Y. and Fasman, G. D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Advances in Enzymology, 47, 45-148.

Curry, J. D. (1990) Biomechanics of mineralized skeletons. In J. G. Carter, Ed., Skeletal biomineralization: Patterns, processes and evolutionary trends, Volume I, p. 11-25. Van Nostrand Reinhold, New York.

Emlet, R. B. (1982) Echinoderm calcite: a mechanical analysis from larval spicules. The Biological Bulletin, 163, 264-275.

Falini, G., Albeck, S., Weiner, S., and Addadi, L. (1996) Control of aragonite or calcite polymorphism by mollusk shell macromolecules. Science, 271, 67-69.

Freeman, J. A. and Wilbur, K. M. (1948) Carbonic anhydrase in molluscs. The Biological Bulletin, 94, 55-59.

Frohman, M. A. (1990) RACE: Rapid amplification of cDNA ends. In M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, Eds., PCR Protocols, p. 28-38. Academic Press, San Diego.

Gordon, J. and Carriker, M. R. (1980) Sclerotized protein in the shell matrix of a bivalve mollusc. Marine Biology, 57, 251-260.

Gu, K., Chang, S. R., Slavin, M. S., Clarkson, B. H., Rutherford, R. B., and Ritchie, H. H. (1998) Human dentin phosphophoryn nucleotide and amino acid sequence. European Journal of Oral Science, 106, 1043-1047.

Hare, P. E. (1963) Amino acids in the proteins from aragonite and calcite in the shells of *Mytilus californianus*. Science, 139, 216-217.

Heuer, A. H. (1998) Microstructural design of the crossed-lamellar shell:

Strombus gigas. In M. Goldberg and C. Robinson Eds., 6th international conference on the chemistry and biology of mineralized tissues, p. X. Vittel, France.

Holden, K. G., Yim, N. C. F., Griggs, L. J., and Weissbach, J. A. (1971) Gel electrophoresis of mucous glycoproteins. I. Effect of gel porosity. Biochemistry, 10, 3105-3109.

Hotta, S. (1969) Infrared spectra and conformation of protein constituting the nacreous layer of molluscan shell. Earth Science; Journal of the Association for Geological Collaboration in Japan, 23, 133-140

Jamieson, J. C. (1953) Phase equilibrium in the system calcite-aragonite. The Journal of Chemical Physics, 21, 1358-1390.

Kasai, H. and Ohta, N. (1981) Relationship between organic matrices and shell structures in recent bivalves. In T. Habe and M. Omori, Ed., Study of molluscan paleobiology, p. 101-106. Kokusai Insatsu, Tokyo. (in Japanese)

Kitano, Y. (1962a) The behavior of various inorganic ions in the separation of calcium carbonate from a bicarbonate solution. Bulletin of the Chemical Society of Japan, 35, 1973-1980.

Kitano, Y. (1962b) A study of the polymorphic formation of calcium carbonate in thermal springs with an emphasis on the effect of temperature. Bulletin of the Chemical Society of Japan, 35, 1980-1985.

Kitano, Y. and Hood, D. W. (1962) Calcium carbonate crystal forms formed from sea water by inorganic processes. The Journal of the Oceanographical Society of Japan, 18, 141-145.

Kitano, Y. and Hood, D. W. (1965) The influence of organic material on the polymorphic crystallization of calcium carbonate. Geochimica et Cosmochimica Acta, 29, 29-41.

Kitano, Y., Kanamori, N., and Tokuyama, A. (1969) Effect of organic matter on solubilities and crystal form of carbonates. American Zoologist, 9, 681-688.

Krampitz, G. P. (1982) Structure of the organic matrix in mollusc shells and avian eggshells. In G. H. Nancollas, Ed., Biological mineralization and demineralization, p. 219-232. Springer-Verlag, Berlin, Heidelberg, New York.

Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of protein. Journal of Molecular Biology, 157, 105-132.

Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature, 227, 680-685.

Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. Science, 227, 1435-1441.

Lowenstam, H. A. (1981) Minerals formed by organisms. Science, 211, 1126-1131.

Lowenstam, H. A. and Weiner, S. (1989) On Biomineralization. Oxford University Press, New York

Mann, S. (1996) Biomineralization and biomimetic materials chemistry. In S. Mann, Ed., Biomimetic materials chemistry, p 1-40. VCH Publishers, New York.

Mann, S., Archibald, D. D., Didymus, J. M., Douglas, T., Heywood, B. R., Meldrum, F. C., and Reeves, N. J. (1993) Crystallization at inorganic-organic interfaces: biominerals and biomimetic synthesis. Science, 261, 1286-1292.

Miyamoto, H., Miyashita, T., Okushima, M., Nakano, S., Morita, T., and Matsushiro, A. (1996) A carbonic anhydrase from the nacreous layer in oyster pearls. Proceedings of National Academy of Science, USA, 93, 9657-9660.

Nissen, H. U. (1969) Crystal orientation and plate structure in echinoid skeletal units. Science, 166, 1150-1152.

Raup, D. M. (1962) The phylogeny of calcite crystallography in echinoids. Journal of Paleontology, 36, 793-810.

Runnegar, B. (1984) Crystallography of the foliated calcite shell layers of bivalve molluscs. Alcheringa, 8, 273-290.

Samata, T. (1988) Studies on the organic matrix in molluscan shells - I. Amino acid composition of the organic matrix in the nacreous and prismatic layers. Venus, 47, 127-140.

Samata, T. and Krampitz, G. (1981) Ca-binding polypeptides in oyster shells. Malacologia, 22, 225-233.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular cloning, second edition. Cold Spring Harbor Laboratory Press, New York

Sarashina, I. and Endo, K. (1998) Primary structure of a soluble matrix protein of scallop shell: Implications for calcium carbonate biomineralization. American Mineralogist, 83, 1510-1515.

Shen, X., Belcher, A. M., Hansma, P. K., Stucky, G. D., and Morse, D. E. (1997) Molecular cloning and characterization of Lustrin A, a matrix protein from shell and pearl nacre of *Haliotis rufescens*. The Journal of Biological Chemistry, 272, 32472-32481.

Sudo, S., Fujikawa, T., Nagakura, T., Ohkubo, T., Sakaguchi, K., Tanaka, M., Nakashima, K., and Takahashi, T. (1997) Structures of mollusc shell framework proteins. Nature, 387, 563-564.

The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science, 282, 2012-2018.

Watabe, N. (1965) Studies on shell formation XI. Crystal-matrix relationships in the inner layers of mollusk shells. Journal of Ultrastructure Research, 12, 351-370.

Watabe, N. and Wilbur, K. M. (1960) Influence of the organic matrix on crystal type in molluscs. Nature, 188, 334.

Weiner, S. (1981) Repeating amino acid sequences in the soluble protein of mollusk shell organic matrices: Their involvement in crystal formation. In A. Veis Ed., The chemistry and biology of mineralized connective tissues: Procedings of the first international conference on the chemistry and biology of mineralized connective tissues, p. 517-521. Elsevier, North-Holland.

Weiner, S. (1983) Mollusk shell formation: Isolation of two organic matrix proteins associated with calcite deposition in the bivalve *Mytilus californianus*. Biochemistry, 22, 4139-4145

Weiner, S. and Addadi, L. (1991) Acidic macromolecules of mineralized tissues: the controllers of crystal formation. Trends in Biochemical Science, 16, 252-256.

Weiner, S. and Hood, L. (1975) Soluble protein of the organic matrix of mollusk shells: a potential template for shell formation. Science, 190, 987-989.

Weiner, S. and Traub, W. (1980) X-ray diffraction study of the insoluble organic matrix of mollusk shells. FEBS Letters, 111, 311-316.

Weiner, S. and Traub, W. (1984) Macromolecules in mollusc shells and their functions in biomineralization. Philosophical Transactions of Royal Society of London, B304, 425-434.

Wheeler, A. P. (1992) Phosphoproteins of oyster (*Crassostrea virginica*) shell organic matrix. In S. Suga and N. Watabe Eds., Hard tissue mineralization and demineralization, p. 171-187. Springer-Verlag, Tokyo, Berlin, Heidelberg, New York.

Wheeler, A. P., George, J. W., and Evans, C. A. (1981) Control of calcium carbonate nucleation and crystal growth by soluble matrix of oyster shell. Science, 212, 1397-1398.

Wilbur, K. M. (1960) Shell structure and mineralization in molluscs. In

R. F. Sognnaes, Ed., Calcification in biological systems, p. 15-40. American Association for the advancement of Science, Washington, D. C.

Wilbur, K. M. and Bernhardt, A. M. (1984) Effects of amino acids, magnesium, and molluscan extrapallial fluid on crystalization of calcium carbonate: *in vitro* experiments. The Biological Bulletin, 166, 251-259.

Figure captions

Figure 1. SDS-PAGE patterns of organic matrices in shells of *P. yessoensis* in 10%(w/v) polyacrylamide gels. Lane A, B, and C represent organic matrix proteins stained with coomassie brilliant blue, periodic acid and Schiff's reagent, and silver, respectively. Lane M represents molecular weight standards.

Figure 2. SDS-PAGE patterns of organic matrices in shells of *P. albicans* in 10%(w/v) polyacrylamide gels. Lane A, B, and C represent organic matrix proteins stained with coomassie brilliant blue, periodic acid and Schiff's reagent, and silver, respectively. Lane M represents molecular weight standards.

Figure 3. SDS-PAGE patterns of organic matrices in shells of *C. farreri* in 10%(w/v) polyacrylamide gels. Lane A represents organic matrix proteins stained with coomassie brilliant blue. Lane M represents molecular weight standards.

Figure 4. Electrophoretic patterns of PCR products in 1%(w/v) agarose gel. Lane A and B represent PCR products amplified using primer pairs of P7-P8 and P7-PT (Table 1.), respectively. Lane M represents molecular weight standards.

Figure 5. Northern blot analysis of MSP-1 transcripts. Poly(A)$^+$ RNA was prepared from whole mantle tissues.

Figure 6. cDNA sequence of MSP-1 and the deduced amino acid sequence. Numbers on the right indicate positions of the nucleotides in the

MSP-1 cDNA sequence. The putative signal peptide is in bold. The vertical arrow designates the cleavage site.The residues determined by Edman degradation are underlined. The initiation codon and the stop codon are boxed. The putative polyadenylation signal (ATTAAA) is also boxed. One-letter abbreviations of amino acids are used.

Figure 7. Modular structure of MSP-1. Numbers on the right indicate positions of the amino acid residues in the MSP-1 sequence. The aspartate-rich domains are boxed. Acidic residues are in bold. One-letter abbreviations of amino acids are used.

Figure 8. Schematic representation of the modular structure of MSP-1.

Figure 9. Hydropathy plot and major domains of MSP-1. The analysis was carried out with the software GENETYX-MAC ver. 9 (Software Development), based on the argorithm of Kyte and Doolittle (1982).

Figure 10. Sequence alignment of aspartate-rich domains (D domains) of MSP-1. Identical amino acids are indicated by asterisks. Numbers in the right indicate positions of the amino acid residues in the MSP-1 sequence. Acidic residues are in bold. Potential *N*-glycosylation sites are boxed. Potential phosphorylation sites are underlined. One-letter abbreviations of amino acids are used.

Figure 11. Sequence alignment of G domains of MSP-1. Identical amino acids are indicated by asterisks. Numbers on the right indicate positions of the amino acid residues in the MSP-1 sequence. Basic residues are in bold. One-letter abbreviations of amino acids are used.

Figure 12. Alignment of the N-terminal amino acid sequence of major soluble proteins in shells of *P. yessoensis* (Py), *P. albicans* (Pa), and *C. farreri* (Cf). Amino acids identical to those in *P. yessoensis* are indicated by asterisks. Acidic residues are in bold. For one-letter abbreviation of amino acids are used.

Figure 13. Alignment of the amino acid sequence of "D" domain of MSP-1 and human phosphophoryn. Numbers on the both sides indicate positions of the amino acid residues in the MSP-1 and human phosphophoryn (Gu et al. 1998) sequences. The residue of phosphophoryn identical to that of MSP-1 sequence is replaced by a dot. A gap is marked as a dash. One-letter abbreviations of amino acids are used.

Figure 14. Alignment of the amino acid sequence of MSP-1 and Lustrin A. Numbers on either side indicate positions of the amino acid residues in the MSP-1 and Lustrin A (a matrix protein of gastropod nacre, Shen et al. 1997) sequences. The residue of Lustrin A identical to that of MSP-1 sequence is replaced by a dot. A gap is marked as a dash. One-letter abbreviations of amino acids are used.

Figure 15. Alignment of the amino acid sequence of MSP-1 and T05C12.10 (a hedgehog gene of *Caenorhabditis elegans*, The *C. elegans* Sequencing Consortium 1998). Numbers on either side indicate positions of the amino acid residues in the MSP-1 and T05C12.10 sequences. The residue of T05C12.10 identical to that of MSP-1 sequence is replaced by a dot. A gap is marked as a dash. One-letter abbreviations of amino acids are used.

Figure 16. X-ray diffraction trace obtained from an inner surface of a disc cut from the near ventral margin of a left valve of *P. yessoensis*.

Figure 17. Crystallographic model for the laths of the Type 1 foliated calcite.

Table 1. Oligonucleotide primers used in this study. Y = T or C, R = A or G, and N = G, A, T, or C. "Position" refer to the corresponding site of the 3' end of each primer to the nucleotide position in the MSP-1 cDNA sequence. Restriction sites are in bold.

Table 2. Results of 3' RACE experiments attempted using various primer pairs (Table 2) and two different methods of cDNA preparations. S-cDNA indicate single-strand cDNA synthesized using the SuperScript preamplification system (Life Technologies) and D-cDNA indicated double-strand cDNA synthesized using the Marathon cDNA amplification kit (Clontech)

Table 3. Amino acid compositions of MSP-1, total soluble matrix of *P. yessoensis* (Kasai and Ohta 1981), and an hplc fraction of soluble matrix of the oyster *Crassostrea virginica* (Wheeler 1992). ND = no data

Figure 1

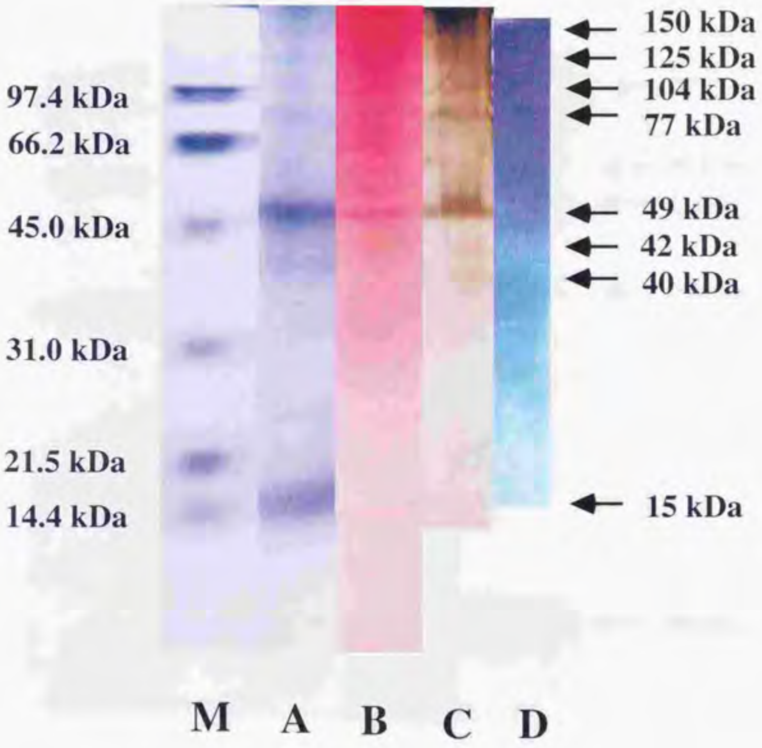
97.4 kDa
66.2 kDa
45.0 kDa
31.0 kDa
21.5 kDa
14.4 kDa
150 kDa
125 kDa
104 kDa
77 kDa
49 kDa
42 kDa
40 kDa
15 kDa
M
A
B
C
D

Figure 2

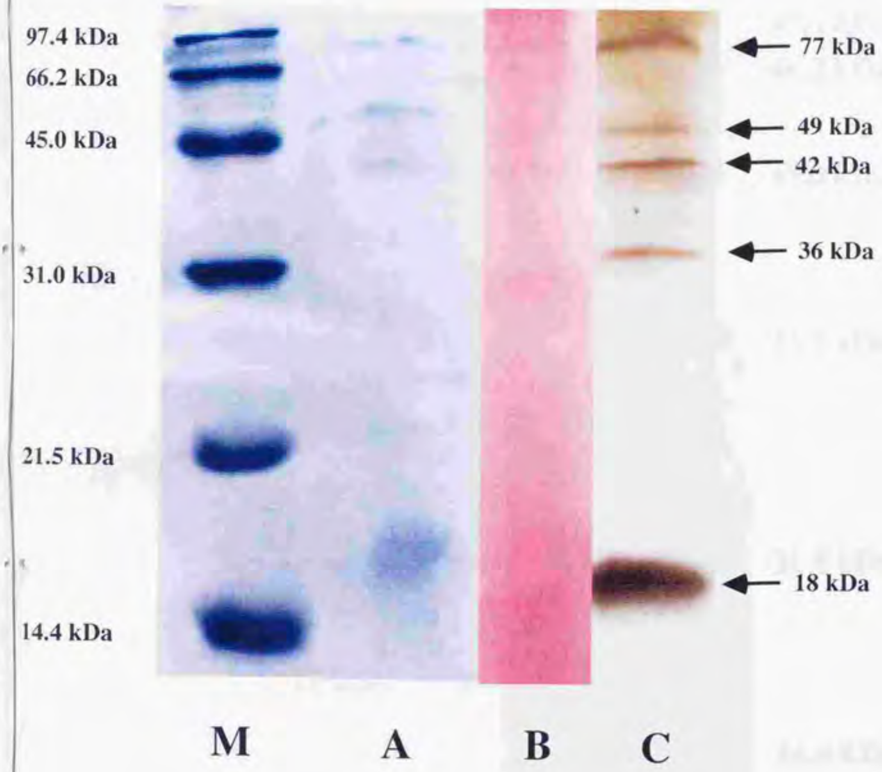
97.4 kDa
66.2 kDa
45.0 kDa
31.0 kDa
21.5 kDa
14.4 kDa
77 kDa
49 kDa
42 kDa
36 kDa
18 kDa
M
A
B
C

Figure 3

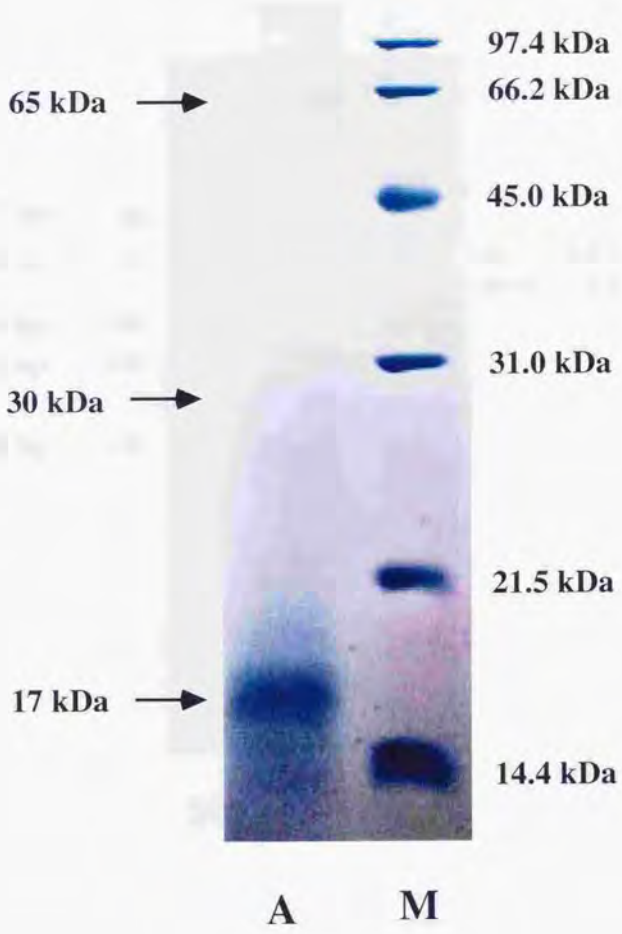
97.4 kDa
65 kDa
66.2 kDa
45.0 kDa
31.0 kDa
30 kDa
21.5 kDa
17 kDa
14.4 kDa
A
M

Figure 4

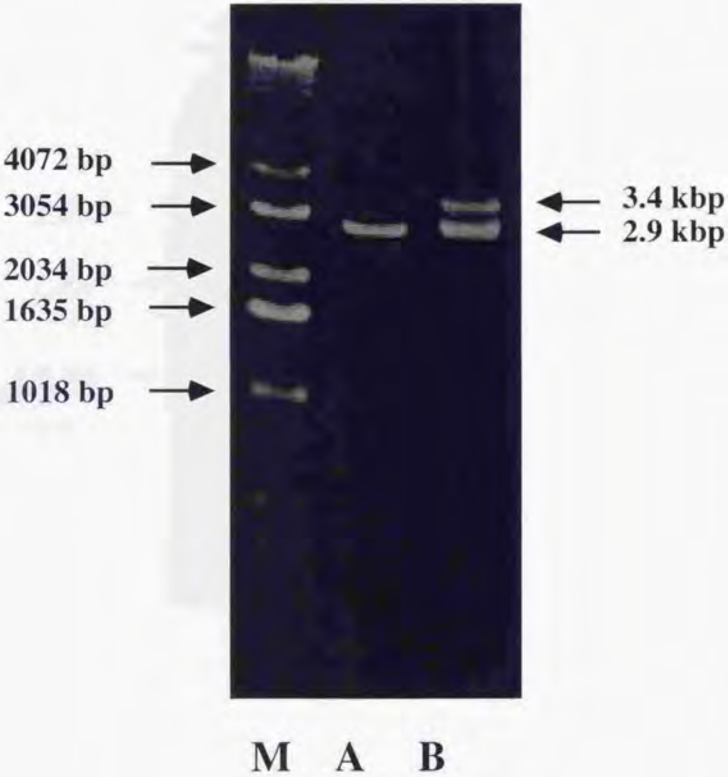

Figure 5

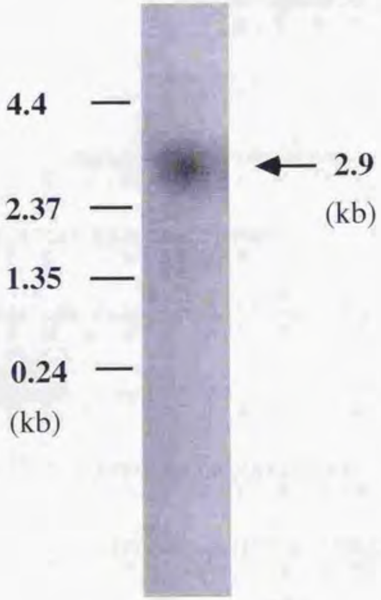

Figure 6

```
        10        20        30        40        50        60
CTCTCCCGGTAACGGACTTTAGCAAACGAGAAAGATGTACATCAAGTTGATCCTGGGTGT
                                   M  Y  I  K  L  I  L  G  V

        70        80        90       100       110       120
CCTTGCTCTCGTGGCATTAGCCGTTTCTGCGCCACTGGATACAGATAAAGATTTAGAATT
  L  A  L  V  A  L  A  V  S  A  P  L  D  T  D  K  D  L  E  F

       130       140       150       160       170       180
TCATCTAGATAGCTTACTAAATGCAGCCGAAGATGGCGGTGGTGGCGATGCTGCTGGCGC
  H  L  D  S  L  L  N  A  A  E  D  G  G  G  G  D  A  A  G  A

       190       200       210       220       230       240
CGAGAAGGCAGCACCAGCAGCAGATCTAAGCGGAGGTAGCAAAGGAGGAAGCAAAGGAAG
  E  K  A  A  P  A  A  D  L  S  G  G  S  K  G  G  S  K  G  S

       250       260       270       280       290       300
TAGCACAAGAAGTAGTAAGGGAGGTAGTAAGGGAGGTAGTAAGGGAGGCAATGGTGGTGG
  S  T  R  S  S  K  G  G  S  K  G  G  S  K  G  G  N  G  G  G

       310       320       330       340       350       360
AGATGCTGATGATTCAAGTAGCTCAAGCGGTTCTGATTCGGGAAGTTCTGGAAGTGACGA
  D  A  D  D  S  S  S  S  S  G  S  D  S  G  S  S  G  S  D  E

       370       380       390       400       410       420
AGAATCAGATGATTCCAGCTCTAGTTCCAGTTCTGGTTCTGGTTCCGGCTCTGGCTCAGG
  E  S  D  D  S  S  S  S  S  S  S  G  S  G  S  G  S  G  S  G

       430       440       450       460       470       480
TTCTGGCTCCGGCTCAAGCTCCAGCTCTGGCTCATCCAGCGATGGATCTGATGATGGTTC
  S  G  S  G  S  S  S  S  S  G  S  S  S  D  G  S  D  D  G  S

       490       500       510       520       530       540
CGATGACGGGTCTGATTCAGGTGACGATGCTGATTCCGCTAATGCTGATGACCTTGATTC
  D  D  G  S  D  S  G  D  D  A  D  S  A  N  A  D  D  L  D  S

       550       560       570       580       590       600
CAATGATTCCGATGATTCCGATAACTCCGGTTCCAATGGCGAGTCTGACTCTGATAACTC
  N  D  S  D  D  S  D  N  S  G  S  N  G  E  S  D  S  D  N  S

       610       620       630       640       650       660
CTCCTCCGACGATGGCGATGGTTCCGATTCCGGTTCCGATTCTGGCAATGATAGTCAGTC
  S  S  D  D  G  D  G  S  D  S  G  S  D  S  G  N  D  S  Q  S

       670       680       690       700       710       720
CGATGACGCTTCTAATAATGATTCTGATGACTCCGATGACTCGGATGATTCTTCTAATGA
  D  D  A  S  N  N  D  S  D  D  S  D  D  S  D  D  S  S  N  D

       730       740       750       760       770       780
TGTAAATGAATCAAATTCTGATGAATCTGGCCCTGGAGGATATGGAGGCAATGGACCAGC
  V  N  E  S  N  S  D  E  S  G  P  G  G  Y  G  G  N  G  P  A

       790       800       810       820       830       840
AGGCAATGGAGGCAAGGGACGCAAGGGAGGCAATGGAGGAGGCAATGGAGGCAATGGAGG
  G  N  G  G  K  G  R  K  G  G  N  G  G  G  N  G  G  N  G  G

       850       860       870       880       890       900
TGATGGATCCAGTTCTAGTTCGAGCTCTGGTTCCGGCTCTGGTTCTGGCTCCGGCTCTGG
  D  G  S  S  S  S  S  S  S  G  S  G  S  G  S  G  S  G  S  G
```

```
       910       920       930       940       950       960
CTCAAGTTCTGGCTCAGGTTCTGGCTCAGGTTCTGGGTCCGGCTCAAGCTCTAGCTCTAG
 S  S  S  G  S  G  S  G  S  G  S  G  S  G  S  S  S  S  S  S

       970       980       990      1000      1010      1020
CTCATCCGGCGATGGATCTGATGATGGTTCCGATGACGGGTCTGATTCAGGTGACGATGC
 S  S  G  D  G  S  D  D  G  S  D  D  G  S  D  S  G  D  D  A

      1030      1040      1050      1060      1070      1080
TAATTCCGCTAATGCTGATGACCTTGATTCCAATGATCCCGATGATTCCGATAACTCCGG
 N  S  A  N  A  D  D  L  D  S  N  D  P  D  D  S  D  N  S  G

      1090      1100      1110      1120      1130      1140
TTCCAATGGCGAGTCTGACTCTGATAACTCTTCCTCCGACGATGGCGATGGTTCCGATTC
 S  N  G  E  S  D  S  D  N  S  S  S  D  D  G  D  G  S  D  S

      1150      1160      1170      1180      1190      1200
CGGTTCCGATTCCGGCAGAGATAGTCAGTCCGATGACGCTTCTAATAATGATTCTGATGA
 G  S  D  S  G  R  D  S  Q  S  D  D  A  S  N  N  D  S  D  D

      1210      1220      1230      1240      1250      1260
CTCCGATGACTCTGATAACTCGTCTACTGATACTGGCGAATCCGATTCCGATGAATCTGG
 S  D  D  S  D  N  S  S  T  D  T  G  E  S  D  S  D  E  S  G

      1270      1280      1290      1300      1310      1320
TCCTGGAGGATATGGAGGCAATGGACCAGCAGGCAATGGAGGTAAGGGACGCAAGGGAGG
 P  G  G  Y  G  G  N  G  P  A  G  N  G  G  K  G  R  K  G  G

      1330      1340      1350      1360      1370      1380
CAATGGAGGAGGCAATGGAGGAGGCAATGGAGGCAATGGAGGTGATGGATCCAGTTCTAG
 N  G  G  G  N  G  G  G  N  G  G  N  G  G  D  G  S  S  S  S

      1390      1400      1410      1420      1430      1440
TGCTAGCTCTGGTTCCGGCTCTGGTTCTGGCTCCGGCTCTGGTTCTGGCTCAGGTTCTGG
 A  S  S  G  S  G  S  G  S  G  S  G  S  G  S  G  S  G  S  G

      1450      1460      1470      1480      1490      1500
CTCAGGTTCTGGCTCAGGTTCTGGGTCCGGCTCAAGCTCTAGCTCTAGCTCATCCGGCGA
 S  G  S  G  S  G  S  G  S  G  S  S  S  S  S  S  S  S  G  D

      1510      1520      1530      1540      1550      1560
TGGATCTGATGATGGTTCCGATGACGGGTCTGATGATGGTGACGATGCTAATTCCGCTAA
 G  S  D  D  G  S  D  D  G  S  D  D  G  D  D  A  N  S  A  N

      1570      1580      1590      1600      1610      1620
TGCTGATGACCTTGATTCCAATGATCCCGATGATTCCGATAACTCCGGTTCCAATGGCGA
 A  D  D  L  D  S  N  D  P  D  D  S  D  N  S  G  S  N  G  E

      1630      1640      1650      1660      1670      1680
GTCTGACTCTGATAACTCTTCCTCCGACGATGGCGATGGTTCCGATTCCGGTTCCGATTC
 S  D  S  D  N  S  S  S  D  D  G  D  G  S  D  S  G  S  D  S

      1690      1700      1710      1720      1730      1740
CGGCAGAGATAGTCAGTCCGATGACGCTTCTAATAATGATTCTGATGACTCCGATGACTC
 G  R  D  S  Q  S  D  D  A  S  N  N  D  S  D  D  S  D  D  S

      1750      1760      1770      1780      1790      1800
TGATAATTCGTCTACTGATACTGGCGAATCCGATTCCGATGAGTCTGGGCCTGGAGGATA
 D  N  S  S  T  D  T  G  E  S  D  S  D  E  S  G  P  G  G  Y

      1810      1820      1830      1840      1850      1860
TGGAGGCAATGGACCAGCAGGCAATGGAGGCAAGGGACGCAAGGGAGGCAAAAGAGGAGG
 G  G  N  G  P  A  G  N  G  G  K  G  R  K  G  G  K  R  G  G

      1870      1880      1890      1900      1910      1920
CAATGGAGGCAATGGCAATGGAGGCAATGGAGGTGATGGATCCAGTTCTAGTTCGAGCTC
 N  G  G  N  G  N  G  G  N  G  G  D  G  S  S  S  S  S  S  S
```

```
      1930      1940      1950      1960      1970      1980
TGGTTCCGGCTCTGGTTCTGACTCCGGCTCTGGCTCAAGTTCTGGCTCAGGTTCTGGCTC
  G  S  G  S  G  S  D  S  G  S  G  S  S  S  G  S  G  S  G  S

      1990      2000      2010      2020      2030      2040
CGGCTCAAGCTCCAGCTCTAGCTCATCCGGCGATGGATCTGATGATGGTTCCGATGACGG
  G  S  S  S  S  S  S  S  S  G  D  G  S  D  D  G  S  D  D  G

      2050      2060      2070      2080      2090      2100
GTCTGATGATGGTGACGATGCTAATTCCGCTAATGCTGATGACCTTGATTCCAATGATCC
  S  D  D  G  D  D  A  N  S  A  N  A  D  D  L  D  S  N  D  P

      2110      2120      2130      2140      2150      2160
CGATGATTCCGATAACTCCGGTTCCAATGGCGAGTCTGACTCTGATAACTCTTCCTCCGA
  D  D  S  D  N  S  G  S  N  G  E  S  D  S  D  N  S  S  S  D

      2170      2180      2190      2200      2210      2220
CGATGGCGATGGTTCCGATTCCGGTTCCGATTCCGGCAGAGATAGTCAGTCCGATGACGC
  D  G  D  G  S  D  S  G  S  D  S  G  R  D  S  Q  S  D  D  A

      2230      2240      2250      2260      2270      2280
TTCTAATAATGATTCTGATGACTCCGATGACTCTGATAACTCGTCTACTGATACTGGCGA
  S  N  N  D  S  D  D  S  D  D  S  D  N  S  S  T  D  T  G  E

      2290      2300      2310      2320      2330      2340
ATCCGATTCCGATGAGTCTGGGCCTGGAGGATATGGAGGCAATGGACCAGCAGGCAATGG
  S  D  S  D  E  S  G  P  G  G  Y  G  G  N  G  P  A  G  N  G

      2350      2360      2370      2380      2390      2400
AGGCAAGGGACGTAAGGGAGGCAAAGGAGGCAGGGGAGGCAATGGAGGCAATGGAGGTGG
  G  K  G  R  K  G  G  K  G  G  R  G  G  N  G  G  N  G  G  G

      2410      2420      2430      2440      2450      2460
TGGATCCAGTTCTAGTTCCGGCTCTGATGCCGATTCTGGTTCTGATTCTGGTTCAAGTGA
  G  S  S  S  S  S  G  S  D  A  D  S  G  S  D  S  G  S  S  E

      2470      2480      2490      2500      2510      2520
AGAGTCTGATAGCGGTTCTGATAGCAGTTCTGGTTCCTCTGGTGGTGATGACGCTGACTC
  E  S  D  S  G  S  D  S  S  S  G  S  S  G  G  D  D  A  D  S

      2530      2540      2550      2560      2570      2580
CTCTTCAAGTTCCTCTTCTTCAGAGGAGGAAGCATAAATGTAATATCTTGAAACTGGACT
  S  S  S  S  S  S  S  E  E  E  A  *

      2590      2600      2610      2620      2630      2640
GGGCATTTGATTATATGGAGAACTATAAGCGGTGTGTTCGCGTCATGTGTAAGGAAAGAC

      2650      2660      2670      2680      2690      2700
AATCAATACTACTGCAGAAATATCGCACTTTGGTATACATGCCCATAGTATATTCCACTA

      2710      2720      2730      2740      2750      2760
CTTACCGCGATGAAATTTAGTTGTAATGATATGATTTGCATTATTGTTAATATATAGATT

      2770      2780      2790      2800      2810      2820
TTTTAAAACATATTTTGGACATTATTATTCACAATCTTAACTATTGGTGCCAGAAAATCG

      2830      2840      2850      2860      2870      2880
CTCTTTATCGATCTGCAACAGTGCCAGCGACCATGGGATGTCCGATTGTCTTGTTGTTAT

      2890      2900      2910      2920      2930      2940
CAGCGATCTATACACTCAGCTCATATACGTTTTCTTGTTTTATTTGATTAAAACGTATGT

      2950      2960      2970      2980
TCATGTAAGTTAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Figure 7

N-terminus	LDTDKDLEFHLDSLLNAAEDGGGGDAAGAEKAAPAADLSGGS	42
Basic	KGGSKGSSTRSSKGGSKGGSKGGN	66
SGD	GGGDADDSSSSSGSDSGSSGSDEESDD	93
SG-1	SSSSSSSGSGSGSGSGSGSSSSSGSSS	122
D-1	DGSDDGSDDGSDSGDDADSANADDLDSNDSDDSDNSGSNGESDSDNSS SDDGDGSDSGSDSGNDSQSDDASNNDSDDSDDSDDSSNDVNESNSDE	217
G-1	SGPGGYGGNGPAGNGGKGRKGGNGGGNGGNGGDG	251
SG-2	SSSSSSSGSGSGSGSGSSSGSGSGSGSGSSSSSSSSG	292
D-2	DGSDDGSDDGSDSGDDANSANADDLDSNDPDDSDNSGSNGESDSDNSS SDDGDGSDSGSDSGRDSQSDDASNNDSDDSDDSDNSSTDTGESDSDE	387
G-2	SGPGGYGGNGPAGNGGKGRKGGNGGGNGGGNGGNGGDG	425
SG-3	SSSSASSGSGSGSGSGSGSGSGSGSGSGSGSGSSSSSSSSG	468
D-3	DGSDDGSDDGSDDGDDANSANADDLDSNDPDDSDNSGSNGESDSDNSS SDDGDGSDSGSDSGRDSQSDDASNNDSDDSDDSDNSSTDTGESDSDE	563
G-3	SGPGGYGGNGPAGNGGKGRKGGKRGGNGGNGNGGNGGDG	602
SG-4	SSSSSSSGSGSGSDSGSGSSSGSGSGSGSSSSSSSSG	639
D-4	DGSDDGSDDGSDDGDDANSANADDLDSNDPDDSDNSGSNGESDSDNSS SDDGDGSDSGSDSGRDSQSDDASNNDSDDSDDSDNSSTDTGESDSDE	734
G-4	SGPGGYGGNGPAGNGGKGRKGGKGGRGGNGGNGGGG	770
C-terminus	SSSSSGSDADSGSDSGSSEESDSGS DSSSGSSGGDDADSSSSSSSSEEEA	820

Figure 8

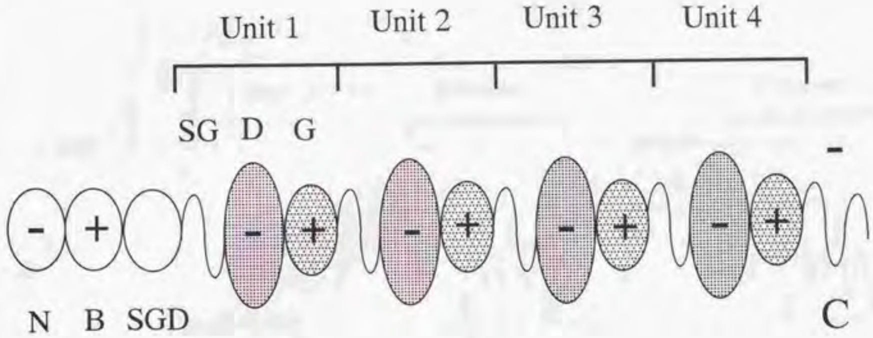

Figure 9

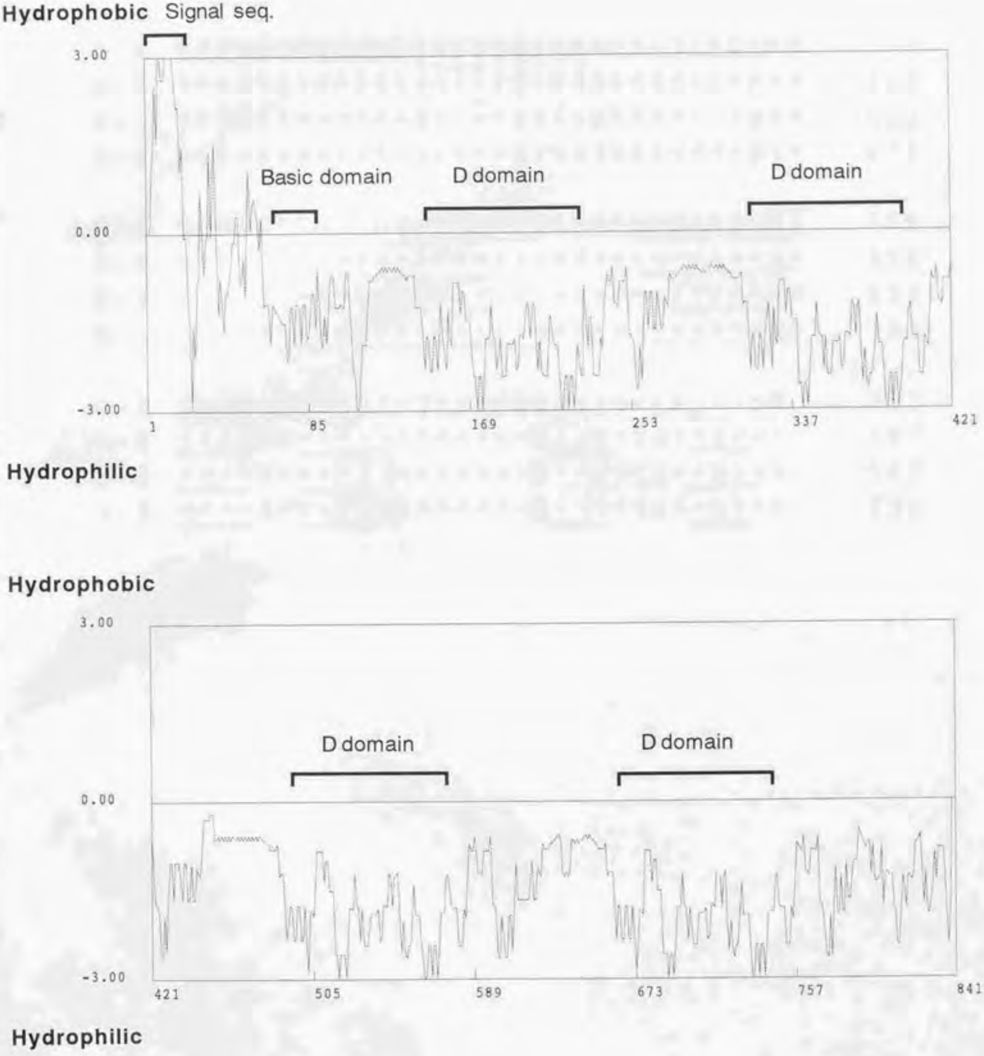
Hydrophobic
Signal seq.
3.00
Basic domain
D domain
D domain
0.00
-3.00
1
85
169
253
337
421
Hydrophilic
Hydrophobic
3.00
D domain
D domain
0.00
-3.00
421
505
589
673
757
841
Hydrophilic

Figure 10

```
D-1  DGSDDGSDDGSDSGDDADSANADDLDSNDSDD    154
D-2  *****************N***********P**    324
D-3  ************D****N***********P**    500
D-4  ************D****N***********P**    671

D-1  SDNSGSNGESDSDNSSSDDGDGSDSGSDSGND    186
D-2  ******************************R*    356
D-3  ******************************R*    532
D-4  ******************************R*    703

D-1  SQSDDASNNDSDDSDDSDDSSNDVNESNSDE     217
D-2  *******************N**T*TG**D***     387
D-3  *******************N**T*TG**D***     563
D-4  *******************N**T*TG**D***     734
```

Figure 11

```
G-1  SGPGGYGGNGPAGNGGKGRKGG-----NGGGNGGNGGDG 251
G-2  **********************NGGG-************ 425
G-3  **********************KRGGNG*N********* 602
G-4  **********************KGGR---********G* 770
```

Figure 12

```
        1   5    10    15    20
Py      LDTDK DLEFH LDSLL NAAED
Pa      S*T*A *TDED EEN
Cf      **P*P *DD
```

Figure 13

```
"D" domain of MSP-1     123 DGSDDGSDDGSDSGDDADSANADDLDSNDSDDSDNSGSNGESDSDNSS 170
Phosphophoryn           117 •S••SD•S•-•N•SS•S••SDS•SS••S•••S••S•N•SDS•••SD-• 162

"D" domain of MSP-1     171 SDDGDGSDSGSDSGNDSQSDDASNNDSDDSDDSDDSSNDVNESNSDE 217
Phosphophoryn           163 ••SS•S•••-•••KS••SKSESDSS•••SKS••S••NSSDSSD•••S 208
```

Figure 14

```
MSP-1         219 GPGGYGGNGPAGNGGKGRKG-GNGGGNGGNGGDGSSSSSSGSGSGSGSGSSSGSGSG  277
Lustrin A     960 •••DEP•VCCWD•RLRPTQ•S•S•S•S•SGS•SS•••••G•T•••••••••••••••••• 1019

MSP-1         278 SGSGSGSSSSSSSSGDGSDDGSDDGSDSGDDANSANADDLDSNDPDDSDNSGSNGESDSD  337
Lustrin A    1020 •S•A•••G•••G••-S••SS••SS••S••SGSG•GSGSSSA•GSGSS•GS•SGSSSGSGS 1078

MSP-1         338 NSSSDDGDGSDSGSDSGRDSQSDDASNNDSDDSDDSDNSSTDTGESDSDESGPGGYGGNG  397
Lustrin A    1079 G•••AS•S••S•••G••SS•G•GSG•GSG•SSGSG•GS••GSSSV•N•-WT•S•SSS•S• 1137

MSP-1         398 PAGNGGKGRKGGNGGGNGGGNGGNGGDGSSSSASSGSGSGSGSGSGSGSGSGSGSG  457
Lustrin A    1138 ----S•SSSWS•S•SSS•T•S•SSSWF••G••SG•••D•S•••S•A•••••S•••S•••• 1193

MSP-1         458 SGSSSSSSSSGDGSDDGSDDGSDDGDDANSANADDLDSNDPDDSDNSGSNGESDSDNSSS  517
Lustrin A    1194 ••---••LWF•S••SS•TGS••SS•SSSG•GSDSSSG•SSGST•GS•SGS•SASGSGTG• 1249

MSP-1         518 DDGDGSDSGSDSGRDSQS  536
Lustrin A    1250 GK•ASY•TDA•••S•NR• 1268
```

Figure 15

```
MSP-1       112 SGSSSSSGSSSDGSDDGSDDGSDSGDDADSANADDLDSNDSDDSDNSGSNGESDSDNSSS 171
T05C12.10   214 •W••W••S•W•TWARHAFNKAAAE•GE•AERIRTRMPIGEKTVAGAA•AT•AAG••K•NI 273

MSP-1       172 DDGDGSDSGSDSGNDSQSDDASNNDSDDSDDSDDSSNDVNESNSDESGPGGYGGNGPAGN 231
T05C12.10   274 NIHVE•N-----••NNN•FEGGRSS•EK••GQLNREI----•G•S•A•A••K••A•AD•A 324

MSP-1       232 GGKGRKGGNGGGNGGNGG-----DGSSSSSS----------SGSGSGSGSGSSSGSGS 276
T05C12.10   325 A•S•AGA•A•A•TN••INITVHT••K•GGNAVAVANANVTVN•A•GV•TT•T•AQT•NE• 384

MSP-1       277 GSGSGSGSSSSSSSG---DGSDDGSDDGSDSGDDANSANADDLDSNDPDDSDNSGSNGE 333
T05C12.10   385 •L•GSA•TDKAGGKK•GHG•SG•S•NNKNK•N•KGKGKGKN•EE•EE•NG•E•GN•KG•N 444

MSP-1       334 SDSDNSSSDDGDGSDSGSDSGRDSQSDDASNNDSDD---SDDSDNSSTDTGESDSDESGP 390
T05C12.10   445 GGNPKGEW•••••DED--•D•T•GG•KESG••GKGKGKG•G•G•GNRNGN•DGNGRPK•D 502

MSP-1       391 GGYGGNGPAGNGGKGRKGGNGGGNGGGNGGNG-GDGSSSSASSGSGSGSGSGSGSGSG 449
T05C12.10   503 •NIKI•IHSPDDNDLLEKDEN•P••K•GA•••N•••DKDNNGK•N•T•D•D•D•N•N•N• 562

MSP-1       450 -SGSGSGSGSGSSSSSSSSGDGSDDGSDDGSDDGDDANSANADDLDSNDPDDSDNSGSNG 508
T05C12.10   563 LT•D•N•T•D•DNNE•---•N•NG••••KN•GA•AGTKPE•REGG•G•GNGTG•GN•DG- 618

MSP-1       509 ESDSDNSSSDDGDGSDSGSDSGRDSQSDDASNNDSDDSDDSDNSSTDTGESDSDESGPGG 568
T05C12.10   619 ---N••GNGSK•L•TG••DGK•EGNK•GTPGKS•GKEDGAGS•G•GNGK•G•GNK••GS• 675
```

```
MSP-1        569 YGGNGPAGNGGKGRKGGKRGGNGGNGNGGN-GGDGSSSSSSSGSGSGSDSGSGSSSGSGS 627
T05C12.10    676 K••A•N-•KS•D•SGD••NN••••T•D•KDKN•K••G•GDNDK••TRA-A•K•NAE•N•K 733

MSP-1        628 GSGSSSSSSSSGDGSDDGSDDGSDDGDDANSANADDLDSNDPDDSDNSGSNGESDSDNSS 687
T05C12.10    734 •N•NDGKG•G•••••GA•GKGDKS•SESG•E•DGK•GKK•EGAGGEAAAGS•GANKGG•D 793

MSP-1        688 SDDGDGSDSGSDSGRD-----------SQSDDASNNDSDDSDDSDNSSTDTGESDSDESG 736
T05C12.10    794 G••D•VDVTDVEV•TKPLTGTKLEELLAKLPNETADGNATG•GNEFGTVQ••AKHNA••S 853

MSP-1        737 PGGYG---GNGPAGNGGKGRKGGKGGRGGNGGNGGGGSSSS--SGSDADSGSDSGSSEES 791
T05C12.10    854 AS•IPLVQARSNTV•••APVPPAP•SGATGS•TS•S•T•E•VTN••G•TESGST••GTTG 913

MSP-1        792 DSGSDSSSGSSGGDDADSSSSSSSSEEEA 820
T05C12.10    914 TGT•GTG•SGT•ASA•RT••IAGDAPQA• 942
```

Figure 16

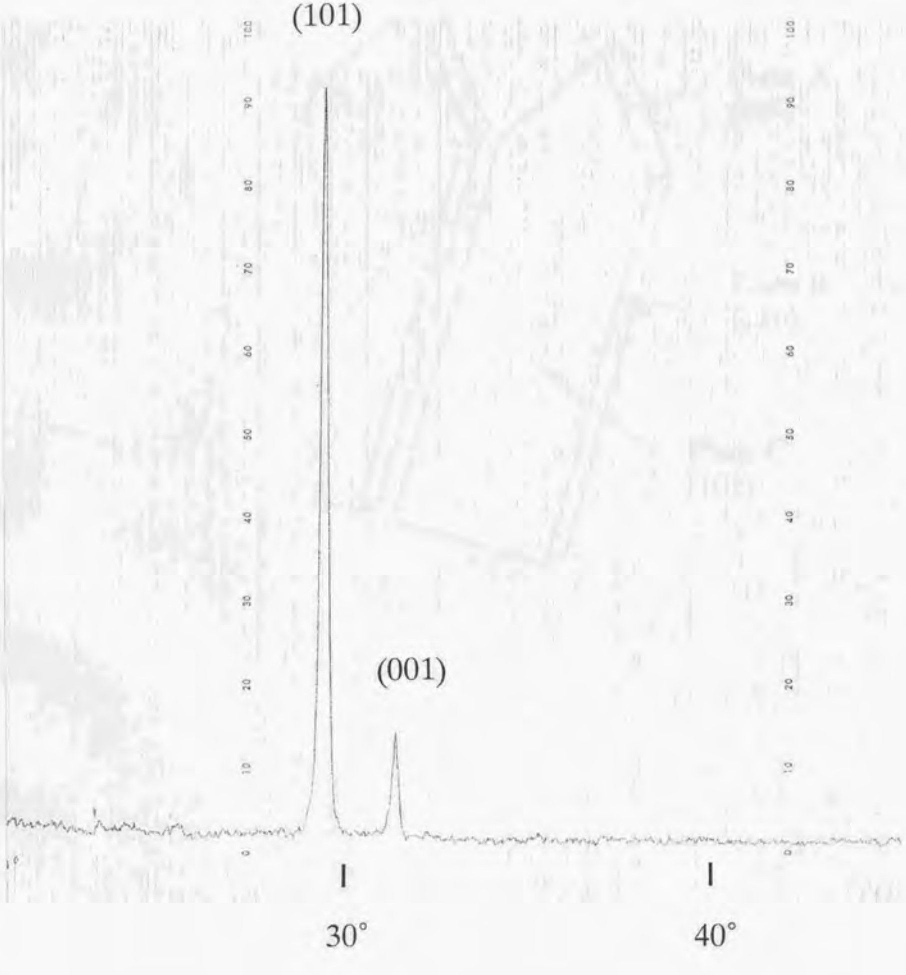

Figure 17

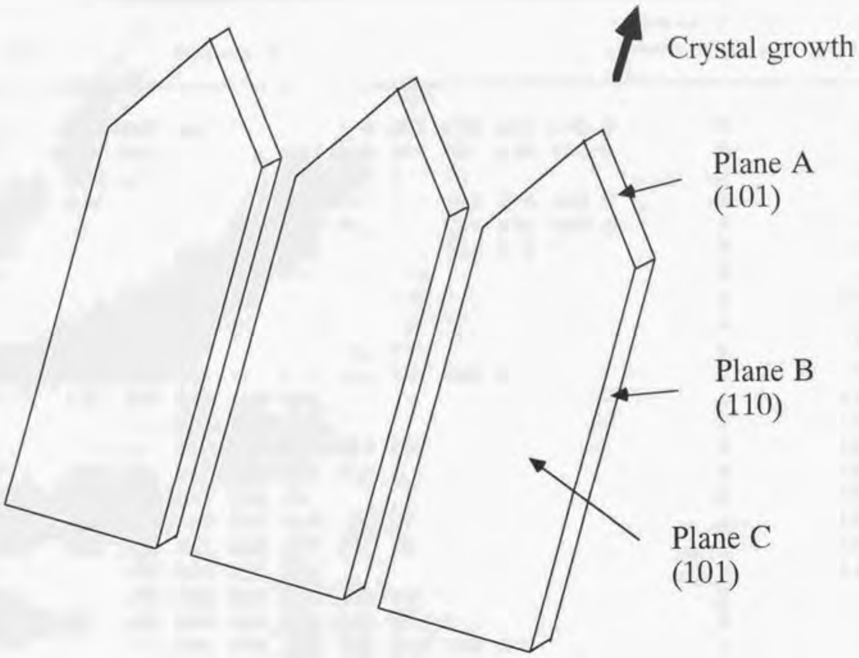

Table 1

Name	Sequence	Sense(S) or Antisense(A)	Position
P1	GAC **TGC AGG GTA CC**Y TNG AYA CNG AYA ARG A	S	111
P2	GA**G GTA CCC TGC AG**Y TNG ART TYC AYY TNG A	S	129
P3	TC**G GAT CC**R TCY TCN GCN GCR TT	A	154
P4	GAC **TGC AGG GTA CC**Y TTG ATA TTG AYA AGG A	S	111
P5	GA**G GTA CCC TGC AG**T TAG AAT TTC ATC TAG A	S	129
P6	GAC **TGC AGG GTA CC**A GCT TAC TAA AYG C	S	144
P7	CTC TCC CGG TAA CGG ACT TTA	S	21
P8	ACA TGA ACA TAC GTT TTA ATC	A	2946
P9	GAT TCA AGT AGC TCA AGC GG	S	330
P10	CAT ATC CTC CAG GGC CAG AT	A	745
P11	GAG GTA CCT TCT GAT GAA TCT GGC C	S	752
P12	CAT ATC CTC CAG GAC	A	1260
P13	GGG TCT GAT GCT GGT	S	1540
P14	CTC CAT TGC CTC CTC TTT TG	A	1849
P15	ACG CAA GGG AGG CAA AAG A	S	1855
P16	TGC CTC CAT TGC CAT	A	1872
P17	AAG CTC CAG CTC TAG CTC AT	S	2006
P18	AAC CGC TAT CAG ACT CTT CA	A	2458
P19	CAA AGG AGG CAG GGG	S	2376
PA	TC**G AAT TCG GAT CCG AGC TC**	A	-
PT	TC**G AAT TCG GAT CCG AGC TC**T T	A	-
AP1	CCA TCC TAA TAC GAC TCA CTA TAG GGC	-	-
AP2	ACT CAC TAT AGG GCT CGA GCG GC	-	-

Table 2

primer pair	template DNA	size of PCR product(s) (kb)
P1 - PA	S-cDNA	not amplified (1)
P2 - PA	PCR prod. of (1)	not amplified
P2 - PA	S-cDNA	not amplified
P4 - PA	S-cDNA	not amplified (2)
P5 - PA	PCR prod. of (2)	not amplified (3)
P5 - PA	S-cDNA	2.1 (4)
P6 - PA	PCR prod. of (2)	not amplified
P6 - PA	PCR prod. of (3)	not amplified
P6 - PA	PCR prod. of (4)	2.1
P6 - PA	S-cDNA	2.1
P4 - AP1	D-cDNA	not amplified (5)
P4 - AP2	PCR prod. of (5)	not amplified (6)
P5 - AP1	D-cDNA	not amplified (7)
P5 - AP1	PCR prod. of (5)	not amplified (8)
P5 - AP2	PCR prod. of (5)	not amplified
P5 - AP2	PCR prod. of (6)	not amplified
P5 - AP2	PCR prod. of (7)	not amplified
P5 - AP2	PCR prod. of (8)	not amplified
P6 - AP1	D-cDNA	not amplified (9)
P6 - AP1	PCR prod. of (5)	not amplified
P6 - AP1	PCR prod. of (7)	not amplified
P6 - AP2	PCR prod. of (5)	not amplified
P6 - AP2	PCR prod. of (7)	not amplified
P6 - AP2	PCR prod. of (9)	not amplified
P4 - PT	S-cDNA	not amplified (10)
P5 - PT	PCR prod. of (10)	2.8, 3.3
P5 - PT	S-cDNA	2.8, 3.3 (11)
P6 - PT	PCR prod. of (11)	2.8, 3.3
P6 - PT	S-cDNA	2.8, 3.3

Table 3

	MSP-1	Soluble Fraction	
Amino Acid	Scallop	Scallop	Oyster
Asx	27.5	25.2	31.5
Asp	20.1		
Asn	7.4		
Thr	1.0	1.6	0.5
Ser	32.3	22.4	28.3
Glx	3.2	6.0	4.2
Glu	2.7		
Gln	0.5		
Pro	1.5	2.4	0.9
Gly	25.0	26.3	30.0
Ala	4.2	6.5	0.9
Cys	0.0	1.4	ND
Val	0.1	1.1	0.2
Met	0.0	0.0	ND
Ile	0.0	0.7	0.1
Leu	1.2	1.6	0.1
Tyr	0.5	0.0	1.7
Phe	0.1	0.6	0.1
Lys	2.1	3.4	0.6
Arg	1.2	0.9	0.3
His	0.1	0.7	ND
Trp	0.0	ND	ND

Kodak Color Control Patches
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
© Kodak, 2007 TM: Kodak
inches
cm
Kodak Gray Scale
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19
C
Y
M
© Kodak, 2007 TM: Kodak