CHEMICAL REVIEWS

Review

"Tuning in" to Mollusk Shell Nacre- and Prismatic-Associated Protein Terminal Sequences. Implications for Biomineralization and the Construction of High Performance Inorganic#Organic Composites

John Spencer Evans

Chem. Rev., **2008**, 108 (11), 4455-4462 • DOI: 10.1021/cr078251e • Publication Date (Web): 13 September 2008

"Tuning in" to Mollusk Shell Nacre- and Prismatic-Associated Protein Terminal Sequences. Implications for Biomineralization and the Construction of High Performance Inorganic–Organic Composites

John Spencer Evans*

Laboratory for Chemical Physics, Center for Biomolecular Materials Spectroscopy, New York University, 345 E. 24th Street, Room 1007, New York, New York, 10010

Received December 11, 2007

Contents

1. Introduction

The field of biomineralization has expanded over the past decade, and there is now a growing database of information on the molecular participants in nucleation and crystal growth processes in organisms. In particular, we now know that proteins play key roles in nucleation and crystal growth of inorganic solids in Nature.[1−25] Using biomineralization proteins as inspiration, the material science and nanotechnology fields have now embarked upon the use of phage and bacterial polypeptide screening and expression systems directed against biological and artificial inorganic materials.[26−33] This has led to the development of polypeptide sequence libraries which can selective bind, interact, and assemble inorganic solids, as well as regulate the nucleation of these inorganic solids in solution.[26−33]

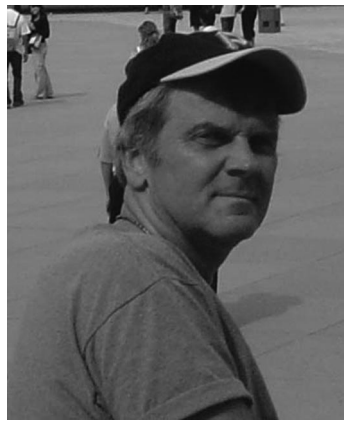
Although these bioinspired laboratory achievements may eventually usher in a new era in materials science and technology, we have only really "leapfrogged" over a major problem. The problem is that we lack a full understanding of how biological organisms developed and exploited polypeptide sequences for associations with minerals. We need to investigate protein-mediated biomineralization in sufficient detail in order to develop sophisticated methods for utilizing proteins and polymer mimics to engineer inorganic materials. The most important step in this journey will begin with the identification and exploration of model biomineralization systems and learning the molecular rules that are at work in these systems. Subsequently, the next step will involve the successful translation of these rules to the laboratory frame where new polypeptides or polymers can be tailored or designed for specific applications. Obviously, the choice of a model system for learning these principles becomes an important factor in the eventual success of this undertaking.

1.1. The Mollusk Shell as a Model System

A relevant model biomineralization system can be found protecting and enhancing the survival of a particular invertebrate marine organism, the mollusk. The mollusk shell is a true composite, consisting of calcium carbonate mineral associated with biopolymers, including lipid assemblies, polysaccharides, and proteins.[34−41] In some mollusks, the evolution of the shell has led to the development of a two-layer composite material, with each layer composed of different polymorphic forms of calcium carbonate (prismatic layer = calcite,[37−41] nacre layer = aragonite[34−36,38,39]), and this is truly one of the most fascinating aspects of the mollusk model system. As shown in Figure 1, both the prismatic and nacre layers contain a number of proteins, some of which are intracrystalline, i.e., incorporated within the inorganic phase, and some which form biofilms around the mineral phase.[37−43] It is now accepted that the mollusk initiates shell construction using amorphous calcium carbonate (ACC) as the starting point for the inorganic phase,[44−46] and this ACC phase eventually transforms to either calcite or aragonite.[44−46] These transformative processes are believed to be guided in part by proteins,[44−46] but how this is accomplished is not yet known. The final composite structures which are formed

* To whom reprints requests should be addressed. Telephone: 2129989605. Fax: 2129954087. E-mail: jse1@nyu.edu.

10.1021/cr078251e CCC: $71.00
Published on Web 09/13/2008



John Spencer Evans is a Professor of Chemistry and Biochemistry at New York University. His research is in the area of protein-mediated biomineralization and genetically engineered materials, with an emphasis on solution and solid-state nuclear magnetic resonance, biophysical techiques, and molecular simulations. He received his Ph.D. degree from the California Institute of Technology, his D.D.S. degree from the University of Illinois, and his Bachelor's degree from Northwestern University.

are very attractive models from a materials standpoint: the prismatic layer is resistant to crack propagation and puncture,[37−39] whereas the nacre layer is more ductile, possesses residual stress, and is fracture resistant (Figure 1).[34−36] Put these elements together and the net result is a highly protective material that can be repaired by the organism if damage occurs.[38] For these reasons, a substantial amount of research has been conducted on the physical and chemical properties of mollusk shells, and this work has yielded a solid database for bioinspired materials research. But we still lack a fundamental understanding of the participation of proteins in mollusk shell nacre and prismatic layer formation. In turn, we are not fully aware of all the biomolecular principles that can be transferred to the laboratory frame for designing new inorganic-associated proteins and polymers.

1.2. Approach and Rationale for Primary Sequence Comparisons

The purpose of this review is to take the first step toward understanding protein-mediated biomineralization processes. To do this, we have to start with the proteins themselves and identify unique molecular traits that functionally distinguish these proteins from all others in the existing protein databases. To limit our focus, we will concentrate our efforts on the proteins of the mollusk biomineralization system, since this model system holds strong potential for the materials sciences and solid-state chemistry communities. Therefore, the main focus of this review will be the comparative analysis of primary sequence traits in mollusk prismatic-[8,9,11−13,42,43] and nacre-[3,4,7,10,14−18,20,21] associated proteins. Here, we argue that the most important molecular traits are encoded within the primary sequence: it dictates not only the linear ordering and choice of chemical groups (i.e., side chains) but also the configuration of these residues in two and three dimensions and the resulting molecular surfaces and electrostatic charges that these residues contribute to the protein. Collectively, these elements establish protein function. For this reason, a comprehensive comparative review of the recent primary sequence entries into the mollusk protein sequence database will ultimately be of great use in understanding the role of biomineralzation proteins in crystal growth, fracture resistance, polymorph selection, and resistance to crack propagation.

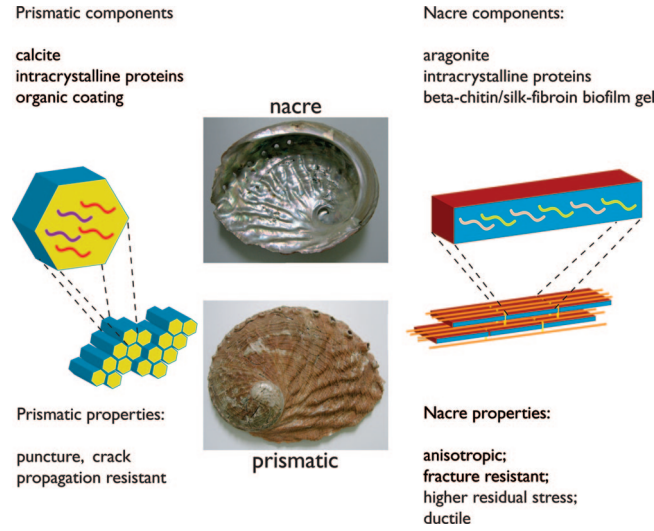


Figure 1. Construction and engineering scheme of mollusk shell layers. The prismatic (outer) and nacreous (inner) regions of the Pacific Red abalone (Haliotis rufescens) shell are presented for visual comparison. The photos at the center of the figure show the macroscopic view of the nacre and prismatic surfaces (note that the prismatic side of the shell has a reddish-stained organic covering that obscures direct visualization of the prismatic mineral layer). The schematics on the right- and left-hand sides provide a micro- and nanolevel cross section representation of the prismatic and nacre architecture along with associated organic components. For the nacre layer, the mineral tablets of aragonite (blue) are surrounded by organic biofilms that are comprised of proteins (red) and polysaccharides (orange); this Biofilm also runs laterally and vertically throughout the mineral phase. Conversely, the mineral phase of the prismatic layer consists of hexagonal calcite rods (yellow) that are shrouded by a protein layer (blue). Note that the mineral phases of both layers contain proteins that are entrapped or occluded in the mineral; these are referred to as intracrystalline proteins. Finally, the known published mechanical properties are listed for both layers.

2. Intended Scope of Sequence Analysis

In the sections that follow, our comparisons will be confined to recent studies (2001−2007) which report either a complete cDNA sequence or a partial sequence determined by protein sequencing methods.[3−21,42,43] The present goal is to contrast the similarities and differences that exist within the nacre- and prismatic-associated sequences classes. From the studies that have been conducted on calcite blocker,[3,4,22,23] aragonite-promoting,[22,25] and calcite morphology modulating[24] mollusk shell-associated protein sequences, it appears that the N- and/or C-terminal regions of these proteins represent putative sites for mineral interaction and/or nucleation. Hence, we will extend our analyses to the terminal sequences of other nacre- and prismatic-associated proteins, with the goal of comparing and contrasting their linear arrangements and amino acid compositions. However, we alert the reader to the fact that other internal regions of these proteins may also be involved in mineralization processes and deserve future consideration.

3. Nacre-Associated Sequences

Table 1 presents candidate nacre-specific proteins and their terminal sequences.[3,4,7,10,14−18,20−25] As one can see, the known in vitro functions of these nacre proteins have been

Table 1. 30 AA Nacre Protein Terminal Sequences[a]

Protein	Species	Function	N or C	Sequence	Net Charge	Name	Ref
AP7	H. *rufescens*	blocker	N	DDNGNYGNGMASVRTQGNTYDDLASLISYL...	-3	AP7N	3
AP24	H. *rufescens*	blocker	N	ADDDEDASSGLCNQYNQNVTTRPNNKPKMF...	-2	AP24N	3
n16	P. *fucata*	promoter	N	AYHKKCGRYSYCWIPYDIERDRYDNGDKKC...	+3	n16N	16
			C	...GLNYLKSLYGGYGNGNGEFWEEYIDERTDN	-4	n16C	16
CaLP	P. *fucata*	promoter	N	EAFHLFDKDGSGSISAEELGTVMRSLGQNP...	-2	CaLPN	15
			C	...DGDGEINYEEFVKMISMKDTDQEEQQQENK	-7	CaLPC	15
n19	P. *fucata*	blocker	N	QLEAVPCDDKSTPYNDDGNGDMAFLDRHKL...	-3	n19N	7
			C	...CASGYFINKFSLQADYNWYQIRYQYRCCKP	+3	n19C	7
PFMG1	P. *fucata*	promoter	N	VCDARKRWWRRATKKVELKPEIGRGGDWKV...	+6	PRMG1N	20
			C	...FSEADENDDEQISTSEFKDFKSRINQCVKD	-5	PRMG1C	20
ACCBP	P. *fucata*	blocker	N	CDYPEAKLLKFLLDDYEKLVRPVPKDGNAT...	-1	ACCBPN	14
			C	...SNINSRFYDCCPEPYDIEFNINIEHNDKDK	-3	ACCBPC	14
Perlucin	H. *laevigata*	promoter	N	GCPLGFHQNRRSCYWFSTIKSSFAEAAGYC...	+3	perlucN	10
Perlustrin	H. *laevigata*	promoter	N	LSCASCENAACPAIGLPCKPSEYVYTPCGC...	-1	perlusN	21
Perlwapin	H. *laevigata*	blocker	N	YGPNLPGCPPGPYPRICARYCHSDRECKAG...	+3	perlwapN	18
			C	...VIYCVGDFDCPGNEKCCGSCPRRCEKPCFD	-1	perlwapC	18
Perlinhibin	H. *laevigata*	blocker	N	ECTIGDSCVVHRHRECRCPRGRAMCDREHHK...	+6	perlinN	17

[a] The individual single-letter amino acid codes are employed, and different residue types are denoted in color (red = anionic; blue = cationic; green = HBDA; Cys, Pro, Gly are underlined). The functions of each protein are based upon reported in vitro experiments: that is, "blocker" refers to the ability to limit or frustrate crystal growth; "promoter" refers to the ability to enhance crystal growth.[3,7,14−18,20,21] The name given to each N- and C-terminal sequence region follows the standard set in previous publications of AP7, AP24, and n16, that is, an abbreviated form of the protein name, followed by "N" if the given sequence is at the N-terminus and "C" if it is located at the C-terminus.[3,22,23] The ellipse (...) is used to indicate that the sequence continues either upstream (i.e., an N-terminal sequence) or downstream (C-terminal sequence) of each respective fragment. The net charge category includes contributions arising from all side chains at pH 7.5 but does not factor in the positive or negative charge of the terminal N-α amine or C-α carboxylate group, respectively, of the polypeptide chain.

divided into two categories: "blockers", which have been reported to inhibit crystal growth, and "promoters", which have been reported to accentuate crystal growth. For comparative purposes, we have limited out attention to 30 AA sequence lengths at either termini, due to the fact that the best-known N- and C-terminal sequences from AP7, AP24, and n16 were created as synthetic peptides and arbitrarily limited to 30 AA to facilitate solid-phase synthesis and postsynthesis yields.[3,4,22,23,25] Obviously, the actual sequence length(s) of mineral modification domains within mollusk shell proteins may be shorter or longer than this arbitrary limit, and the reader is urged to examine the published nacre protein sequences for additional information.

3.1. Hydrophilic and Charged Clusters

Many of the nacre protein terminal sequences possess short clusters (2−3 contiguous residues) of anionic (Asp, Glu), cationic (Lys, His, Arg), or hydrogen-bonding donor/acceptor (HBDA) (Ser, Thr, Tyr, Gln, Asn) residues (Table 1). The HBDA class is believed to be involved in interactions with mineral-associated water clusters during the crystal growth process.[22,23] The presence of anionic, cationic, and HBDA residues predominates most nacre protein terminal sequence compositions. The exceptions to this are AP24N, which possesses a pentaanionic N-terminal cluster, and perlucN, perlusN, and perlwapN, which possess Pro-, Gly-, Cys-containing sequence blocks. In general, the occurrence of Asp and Glu in the nacre terminal sequences ranges from 3:2 to nearly 1:1 Asp/Glu; that is, the ratio approaches unity. With an approximate balance of anionic, cation, and HBDA residues per sequence, we would expect each sequence to be capable of binding to Ca(II), CO_2^{-3} ions, and water molecules in similar proportions, either in solution or at the mineral interface.[22,23] The ratio of cationic to anionic residues appears to vary within the nacre terminal sequence ensemble, and this may have functional implications, for as one can see in Table 1, the net electrostatic charge of nacre terminal sequences can either be positive or negative. Moreover, as shown in Table 1, the magnitude of the net electrostatic charge also varies considerably. Thus, variation in charge sign and magnitude are key descriptors of these nacre terminal polypeptide sequences.

Another important feature of the nacre ensemble is the positioning of cationic and anionic amino acids relative to one another (Table 1). In general, we note the intermixing of oppositely charged residues. This situation creates a heterogeneous (−), (+) charged surface along the length of each terminal sequence.[24] An examination of Table 1 reveals that the length spacing of charged clusters varies in each terminal sequence and even within the N- and C-terminal sequences of the same nacre protein. A similar charge heterogeneity is also observed for the placement and location of HBDA residues. Interestingly, these phenomena occur to a lesser degree in the *H. Laevigata* nacre sequences. Thus, rather than segregate residues of like charge or hydrogen-bonding capabilities away from one another, the nacre terminal sequences evolved an integrated usage of (−), (+), and HBDA residue types together. This may be important for the function of these nacre protein sequences.

3.2. Other Residues and Their Significance

Another area of interest is the content of Gly, Pro, and Cys within the nacre terminal sequence ensemble. AP7N, n16C, and CaLPN possess four or more Gly residues per 30 AA (Table 1); Gly is typically associated with the presence of loop structures and is typically located within sequence regions that have a substantial degree of chain motion.[47−49] Conversely, the perlusN and perlwapN terminal sequences contain four or more Pro residues per 30 AA, and it is known that the Pro imido ring introduces a kink or bend in a protein sequence and a more limited degree of motion.[50−52] Finally, the terminal sequences n16N, perlusN, perlwapN, perlwapC, and perlinC possess three or more Cys residues, an occurrence that has interesting implications in terms of establishing intrachain association (i.e., disulfide bonds) and folding. In the case of the *H. laevigata* nacre proteins, disulfide bonds are reported to exist within these nacre-associated sequences;[17,18] however, in other proteins, such as n16, Cys residues are evident, but disulfide bonds have not yet been identified, and thus the presence of Cys within this sequence is not clearly

understood.[16] Overall, the occurrence of Pro, Gly, and Cys may have evolved as a means of "tuning" the conformation, motion, and stability of each sequence region. This, in turn, may have "tuned" the functionality of these regions.

3.3. Structural Features of Nacre Protein Terminal Sequences

To date, very few structural studies have been conducted on nacre-specific polypeptides, and the same could be said for biomineralization proteins in general. Typically, this problem is due to either poor protein yields from nacre tissue or insufficient expression of recombinant nacre proteins for structural studies.

However, a number of studies have been conducted with synthetically created N- and C-terminal fragments of nacre proteins, AP7,[23,51,52] AP24,[23,52] and n16,[25] and from these studies we can glean some important structural insights. First, the 30 AA terminal sequences of n16, AP7, and AP24 all exist as conformationally unfolded species in solution, with some interesting differences.[23,25,51,52] For example, n16N and n16C demonstrate some degree of random-coil and β strand conformation,[25] whereas AP7N and AP24N both possess some degree of random coil-like conformation that exists in an equilibrium with other secondary structures, such as α helix.[23,51,53,54] These nacre-associated polypeptides are structurally unstable and would require external forces to stabilize their internal structures.[23,25,53,54] This has been recently demonstrated in 2,2,2-trifluoroethanol solvent induction studies, where n16N, AP7N, and AP24N form folded structures in the presence of this solvent.[55] By analogy, other external influences such as ionic clusters or mineral surfaces may also act as polypeptide folding "inductants" that lead to the structural stabilization of labile polypeptides during the mineralization process. External stabilization compensates for the lack of internal polypeptide stabilization,[22,23,55] and this entropically driven process achieves internal polypeptide ordering. Thus, conformational transformation or reordering appears to be an integral feature of the terminal sequences derived from AP7, AP24, and n16, and it would be interesting to learn if the same holds true for the other nacre terminal sequences of Table 1.

Although the structures of the other nacre-associated polypeptides are not known, we speculate that this ensemble of polypeptides will most likely adopt unfolded, conformationally labile structures in solution as well and exhibit the capabilities of conformational transformation in the presence of external forces. Structurally, N-terminal nacre sequences that feature a high degree of surface charge and/or the presence of Pro, Ala, Gln residues would most likely adopt some precentage of polyproline type II (PPII) secondary structures.[6,56−58] The PPII secondary structure is an extended helix,[6,56−58] and this molecular surface promotes interactions with other surfaces.[6,56−58] Conversely, those N-terminal sequences with significant Gly content would be expected to form random coil or looplike structures in solution.[47−49] Given that the development of the nacre layer initiates from a disordered, amorphous phase,[44−46] it may be that conformationally labile, unfolded polypeptides evolved to allow adaptation at disordered interfaces and/or to transport ion clusters to these interfaces. Obviously, other secondary structures may also exist within this sequence ensemble (Table 1) and further research will be needed to correlate polypeptide structure with functional attributes.

3.4. Expected Nacre Polypeptide Surface Topologies

Planar surfaces have been shown to be important for macromolecule-directed nucleation and inhibition of crystal growth.[23,55] n16N, AP7N, and AP24N adopt some degree of planar structure in solution, with n16N and AP24N demonstrating a higher degree of planarity.[23,55] The extent of polypeptide surface planarity may be important, in that both AP7N and AP24N induce the same in vitro calcium carbonate growth phenomena, but with different kinetics: AP24N exhibits faster rates of nucleation, obtuse step acceleration, and acute step inhibition than AP7N.[22,23] Hence, there may be a link between protein surface topology and mineralization function and kinetics, and this adds another layer of complexity and control to the protein-mediated process of aragonite formation and stabilization in the nacre layer.

Based upon amino acid content, we can make some qualitative predictions of the molecular surfaces of the nacre sequences in Table 1. In particular, the presence of dispersed (−), (+), and HBDA regions would lead to the formation of heterogeneously distributed charges and hydrogen bonding sites over most of these terminal sequences. Heterogeneous charge spacing could be important for coassembling Ca (II), CO_2^{-3} ionic clusters on the surface of these sequences and/or for recognizing Ca, CO_2^{-3} clusters at the mineral interface, particularly for a disordered precursor phase such as ACC.[42−44] Similarly, the heterogeneity in HBDA cluster region location and spacing (Table 1) may also be an important feature for protein-mediated binding of water molecules at mineral interfaces or with solvated ion clusters.[22,23] The fact that charge and HBDA spacing varies from sequence to sequence (Table 1) suggests that these features "tune" the function(s) of the sequences. As an example, some sequences may utilize one type of charge/HBDA spacing to create molecular surfaces that are adept at blocking calcite growth, whereas other sequences may employ different charge/HBDA spacings that generate molecular surfaces that are more suited to promoting nucleation, phase stabilization, or other processes that are important to nacre formation (Table 1).

4. Prismatic-Associated Sequences

Table 2 presents primary terminal sequences obtained for prismatic associated mollusk shell proteins.[8,9,11−13,42,43] This sequence list is not as extensive as that presented in Table 1, and this is primarily due to difficulties in obtaining complete cDNA-derived sequences for these proteins, some of which are highly redundant in amino acid usage and are polyelectrolytes.[8,9,11−13,42,43] The known in vitro functions of these prismatic proteins can be grouped as follows: blocker, promoter, and a third category, morphology, which indicates that crystal shape is affected.[8,9,11−13,42,43] At this time, the function of the scallop shell MSP isoform proteins (henceforth referred to as MSP1,2) is not known. Given that the first terminal sequence studies of Asprich were conducted with approximately 40 AA long synthetic peptides,[24] we have chosen to use this length for defining other prismatic sequences. Obviously, many of the conclusions that are reached regarding this limited set of sequences will require reassessment once other prismatic sequences become available and molecular trends begin to emerge from the larger pool of published sequences.

Table 2. 40 AA Prismatic Protein Terminal Sequences[a]

Protein	Species	Function	N or C	Sequence	Net Charge	Name	Ref
Asprich	A. *rigida*	shape	N	KPVFKRSLSDPSDDGGANDVADDVEADAADLEEDVDQDVD...	-12	Fragment-1	11
			C	...ADEADADEADADEADADNDAADETDAADVGTEAEDVADDE	-18	Fragment-2	11
Aspein	P. *fucata*	promoter	N	FPVADQTTNELGSSGAAAAGAVVSEPSDAGDAADAGDADA...	-8	aspeinN	13
			C	...DSGDDDDDDDDDDGDDGDDDSGDDDGDDSDDDDDDDDDQ	-31	aspeinC	13
Prismalin14	P. *fucata*	inhibitor	N	QYFFRGGDDDNGFFGGDDDNGYFGYFPRFSYPIYRPIYRI...	-2	pr14N	12
			C	...GLYGGYGGFGGYGYRPFSYGYNPFSYGYYGFGDDDGGFDD	-5	pr14C	12
Calprismin	P. *nobilis*	shape	N	ITIEQGELVLNCISDKMMCEFEAAAEMVVPSPCPSEYDTC...	-7	calpN	9
MPSP	C. *gigas*	promoter	N	TADGDDDDDEVDATLYENILGRDSVHDDIHESTICVDDAD...	-12	mpspN	8
			C	...DDSNVATRNSVAAKYDGIVTFDDLRVGEVGSEDGDDGKPV	-4	mpspC	8
MSP1, 2	*P. yessoensis*	unknown	N	LDTDKDLEFLDSLLNAAEDGGGGDAAGAEEAAPAADLSGG...	-10	msp1N	42,43
			C	...SGSDSGSSEESDSGSDSSSGSSGGDDADSSSSSSSSEEEA	-11	msp1C	42,43

[a] The individual single-letter amino acid codes are employed, and we have denoted different residue types in color (red = anionic; blue = cationic; green = HBDA; Cys, Pro, Gly are underlined). The functional classification of each protein is based upon reported in vitro experiments; that is, "blocker" refers to the ability to limit or frustrate crystal growth; "promoter" refers to the ability to enhance crystal growth, and "shape" refers to the ability of a polypeptide to affect cyrstal morphologies.[8,9,11−13] With the exception of Asprich, the name given to each N- and C-terminal sequence region follows the standard set in previous publications of AP7, AP24, and n16, i.e., an abbreviated form of the protein name, followed by "N" if the given sequence is at the N-terminus and "C" if it is located at the C-terminus.[3,22,23] The ellipse (...) is used to indicate that the sequence continues either upstream (i.e., an N-terminal sequence) or downstream (a C-terminal sequence) of each respective fragment. The net charge category includes contributions arising from all side chains at pH 7.5 but does not factor in the positive or negative charge of the terminal N-α amine or C-α carboxylate group, respectively, of the poypeptide chain. Note that MSP1,2 represents the identical terminal sequences of the mollusk shell protein (MSP) 1 and 2 isoforms found in the prismatic layer of the scallop, *P. yesoensis*.[42,43]

4.1. Hydrophilic and Charged Clusters

The most striking feature of the prismatic terminal sequence series is the predominance of anionic residues. The content of Asp and Glu residues is at least 10% of the total amino acid composition in MSP1,2, CalpN, Pr14N, and Pr14C and reaches 80% in aspeinC (Table 2). Anionic amino acids typically exist as contiguous homogeneous clusters of two or more residues in these sequences. There is a definite trend toward Asp usage, with Asp/Glu ratios ranging from 1:1 to 6:1, and it is interesting to note that Glu is absent from aspeinC, Pr14N, and Pr14C sequences. The high usage of anionic residues results in the generation of net negative charges for all prismatic terminal sequences (Table 2). Thus, variation in charge magnitude but not sign is one of the major descriptors of the prismatic series (Table 1).

Another feature of interest is the occurrence of cationic amino acids.[8,9,11−13] Positively charged residues are clearly underrepresented in most of the prismatic sequence ensemble (Table 2), and in the case of the aspeinN, aspeinC, and msp1C terminal fragments, they are totally absent. Interestingly, where cationic residues do occur, they appear to follow one of two placement rules: they are either segregated from anionic residues (e.g., Fragment-1, Pr14N, Pr14C), or they occur as a single residue within close proximity to an anionic residue (e.g., CalpN, mpspN, mpspC, msp1N). In the case of Pr14N and Fragment-1, the segregation of positively charged residues creates a cationic region at one end of the terminal sequence (Table 2). These features are yet another unique trait of the prismatic ensemble that are not typically observed in the nacre terminal sequences (Table 1).

Finally, we note that prismatic sequences exhibit variations in HBDA content.[8,9,11−13] As shown in Table 2, HBDA residues are relatively scarce in the terminal sequences of Aspein, Asprich, and the N-terminal region of MSP1,2. However, they are present in the remaining members of the ensemble. Interestingly, where HBDA residues do occur in prismatic sequences, the HBDA cluster size is approximately the same as that found in the majority of nacre terminal sequences (i.e., 2−3 amino acids, Tables 1 and 2). This suggests that there is an optimal HBDA cluster size for both nacre and prismatic protein-mediated crystal growth and nucleation processes. An exception to this is the msp1C fragment (Table 2), which contains -SS-, -SSS-, and -$(S)_8$- clusters. Although it is not clear why these HBDA clusters are larger than normal in the msp1C fragment, it has been suggested that these Ser clusters are either part of a "serine loop" structure motif or represent potential sites for phosphorylation.[42,43]

4.2. Other Residues and Their Significance

The prismatic sequence ensemble has a low Cys content,[8,9,11−13,42,43] where one or two Cys residues/40 AA are generally noted. In the terminal sequences of MSP1,2, Asprich, Aspein, and Prismalin-14, Cys is absent. Thus, the presence of disulfide linkages within prismatic sequences would be expected to be low, in stark contrast to the case for nacre-associated *H. laevigata* nacre sequences, which are reported to possess disulfide linkages (Table 1).[17,18] As we shall discuss later on, this may have implications for the secondary structures and surface topologies of the prismatic terminal sequences.

There are significant variations in the Gly and Pro content within the prismatic associated terminal sequences (Table 2). There are those terminal sequences which are deficient in these residues (e.g., Fragment-1, Fragment-2, aspeinC, mpspN) and those which possess significant percentages (e.g., msp1N, msp1C, aspeinN, Pr14N, Pr14C, CalpN). In general, Gly residues promote the formation of looplike regions,[45−48] and Pro residues promote kink or bend regions in the polypeptide chain.[50−52] Beyond this generalization, there are a few unusual examples of Gly and Pro occurrence within the prismatic sequence ensemble. The first is Prismalin14, which utilizes Pro and Gly in different terminal regions: Pr14N contains a Pro-repeat, -PIYR-, and Pr14C contains a -GGY- repeat. The second example is MSP1,2, where the N- and C-terminal sequences feature tetraGly sequence blocks (Table 2), which have been hypothesized to function as "glycine loop" structural motifs.[42,43]

Of additional interest is the usage of hydrophobic Ala and Val residues within the prismatic sequence ensemble. In those sequences which have a high percentage of Asp or Glu (MSP1,2, Asprich, Aspein), Ala and Val tend to predominate

in certain sequence blocks (Table 2). This phenomenon is not observed within the nacre sequence ensemble (Table 1), where larger, sterically bulky side chains, such as Leu, Ile, Phe, and Trp, are typically featured. Given that Ala and Val are smaller in their van der Waals (vdW) volumes compared to Ile, Leu, and aromatic ring residues, there may be a functional advantage for polyanionic prismatic sequences to accommodate smaller vdW volume neutral hydrophobic side chains such as Ala and Val.

4.3. Structural Features of Prismatic Protein Terminal Sequences

Unfortunately, there is little data available for prismatic protein sequences and their structural preferences. As of this writing, only the Asprich N- and C-terminal sequences have been examined for conformational preferences, and this polyanionic sequence has been observed to form a combination of random coil and PPII structure in solution.[24,59] Nonetheless, we can predict some general structural trends based upon amino acid usage within the prismatic ensemble (Table 2). It has been observed that polyelectrolyte polypeptides do exhibit a preference for PPII extended helical structure.[57,60] That being the case, it is likely that other polyelectrolyte prismatic terminal sequences, such as Fragment-1, aspeinN, aspeinC, Pr14N, Pr14C, msp1N, msp1C, mpspN, and mpspC also possess some mixture of RC and PPII structural content. As mentioned in the preceding section, the Gly-rich Pr14C and msp1N, 1C sequences may possess some degree of loop structure as a result of the placement of numerous Gly residues along its length.[42,43,47−49] Pro-rich regions also exist within CalpN and Pr14N, and the presence of Pro may give rise to either PPII[53,56−58,60] or turnlike structures[48−50] within these sequences. The presence of RC, turnlike, loop, and PPII structures may convey functional advantages to prismatic terminal sequences within the framework of prismatic calcite formation.

From the foregoing, the prismatic terminal sequence library would be expected to adopt unfolded, conformationally labile structures in solution. As discussed earlier for their nacre counterparts, prismatic sequences are most likely inherently unstable and require external interactions to stabilize their internal structures.[23,25,53,54] This inherent instability would act as an entropic driving force for polypeptide associations with ion clusters or mineral surfaces. These external associations lead to internal stabilization of the polypeptide sequence. Thus, both prismatic- and nacre-associated terminal sequences would be expected to employ conformational transformation or entropy-based reordering events to achieve their functional goals within their respective mineralization environments.

4.4. Expected Prismatic Polypeptide Surface Topologies

Given the amino acid composition and linear arrangments of anionic and cationic amino acids (Table 2), the molecular surfaces of the prismatic associate sequences will be polyelectolyte in nature and possess sequence cluster regions which are exclusively (−) or (+) in nature. Where polyanionic regions predominate, we would expect these sequence regions to be adept at condensing Ca(II) ions at their surfaces. By the same token, where we find cationic and HBDA sequence clusters, we would expect CO_3^{2-} and water interactions to be fostered in these regions. Given that polyanionic sequence regions predominate in most of the prismatic terminal sequences (Table 2), it follows that sequence regions for CO_3^{2-} and water binding are somewhat limited compared to Ca(II) ion binding regions. How all of this plays out in solution and on mineral surfaces is not clear, but we would anticipate that prismatic terminal sequences may preferentially recognize and bind to mineral surfaces where Ca(II) ions are featured, with some secondary interactions directed by the presence of CO_3^{2-}. Likewise, the stoichiometry of ion cluster binding by a given prismatic sequence would appear to be skewed toward Ca(II) content over CO_3^{-2}.

5. Summary and Conclusions

5.1. The Biological Perspective

It is clear that nacre- and prismatic-associated terminal sequences exhibit important similarities and differences (Tables 1 and 2). The major similarity for both classes is structural. Prior to their interaction with mineral surfaces and/or ionic and water clusters, we would predict that the majority of the nacre and prismatic terminal sequences will exist as unfolded, conformationally labile species in solution. However, protein−mineral and protein−cluster interaction processes would initiate structural reordering to some degree, and then conformationally unstable regions of each sequence would become stabilized. This hypothesis needs to be experimentally tested, and it is hoped that in the near future methods and suitable mollusk shell biomineralization protein candidates will become available for this purpose.

What appears to distinguish prismatic from nacre sequences are two important features: electrostatics and the usage of Pro, Gly, and Cys. We will discuss electrostatics first. In general, the nacre sequences possess a heterogeneously distributed mixture of anionic and cationic residues that gives rise to either a net (+) or (−) charge on the corresponding polypeptide (Table 1). Thus, compared to prismatic sequences, nacre terminal sequences will possess molecular surfaces with a heterogeneous charge distribution of (+) and (−). In comparison, the prismatic sequence library possesses multiple anionic residues and net anionic charge. When cationic residues do occur in these prismatic sequences, they tend to be segregated from anionic residues. As a consequence, prismatic terminal sequences have a greater degree of (+), (−) charge separation and the resulting molecular surfaces will reflect this. These differences in charge distribution will have an impact on polypeptide−mineral surface recognition and possibly define the potential sites on the mineral surface where nacre or prismatic terminal sequences preferentially interact. For similar reasons, ionic clustering on polypeptide molecular surfaces would also be expected to be different for nacre and prismatic terminal sequences. In particular, those prismatic terminal sequences which are polyelectrolyte in nature will foster significant counterion condensation at their molecular surfaces and offer homogeneous multiple sites for mineral surface or ion interactions, compared to their nacre counterparts. Collectively, these traits can differentially "tune" the recognition, interaction, and specificity of nacre and prismatic sequence binding to mineral surfaces and influence the stoichiometry (i.e., Ca(II)/CO_2^{3-}) of bound ions on each type of sequence class.

Let us now consider the occurrence of Cys, Pro, and Gly in both terminal sequence ensembles and their potential

impact. Compared to prismatic-specific sequences, nacre-specific terminal sequences utilize Cys residues to a greater extent. Although the significance of Cys is not completely understood, reports of internal disulfide linkages and their polypeptide folding implications[17,18] suggest that Cys may play an important role in defining the conformation, stability, and function of nacre-specific terminal sequences. Conversely, the fact that prismatic sequences are deficient in Cys suggests that disulfide linkages are rare within the prismatic ensemble and/or that Cys is not a significant requirement for either the functional or structural aspects of these sequences. Gly and Pro are found in both nacre and prismatic terminal sequences, suggesting that there is a need to "tune" the molecular motion or conformation within both types of sequences, as well as introduce loop- $(Gly)^{47-49}$ or PPII-like (Pro, Ala, Gln)[50−52,56−60] structural regions for functionality. With the exception of the prismatic-associated MSP1,2 protein family, there appears to be a higher occurrence of Gly and/or Pro in the nacre sequence ensemble, which would indicate that nacre terminal sequences have a greater reliance on these residues for functional purposes.

It is hoped that this review of mollusk shell terminal sequences will encourage further scientific investigations of the structural and functional aspects of these sequences and the parent proteins from which they are derived. Obviously, nonterminal sequence regions have been excluded from consideration in this review, but it would be foolish to suggest that these regions do not play a role in the mollusk shell mineralization process on some level. Over time, we hope that comparative sequence studies and structure/function investigations will be conducted on these other sequence regions to provide a more comprehensive view.

5.2. Implications for Protein-Based Genetically Engineered Materials

It is not too early to begin the process of exploiting the molecular features of mollusk protein sequences for inorganic materials design, processing, and synthesis. These studies would utilize either genetically expressed or chemically synthesized polypeptides or protein-like polymers that mimic the molecular qualities listed below and elsewhere.[26−33,55]

5.2.1. Conformational Instability

The fact that biomineralization sequences exist as unfolded species in solution and lack internal stabilization could be used a starting point in the design of inorganic-associated proteins and polymers. Rather than rely on polypeptide sequences that already adopt stable secondary structures such as α helix and β hairpin, it would seem more advantageous to either start with an unstable sequence or take a folded sequence and introduce destabilizing residues (e.g., Gly, Pro), such that this modified sequence becomes unstable. With regard to Cys−Cys disulfide bonding, some of the nacre sequences appear to utilize this bonding to establish points of intrachain contact, which may have an impact on chain motion or overall chain configuration. Again, this may be of use in "tuning" the conformation of mimics for material surface interaction and the nucleation of inorganic solids.

5.2.2. Interactive Residues

There is a tendency within the biomineralization community to think of "interactive" residues as those with anionic charge (i.e., Glu, Asp, phosphorylated Ser, and Thr). However, this thinking is too restrictive and ignores the contributions of cationic and HBDA residues, both of which play an important role in surface interactions and nucleation. Hence, in order to achieve desired and enhanced interactions with material surfaces and nucleation foci, it is recommended that true biomineralization mimics need to incorporate anionic, cationic, and HBDA residues in proper positions and ratios. Moreover, biomineralization proteins utilize terminal sequence regions for mineral interaction and modification, and thus protein and polymer mimics of these sequences should employ terminal segments for interaction as well, with other internal segments to be utilized for other purposes.

5.2.3. Polypeptide Surface Design

Recent studies indicate that planar surfaces on polypeptides appear to be important for nucleation and polypeptide−surface interactions.[23,55] Thus, engineering planar regions into proteins or polymers may lead to enhancement of interactions with material surfaces or participation in nucleation processes. Moreover, we have seen how dispersal or segregation of (−), (+) charge residues can exist in a series of mineral-associated polypeptides (Tables 1 and 2). It would be interesting to utilize these principles of charge integration or separation in the design of new proteins and polymers whose tasks would involve interaction with inorganic solids and/or participation in nucleation. In some instances, polyelectrolyte domains may be important choices for facilitating counterion condensation or "tuning" interactions between polypeptides and specific interfaces.

6. Acknowledgments

This work was supported by funding from the Department of Energy (DE-FG02-03ER46099) and the National Science Foundation (DMR-0704148). We thank the "Oracle" for helpful discussions and defining the theme of this review. This paper represents Contribution Number 42 from the Laboratory for Chemical Physics, New York University.

7. Note Added in Proof

The family of intrinsically disordered or unstructured proteins possess many of the structural and functional characteristics that are found in biomineralization sequences. We believe that many biomineralization proteins should be categorized within this family of proteins, and future work should consider utilizing appropriate approaches in studying these proteins.

8. References

(1) Mann, S.; Webb, J.; Williams, R. J. P. *Biomineralization: Chemical and Biochemical Perspectives*; VCH: Weinheim, Germany, 1989; p 1.
(2) Lowenstam, H. A.; Weiner, S.; *On Biomineralization*, Oxford Press: New York, NY, 1989; p 1.
(3) Michenfelder, M.; Fu, G.; Weaver, J. C.; Wustman, B. A.; Taranto, L.; Evans, J. S.; Morse, D. E. *Biopolymers* **2003**, *70*, 522; errata **2004**, *73*, 291.
(4) Kim, I. W.; Collino, S.; Morse, D. E.; Evans, J. S. *Cryst. Growth Des.* **2006**, *6*, 1078.
(5) Fu, G.; Qiu, S. R.; Orme, C. A.; Morse, D. E.; DeYoreo, J. J. *Adv. Mater.* **2007**, *17*, 2678.
(6) Kulp, J. L., III; Minamisawa, T.; Shiba, K.; Evans, J. S. *Langmuir* **2007**, *23*, 3857.

(7) Yano, M.; Nagai, K.; Morimoto, K.; Miyamoto, H. *Biochem. Biophys. Res. Commun.* **2007**, *362*, 158.
(8) Lee, S. W.; Kim, Y. M.; Choi, H. S.; Yang, J. M.; Choi, C. S. *Protein J.* **2006**, *25*, 288.
(9) Marin, F.; Armons, R.; Guichard, N.; Stigter, M.; Hecker, A.; Luquet, G.; Layrolle, P.; Alcaraz, G.; Riondet, C.; Westbroek, P. *J. Biol. Chem.* **2005**, *280*, 33895.
(10) Weiss, I. M.; Kaufmann, S.; Mann, K.; Fritz, M. *Biochem. Biophys. Res. Commun.* **2000**, *267*, 17.
(11) Gotliv, B.-A.; Kessler, N.; Sumerel, J. L.; Morse, D. E.; Tuross, N.; Addadi, L.; Weiner, S. *ChemBioChem* **2005**, *6*, 304.
(12) Suzuki, M.; Murayama, E.; Inoue, H.; Ozaki, N.; Tohse, H.; Kogure, T.; Nagasawa, H. *Biochem. J.* **2004**, *382*, 205.
(13) Tsukamoto, D.; Sarashina, I.; Endo, K. *Biochem. Biophys. Res. Commun.* **2004**, *320*, 1175.
(14) Ma, Z.; Huang, J.; Sun, J.; Wang, G.; Li, C.; Xi, L.; Zhang, R. *J. Biol. Chem.* **2007**, *282*, 23253.
(15) Li, S.; Xie, L.; Ma, Z.; Zhang, R. *FEBS J.* **2005**, *272*, 4899.
(16) Samata, T.; Hayashi, N.; Kono, M.; Hasegawa, K.; Horita, C.; Akera, S. *FEBS Lett.* **1999**, *462*, 225.
(17) Mann, K.; Siedler, F.; Treccani, L.; Heinemann, F.; Fritz, M. *Biophys. J.* **2007**, *93*, 1246.
(18) Treccani, L.; Mann, K.; Heinemann, F.; Fritz, M. *Biophys. J.* **2006**, *91*, 2601.
(19) Pokroy, B.; Zolotoyabko, E.; Adir, N. *Biomacromolecules* **2006**, *7*, 550.
(20) Liu, H.-L.; Liu, S.-F.; Ge, Y.-J.; Liu, J.; Wang, X.-Y.; Xie, L.-P.; Zhang, R.-Q.; Wang, Z. *Biochemistry* **2007**, *46*, 844.
(21) Weiss, I. M.; Gohring, W.; Fritz, M.; Mann, K. *Biochem. Biophys. Res. Commun.* **2001**, *285*, 244.
(22) Kim, I. W.; Darragh, M. R.; Orme, C.; Evans, J. S. *Cryst. Growth Des.* **2006**, *5*, 5.
(23) Collino, S.; Evans, J. S. *Biomacromolecules* **2007**, *8*, 1686.
(24) Collino, S.; Kim, I. W.; Evans, J. S. *Cryst. Growth Des.* **2006**, *6*, 839–842.
(25) Kim, I. W.; DiMasi, E.; Evans, J. S. *Cryst. Growth Des.* **2004**, *4*, 1113.
(26) Seker, O. U. S.; Wilson, B.; Dincer, I. S.; Kim, I. W.; Oren, E. E.; Evans, J. S.; Tamerler, C.; Sarikaya, M. *Langmuir* **2007**, *23*, 7895.
(27) Sarikaya, M.; Tamerler, C.; Jen, A. K. Y.; Schulten, K.; Baneyx, F. *Nat. Mater.* **2003**, *2*, 577.
(28) Naik, R. R.; Stringer, S. J.; Agarwal, G.; Jones, S. E.; Stone, M. O. *Nat. Mater.* **2002**, *1*, 169.
(29) Thai, C. K.; Dai, H. X.; Sastry, M. S. R.; Sarikaya, M.; Schwartz, D. T.; Baneyx, F. *Biotechnol. Bioeng.* **2004**, *87*, 129.
(30) Naik, R. R.; Brott, L. L.; Clarson, S. J. *J. Nanosci. Nanotechnol.* **2002**, *2*, 95.
(31) Sano, K. I.; Sasaki, H.; Shiba, K. *Langmuir* **2005**, *21*, 3090.
(32) Chen, H. B.; Su, X. D.; Neoh, K. G.; Choe, W. S. *Anal. Chem.* **2006**, *78*, 4872.
(33) Tamerler, C.; Duman, M.; Oren, E. E.; Gungormus, M.; Xiong, X.; Kacar, T.; Parviz, B. A.; Sarikaya, M. *Small* **2006**, *2*, 1372.
(34) Li, X.; Xu, Z-H.; Wang, R. *Nano Lett.* **2006**, *6*, 2301.
(35) Li, X.; Chang, W.-C.; Chao, Y. J.; Wang, R.; Chang, M. *Nano Lett.* **2004**, *4*, 613.
(36) Chateigner, D.; Hedegaard, C.; Wenk, H.-R. *J. Struct. Geol.* **2000**, *22*, 1723.
(37) Li, X.; Nardi, P. *Nanotechnology* **2004**, *15*, 211.
(38) Su, X.-W.; Zhang, D.-M.; Heuer, A. H. *Chem. Mater.* **2004**, *16*, 581.
(39) Eichhorn, S. J.; Scurr, D. J.; Mummery, P. M.; Golshan, M.; Thompson, S. P.; Cernik, R. J. *J. Mater. Chem.* **2005**, *15*, 947.
(40) Dauphin, Y.; Cuif, J. P.; Doucet, J.; Salome, M.; Susini, J.; Williams, C. T. *Marine Biol.* **2003**, *142*, 299.
(41) Feng, Q. L.; Li, H. B.; Pu, G.; Zhang, D. M.; Cui, F. Z.; Li, D. *J. Mater. Sci.* **2000**, *35*, 3337.
(42) Hasegawa, Y.; Uchiyama, K. *Fish. Sci.* **2005**, *71*, 1174.
(43) Sarashina, I.; Endo, K. *Mar. Biotechnol.* **2001**, *3*, 362–369.
(44) Addadi, L.; Raz, S.; Weiner, S. *Adv. Mater.* **2003**, *15*, 959.
(45) Weiss, I.; Tuross, N.; Addadi, L.; Weiner, S. *J. Exp. Zool.* **2002**, *293*, 478.
(46) Gerhke, N.; Nassif, N.; Pinna, N.; Antonetti, M.; Gupta, H. S.; Colfen, H. *Chem. Mater.* **2005**, *17*, 6514.
(47) Krieger, F.; Moglich, A.; Kiefhaber, T. *J. Am. Chem. Soc.* **2005**, *127*, 3346.
(48) Wustman, B.; Santos, R.; Zhang, B.; Evans, J. S. *Biopolymers* **2002**, *65*, 1305.
(49) Cammers-Goodwin, A.; Allen, T. J.; Oslick, S. L.; McClure, K. F.; Lee, J. H.; Kemp, D. S. *J. Am. Chem. Soc.* **1996**, *118*, 3082.
(50) Dumy, P.; Keller, M.; Ryan, D. E.; Rohwedder, B.; Wohr, T.; Mutter, M. *J. Am. Chem. Soc.* **1997**, *119*, 918.
(51) Butcher, D. J.; Nedved, M. L.; Neiss, T. G.; Moe, G. R. *Biochemistry* **1996**, *35*, 698.
(52) Bhattacharyya, R.; Chakrabarti, P. *J. Mol. Biol.* **2003**, *331*, 925.
(53) Kim, I. W.; Morse, D. E.; Evans, J. S. *Langmuir* **2004**, *20*, 11664.
(54) Wustman, B. A.; Morse, D. E.; Evans, J. S. *Biopolymers* **2004**, *74*, 363.
(55) Collino, S.; Evans, J. S. *Biomacromolecules*, in press.
(56) Cubellis, M. V.; Caillez, F.; Blundell, T. L.; Lovell, S. C. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 880.
(57) Kentsis, A.; Mezei, M.; Gindin, T.; Osman, R. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 493.
(58) Chellgren, B. W.; Creamer, T. P. *Biochemistry* **2004**, *43*, 5864.
(59) Delak, K.; Giocondi, J.; Orme, C.; Evans, J. S. *Cryst. Growth Des.*, submitted.
(60) Eker, F.; Griebenow, K.; Cao, X.; Nafie, L. A.; Schweitzer-Stenner, R. *Biochemistry* **2004**, *43*, 613.

CR078251E