BIOINFORMATICS DISCOVERY NOTE

Vol. 28 no. 24 2012, pages 3182–3185
doi:10.1093/bioinformatics/bts604

Sequence analysis

Advance Access publication October 11, 2012

Aragonite-associated biomineralization proteins are disordered and contain interactive motifs

John Spencer Evans*

Laboratory for Chemical Physics, Department of Basic Sciences and Craniofacial Biology, New York University, New York, NY 10012, USA

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The formation of aragonite mineral in the mollusk shell or pearl nacre requires the participation of a diverse set of proteins that form the mineralized extracellular matrix. Although self-assembly processes have been identified for several nacre proteins, these proteins do not contain known globular protein–protein binding domains. Thus, we hypothesize that other sequence features are responsible for nacre matrix protein–protein assembly processes and ultimately aragonite biosynthesis.

Results: Of 39 mollusk aragonite-associated protein sequences, 100% contain at least one region of intrinsic disorder or unfolding, with the highest percentages found in framework and pearl-associated proteins relative to the intracrystalline proteins. In some instances, these intrinsically disordered regions were identified as bind/fold sequences, and a limited number correlate with known biomineral-relevant sequences. Interestingly, 95% of the aragonite-associated protein sequences were found to contain at least one occurrence of amyloid-like or cross-β strand aggregation-prone supersecondary motifs, and this correlates with known aggregation and aragonite formation functions in three experimentally tested protein sequences. Collectively, our findings indicate that aragonite-associated proteins have evolved signature sequence traits of intrinsic disorder and aggregation-prone regions that are important for their role(s) in matrix assembly and mineralization.

Contact: jse1@nyu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 26, 2012; revised on September 6, 2012; accepted on October 4, 2012

1 INTRODUCTION

The nacre layer of the mollusk shell and pearl deposit are both high-performance composites consisting of the calcium carbonate polymorph, aragonite and polysaccharides and proteins (Falini *et al*., 2011; Levi-Kalisman *et al*., 2001). The protein and polysaccharide macromolecular components form a 3D lamellar network around each single-crystal aragonite tablet that contributes to the architecture, organization and molecular properties of the nacre layer (Falini *et al*., 2011; Smith *et al*., 1999). Although we are still in the early stages of understanding the role(s) that proteins play in nacre construction, it is clear that some proteins are self-associative and participate in the formation and stabilization of fracture-resistant single-crystal aragonite (Amos *et al*., 2010, 2011; Marie *et al*., 2010; Ponce and Evans, 2011), whereas other proteins have different responsibilities, such as enzyme catalysis (Berland *et al*., 2011), protease activities (Marie *et al*., 2010) and cell receptor binding (Weiss *et al*., 2001). Thus far, three different classes of nacre-associated proteins have been identified in mollusks. The framework or interlamellar protein family is associated with the β-chitin polysaccharide-containing water-insoluble tri-layer matrix that surrounds each aragonite tablet (Berland *et al*., 2011; Samata *et al*., 1999; Suzuki *et al*., 2010, 2011). In contrast, the intracrystalline protein family is extracted as a water-soluble fraction and is intercalated or found in organic inclusions within aragonite (Marie *et al*., 2010; Michenfelder *et al*., 2003; Suzuki *et al*., 2011). Lastly, there is a family of proteins associated with the formation of the nacre pearl deposit (Liu *et al*., 2007). Collectively, few of these nacre proteins possess known globular binding domains that are typically associated with protein assembly (Berland *et al*., 2011; Marie *et al*., 2010; Michenfelder *et al*., 2003; Samata *et al*., 1999; Suzuki *et al*., 2010, 2011; Weiss *et al*., 2001). Hence, there must be other sequence features that drive nacre protein matrix assembly, which, in turn, drives aragonite formation.

Recent bioinformatics studies revealed that intrinsically disordered and aggregation-prone domains exist within the diverse set of human extracellular matrix protein (HECMP) sequences (Peysselon *et al*., 2011). These domains are believed to be responsible for observed matrix assembly and hierarchal ordering of the HECM. We hypothesize that similar sequence features must be at work in the aragonite extracellular matrix protein (AECMP) scenario as well. With this in mind, we initiated a bioinformatics study of 39 AECMP sequences arising from seven different molluscan species (Supplementary Table S1). These sequences were analyzed at the global level for intrinsic disorder and aggregation propensities. Our study indicates that all AECMP sequences possess one or more disordered sequence regions, with 51% predicted to possess one or more bind/fold recognition regions. Interestingly, 95% of the AECMPs were found to possess aggregation-prone, amyloid- or prion-like cross-β strand supersecondary motifs. We conclude that, as per the human ECM, intrinsically disordered and aggregation-prone sequences are key molecular features that contribute to the formation and function of mollusk nacre.

*To whom correspondence should be addressed.

2 METHODS

The AECMP library was compiled from the complete primary sequences of 39 published aragonite-associated proteins obtained from the GenBank, UniProt or Swiss-Prot using 'pearl', 'nacre', 'matrix', 'aragonite' and 'mollusca' as key search terms (Supplementary Table S1) for sequences deposited prior to March 2012. All isoforms of the n16 (Samata *et al.*, 1999; Suzuki *et al.*, 2010) and Pinctada fucata mantle gene (PFMG) (Liu *et al.*, 2007) families were included in this sequence library. All isoforms of the n16 (Samata *et al.*, 1999; Suzuki *et al.*, 2010) and PFMG (Liu *et al.*, 2007) families were included in this sequence library. All sequences were further subdivided into three categories (11 intracrystalline, 16 framework and 12 pearl) based on known published data or National Center for Biotechnology Information sequence file identifiers that denote their tissue localization and/or their extractability in aqueous media from the nacre matrix (intracrystalline = soluble, framework = insoluble) or the oyster pearl ('pearl' = pearl associated) (Falini *et al.*, 2011; Liu *et al.*, 2007; Marie *et al.*, 2010; Michenfelder *et al.*, 2003; Samata *et al.*, 1999; Suzuki *et al.*, 2010, 2011; Weiss *et al.*, 2000, 2001). Putative signal peptide regions were identified using ExPASy Signal P software (Petersen *et al.*, 2011), and these signal regions were deleted from each DNA-derived primary sequence prior to the analyses described below.

To determine the percentage and the location of disordered sequence regions within AECMPs, we used the IUP_PRED (Dosztányi *et al.*, 2005), GLOBPLOT (Linding *et al.*, 2003) and DISOPRED (Ward *et al.*, 2004) prediction algorithms with standard defaults. Subsequently, we used a suite of programs to globally identify putative sequence regions that exhibit association propensities. The ANCHOR protein–protein binding algorithm was applied (default parameters) to locate IUP-identified disordered regions that have the propensity to energetically partner with another protein (Mészáros *et al.*, 2009). This algorithm was originally used to identify sensitive disordered regions that fold when contacting a globular protein. However, in our instance, we are using these methods to probe the potential folding sensitivity or propensity of aragonite-associated sequence domains without regard for the nature of the stabilizing target. The location of ANCHOR domains were correlated against known biomineralization-relevant sequences (Supplementary Table S2 and S3). The TANGO algorithm (Linding *et al.*, 2004) was used to identify aggregation propensities. Finally, to test the relative accuracy of cross-β strand amyloid-like sequence predictions, we used the predictive algorithms BETASCAN (Bryan *et al.*, 2009), AGGRESCAN (Conchillo-Sole *et al.*, 2007) and FOLD_AMYLOID (Garbuzynsky *et al.*, 2010) on aragonite protein sequences AP7, PFMG1 and n16-3 that have been experimentally confirmed to be intrinsically disordered, contain extended β strand structures and stabilize aragonite *in vitro* (Amos *et al.*, 2010, 2011; Keene *et al.*, 2010).

3 RESULTS

3.1 Intrinsic disorder and bind/fold regions within AECMPs

Using IUP, GLOBPLOT and DISOPRED, we found that all AECMPs contain intrinsically disordered regions, with the framework and intracrystalline classes containing the highest and lowest percentages, respectively (Supplementary Fig. S1). In some cases, it has been documented that some unstable disordered domains will fold when they bind to targets or are influenced by environmental factors (Amos *et al.*, 2010, 2011; Tompa, 2002; Uversky and Dunker, 2010; Ward *et al.*, 2004; Xie *et al.*, 2007). These unique 'bind/fold' domains can be identified using the ANCHOR propensity scoring (Fig. 1) (Mészáros *et al.*, 2009). This scoring method identified a subset of 20 proteins (51%) whose probabilities were >0.5 (Fig. 1). With few exceptions (i.e. gly 30, mpn88 and mpn), these predicted bind/fold regions are <20 amino acids in length. We find fewer bind/fold positive proteins within the intracrystalline class (36%) compared with the framework (56%) and pearl-associated (58%) classes (Fig. 1), and this correlates with their intrinsic disorder score rankings (Supplementary Fig. S1). Two points should be noted. First, ANCHOR-positive intracrystalline and pearl proteins possess fewer bind/fold regions compared with framework proteins. We interpret this to mean that intracrystalline and pearl proteins may be limited in terms of extrinsic target interactions. Second, we note that framework proteins contain five or more predicted ANCHOR binding regions, with some >20 amino acids in length (i.e. Pif 80, bmsp120, bmsp100, mpn88 and mpn). This is consistent with the fact that some framework proteins are known to interact with other AECMPs and with the β-chitin polysaccharide layer (Keene *et al.*, 2010; Samata *et al.*, 1999; Suzuki *et al.*, 2010). Thus, having more than one or longer bind/fold regions would be essential for framework proteins to interact with multiple matrix targets.

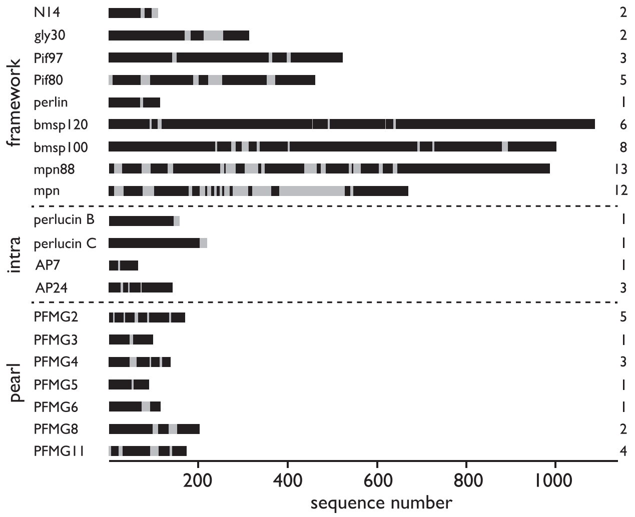

Fig. 1. AECMP ANCHOR bind/fold intrinsically disordered domain profiles. Calculations were performed using standard program defaults. Black bars represent the full-length mature sequence, and gray regions denote the location and length of ANCHOR regions, with the number of identified ANCHOR domains listed on the right

To gain further insight into the identity of these predicted ANCHOR domains, we correlated the location of these bind/fold regions with the occurrence of important biomineral-relevant sequences (Supplementary Tables S2 and S3). We note the following trends: (i) in the framework Pif80, mpn88, mpn, bmsp100 and the pearl-associated PFMG8 sequences, we observe a correlation between some biomineral-associated Pro, Gly-containing repeat regions and ANCHOR domain locations, (ii) PFMG2 contains a Ca (II) calponin-like sequence region that correlates with the location of a predicted ANCHOR binding domain, (iii) AP7 possesses an ANCHOR region that overlaps with a known self-assembly/aragonite stabilization sequence (Fig. 1) (Amos *et al.*, 2010, 2011). Hence, only a limited subset of seven ANCHOR-positive AECMPs (35%) exhibited a correlation between biomineral-relevant domain sequences and the location of a predicted bind/fold region. This suggests that the majority of the predicted bind/fold regions may be unique domains designed for matrix targets that have yet to be identified.

3.2 Aggregation-prone sequences and their occurrence within AECMPs

It has been postulated that amyloid-like sequences evolved as early elements in protein folding and agglomeration processes (Greenwald and Riek, 2012). It is known that these supersecondary short motifs adopt an extended β strand structure and have been shown to be important for initiating self-assembly (Bryan *et al.*, 2009; Conchillo-Sole *et al.*, 2007; Garbuzynsky *et al.*, 2010; Linding *et al.*, 2004). Given that the mollusks evolved millions of years ago (Berland *et al.*, 2011), it is possible that these motifs are present in the existing AECMP proteome. If true, then this would represent a significant discovery that could help explain nacre ECM assembly and subsequent mineral formation.

To assess this, we used TANGO to globally predict the location and number of cross-β strand aggregation-prone regions in each of the 39 protein sequences (Fig. 2). Surprisingly, we note that 37 sequences (95%) were identified as possessing one or more cross-β strand regions. The intracrystalline and pearl-associated sequences are similar in terms of the average number of cross-β strand motifs (2.7 and 2.5 motifs/sequence, respectively), but once again, the framework protein sequences distinguish themselves with higher TANGO scoring (4.6 motifs/sequence, with the highest occurrence noted for bmsp100, bmsp120, gly30, Pif and mpn88) and TANGO sequence lengths. Using BETASCAN, AGGRESCAN and FOLD_AMYLOID, we verified that the self-assembling AP7 (intracrystalline), PFMG1 (pearl) and n16.3 (framework) sequences each contain two or more cross-β strand amyloid-like supersecondary motifs (Supplementary Fig. S3). This correlates with the presence of extended β strand structure in these protein sequences (Amos *et al.*, 2011, 2010; Keene *et al.*, 2010; Ponce and Evans, 2011). We conclude that the majority of AECMPs contain one or more cross-β strand supersecondary motifs. Hypothetically, these motifs would increase the aggregation potential of these sequences and drive matrix assembly processes in tandem with intrinsic disorder (Figs 2 and 3).

4 DISCUSSION

In this study, we confirm that 100% of the 39 studied nacre proteins contain one or more regions of intrinsic disorder (Supplementary Fig. S1), and 95% of these same proteins possess one or more interactive regions, such as bind/fold or amyloid-like motifs (Figs 1 and 2). Hence, we conclude that, like their HECMP counterparts, the AECMPs have evolved signature molecular traits of intrinsically disordered and aggregation-prone 'interactive' sequences that enable matrix assembly (Fig. 3). As expected, each nacre protein class possesses distinguishing features. Framework and pearl-associated classes possess the highest percentages of intrinsic disorder, bind/fold regions and amyloid-like sequences (Figs 1 and 2, Supplementary Fig. S1). Conversely, the intracrystalline class possesses lower percentages of intrinsic disorder (Supplementary Fig. S1) and fewer bind/fold and aggregation-prone sites (Figs 1 and 2). Thus, proteins with higher percentages of 'interactive' regions, such as the framework or pearl subclasses, may be involved with multiple targets (e.g. silk-like fibroin proteins, β-chitin polysaccharide and other framework proteins) and thus require longer length and/or more numerous interactive regions to participate in intermolecular interactions with diverse targets. Conversely, the intracrystalline proteins may be designed for a limited number of protein– matrix interactions within organic inclusions and hence require fewer 'interactive' regions to achieve this. We believe that the presence and extent of intrinsic disorder and interactive sites in a given AECMP are related to (i) the functional attributes of a given protein and (ii) the number and nature of potential matrix targets that a given protein interacts with.

The most unexpected finding was the presence of amyloid-like cross-β strand supersecondary motifs (Fig. 2), a class of aggregation-prone sequences that have been linked with the evolutionary development of protein folding and assembly.

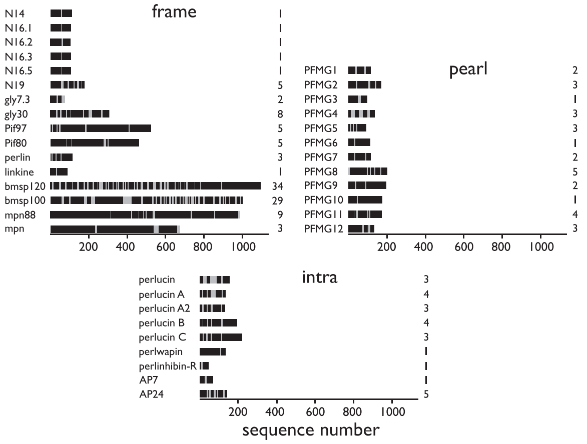

Fig. 2. AECMP TANGO-positive cross-β strand aggregation profiles. Black bars represent the full-length mature sequences, and gray regions denote the location and length of TANGO regions, with the number of identified cross-β strand aggregation domains listed on the right. The following parameters were used: ionic strength = 0.1 M, 289 K (seawater temperature) and pH 8.0 (to mimic *in vitro* mineralization assay conditions) (Amos *et al.*, 2011; Keene *et al.*, 2010)

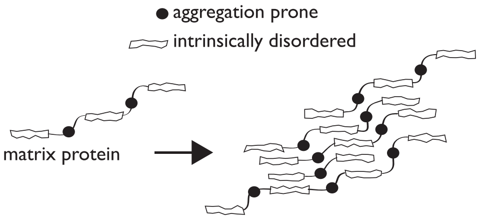

Fig. 3. Hypothetical scheme for AECMP self-assembly via aggregation-prone and intrinsically disordered domain interactions. We assume that there are complementary domain pairings and that both types of domains are simultaneously active during assembly, but this may not be universally true

Similar to what we observed for intrinsically disordered regions, the occurrence of TANGO-identified aggregation-prone regions (i.e. average number of motifs) follows the relationship frame > pearl ∼ intra (Fig. 2). We conclude that although amyloid-like motifs are important for the function of all AEMCPs, they appear to be critical for the framework-specific lamellar layer, possibly playing major roles in protein–polysaccharide recognition, protein–protein assembly and elastomeric behavior under force (Smith *et al.*, 1999). In conclusion, we believe that the sequence location and number of intrinsically disordered and amyloid-like supersecondary motifs may be important for aggregation (Fig. 3), protein orientation and assembly stability, and may also play a role in the recognition and interaction of a given protein with other specific matrix component(s) during the nacre biomineralization process. These sequence features will now become the subject of experiments aimed at confirming their functional roles in nacre assembly.

ACKNOWLEDGEMENT

This article represents contribution number 66 from the Laboratory for Chemical Physics, New York University.

Funding: The US Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award DE-FG02-03ER46099.

Conflict of Interest: none declared.

REFERENCES

Amos,F.F. *et al.* (2010) The N- and C-terminal regions of the pearl-associated EF Hand protein, PFMG1, promote the formation of the aragonite polymorph in vitro. *Cryst. Growth Des.*, **10**, 4211–4216.

Amos,F.F. *et al.* (2011) A C-RING-like domain participates in protein self-assembly and mineral nucleation. *Biochemistry*, **50**, 8880–8887.

Berland,S. *et al.* (2011) Coupling proteomics and transcriptomics for the identification of novel and variant forms of mollusk shell proteins: a study with *P. margaritifera*. *Chembiochem*, **12**, 950–961.

Bryan,A.W. *et al.* (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilities analysis. *PLoS Comput. Biol.*, **5**, 1–11.

Conchillo-Sole,O. *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65–82.

Dosztányi,Z. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Falini,G. *et al.* (2011) The interstitial crystal-nucleating sheet in molluscan Haliotis rufescens shell: a bio-polymeric composite. *J. Struct. Biol.*, **173**, 128–137.

Garbuzynsky,S.O. *et al.* (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, **26**, 326–332.

Greenwald,J. and Riek,J. (2012) On the possible amyloid origin of protein folds. *J. Mol. Biol.*, **421**, 417–426.

Keene,E.C. *et al.* (2010) Matrix interactions in biomineralization: aragonite nucleation by an intrinsically disordered nacre polypeptide, n16N, associated with a β-chitin substrate. *Cryst. Growth Des.*, **10**, 1383–1389.

Levi-Kalisman,Y. *et al.* (2001) Structure of the nacreous organic matrix of a bivalve mollusk shell examined in the hydrated state using cryo-TEM. *J. Struct. Biol.*, **135**, 8–17.

Linding,R. *et al.* (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

Linding,R. *et al.* (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins (TANGO). *J. Mol. Biol.*, **342**, 345–353.

Liu,H.L. *et al.* (2007) Identification and characterization of a biomineralization related gene *PFMG1* highly expressed in the mantle of *Pinctada fucata*. *Biochemistry*, **46**, 844–851.

Marchler-Bauer,A. *et al.* (2011) CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, 225–229.

Marie,B. *et al.* (2010) Proteomic analysis of the acid-soluble nacre matrix of the bivalve Unio pictorum: detection of novel carbonic anhydrase and putative protease inhibitor proteins. *Chembiochem*, **11**, 2138–2147.

Mészáros,B. *et al.* (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, 1–18.

Michenfelder,M. *et al.* (2003) Characterization of two molluscan crystal-modulating biomineralization proteins and identification of putative mineral binding domains [published erratum appears in *Biopolymers* (2004), **73**, 291]. *Biopolymers*, **70**, 522–533.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Peysselon,F. *et al.* (2011) Intrinsic disorder of the extracellular matrix. *Mol. Biosyst.*, **7**, 3353–3365.

Ponce,C.B. and Evans,J.S. (2011) Polymorph crystal selection by n16, an intrinsically disordered nacre framework protein. *Cryst. Growth Des.*, **11**, 4690–4696.

Samata,T. *et al.* (1999) A new matrix protein family related to the nacreous layer formation in *Pinctada fucata*. *FEBS Lett.*, **462**, 225–229.

Smith,B.L. *et al.* (1999) Molecular mechanistic origin of the toughness of natural adhesives, fibres, and composites. *Nature*, **399**, 761–763.

Suzuki,M. *et al.* (2010) An acidic matrix protein Pif is a key macromolecule for nacre formation. *Science*, **325**, 1388–1390.

Suzuki,M. *et al.* (2011) Identification and characterization of a calcium carbonate-binding protein, blue mussle shell protein (BMSP), from the nacreous layer. *Chembiochem*, **12**, 2478–2487.

Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.

Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta*, **1804**, 1231–1264.

Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Weiss,I.M. *et al.* (2000) Purification and characterization of perlucin and perlustrin, two new proteins from the shell of the mollusc *Haloitis laevigata*. *Biochem. Biophys. Res. Commun.*, **267**, 17–21.

Weiss,I.M. *et al.* (2001) Perlustrin, a *Haloitis laevigata* (abalone) nacre protein, is homologous to the insulin-like growth factor binding protein N-terminal module of vertebrates. *Biochem. Biophys. Res. Commun.*, **285**, 244–249.

Xie,H. *et al.* (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.*, **6**, 1882–1898.