Chapter 3

The Skeletal Proteome of the Coral *Acropora millepora*: the Evolution of Calcification by Co-option and Domain Shuffling

The work presented in this chapter was published in *Molecular Biology and Evolution*, 2013, 30(9), 2099–2112. doi: 10.1093/molbev/mst109
Chapter 3 The Skeletal Proteome of the Coral *Acropora millepora*: the Evolution of Calcification by Co-option and Domain Shuffling

Authors:

Paula Ramos-Silva\textsuperscript{a,b}, Jaap Kaandorp\textsuperscript{b}, Lotte Huisman\textsuperscript{b,e}, Benjamin Marie\textsuperscript{c}, Isabelle Zanella-Cléon\textsuperscript{d}, Nathalie Guichard\textsuperscript{d}, David J. Miller\textsuperscript{e} and Frédéric Marin\textsuperscript{a}

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

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

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

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

\textsuperscript{e} ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia
3.1 Introduction

In corals, biocalcification is a major function that may be drastically affected by Ocean Acidification (OA). Scleractinian corals grow by building up aragonitic exoskeletons that provide support and protection for soft tissues. Although this process has been extensively studied, the molecular basis of biocalcification is poorly understood. Notably lacking is a comprehensive catalogue of the skeleton-occluded proteins - the skeletal organic matrix proteins (SOMPs) that are thought to regulate the mineral deposition.

Using a combination of proteomics and transcriptomics, we report the first survey of such proteins in the staghorn coral *Acropora millepora*. The organic matrix extracted from the coral skeleton was analyzed by mass spectrometry and bioinformatics, enabling the identification of 36 SOMPs. These results provide novel insights into the molecular basis of coral calcification and the macroevolution of metazoan calcifying systems, while establishing a platform for studying the impact of OA at molecular level. Besides secreted proteins, extracellular regions of transmembrane proteins are also present, suggesting a close control of aragonite deposition by the calicoblastic epithelium. In addition to the expected SOMPs (Asp/Glu-rich, Galaxins), the skeletal repertoire included several proteins containing known extracellular matrix domains.

From an evolutionary perspective, the number of coral-specific proteins is low, many SOMPs having counterparts in the non-calcifying cnidarians. Extending the comparison to the skeletal organic matrix proteomes of other metazoans allowed the identification of a pool of functional domains shared between phyla. The data suggest that cooption and domain shuffling may be general mechanisms by which the trait of calcification has evolved.
Data deposition: The protein sequence data reported in this paper was submitted to the UniProt Knowledgebase under the accession numbers B3EWY[6-9], B3EWZ[0-9], B3EX0[0-2], B7W112, B7W114, B7WFQ1, B8RJM0, B8UU51, B8UU59, G8HTB6 B8UU74, D9IQ16, B8UU78, B8V7P3, B8V7Q1, B8V7R6, B8V7S0, B7T7N1, B8VIV4, B8VIU6, B8VIW9, B8VIX3, B8WI85.
### 3.2 Background

It is generally accepted that anthropogenic CO$_2$ emissions cause deleterious effects, not only on the atmosphere (*i.e.* global warming), but also on seawater chemistry. Since pre-industrial times, the average pH of seawater has fallen from 8.2 to 8.1, and this ocean acidification (OA) is thought to have significantly affected marine organisms in a variety of ways, including impacts on important physiological functions such as calcification [128–130]. This process is essential for a wide range of marine metazoans and fulfills a diverse array of important physiological and ecological roles [131]. In the neritic domain of tropical and sub-tropical regions, the most prominent calcium carbonate producers are cnidarians, in particular scleractinians (stony corals), where coral reefs have an annual net production ranging from 5 to 126 mol CaCO$_3$ m$^{-2}$ year$^{-1}$ [132,133]. Recent studies have shown that OA and increasing sea surface temperatures decrease coral calcification rates [70,128,134–138]. However, most of the data supporting these negative effects are obtained at the organism/colony level and only few studies have considered the underlying molecular and cellular mechanisms. Of particular significance is the fact that the molecular machinery of coral biomineralization (the 'calcification toolkit') is, as yet, poorly characterized. Thus a better understanding of the process is required, in particular at the level of the mineralizing space, where CaCO$_3$ crystals are formed and shaped [33]. A number of candidate genes for roles in coral calcification have been identified for being expressed in the calicoblastic ectoderm [139–141]. This number was recently increased with high-throughput analyses [142–144], while microarray studies led to the identification of the SCRs, a family of Small Cysteine-Rich proteins showing molecular features suggestive of
an involvement in calcification processes [145]. Amongst the most obvious candidates for roles in coral calcification are the genes that encode skeletal organic matrix proteins (SOMPs), i.e., proteins secreted by the calcifying tissues – the calicoblastic ectoderm in corals – which are occluded in the skeleton during its formation [28]. These proteins are the main components of the skeletal organic matrix (OM), which in corals contains also polysaccharides [32], glycoproteins [146] and lipids [147].

The SOMPs have been characterized biochemically ‘in bulk’ (i.e. without fractionation) in different coral species [148–151], the compositional analyses have shown that it is enriched in aspartic acid and, in lesser extent, in glutamic acid and glycine. This feature, together with the presence of saccharidic moieties, makes the OM unusually acidic. Moreover, its interaction with calcium carbonate was observed experimentally in the scleractinians Balanophyllia europaea [152], Acropora digitifera, Lophelia pertusa, Montipora caliculata [153] as well as by OM produced by coral cells of Stylophora pistillata [154], suggesting an important role of the OM in coral calcification.

While several organic matrix proteins have been extensively studied in other calcifying metazoans, such as mollusks [22] and sea urchins [155], their characterization is still in its infancy for corals. Only recently an approach combining proteomics with genomics was applied unveiling a set of partial and complete sequences of proteins in the skeleton of the hexacoral Stylophora pistillata [156]. Prior to this study, only the full sequence of two SOMPs, Galaxin from the scleractinian Galaxea fascicularis [56], and Scleritin from the octocoral
Corallium rubrum [157] were published and a number of other partial sequences from octocorals [158–161] and scleractinians [151].

One prerequisite for understanding the process of coral calcification is a comprehensive survey of the skeletal matrix proteins. In the present paper, we address this requirement, by describing 36 extracellular proteins that constitute the SOMP repertoire, or ‘skeletome’, of the scleractinian coral Acropora millepora.

This is the first large-scale survey of the proteins present in the skeleton of a member of the acroporid family and the second in cnidarians. These proteins give new insights on the molecular tools required for controlling the deposition of the calcium carbonate and provide a platform for investigating the impact of environmental factors on calcification at the molecular level. In addition, they open novel perspectives on the mechanisms of evolution of SOMPs within the Cnidaria and, more broadly, within other calcifying metazoans.

3.3 Materials and Methods

3.3.1 Skeletal collection and SEM observations

Acropora millepora colonies were collected at the Great Barrier Reef in Australia (Pioneer Bay, Orpheus Island) in November 2010, prior to the annual spawning event. Mother colonies that died after spawning were used to collect the skeletal material; animal tissue, symbionts and other microorganisms were removed by immersion in NaOCl (5%, vol/vol) for 72 h. The skeletal material was then rinsed with purified water, dried and mechanically fragmented. The skeleton microstructure was observed with a tabletop Scanning Electron Microscope (SEM, Hitachi TM-1000) under an acceleration voltage of 15 kV. Mirror polished
transverse and longitudinal sections were observed after etching in EDTA (1% w/vol, 3 min.), rinsed with water and dried. The skeletal fragments were then split in two batches. The first one was not submitted to further treatment before mineral dissolution for matrix extraction while the second batch was treated further (see below).

3.3.2 Organic matrix extraction

Skeletal fragments of the second batch were reduced to powder (Fritsch Pulverisette 14), which was subsequently sieved (particle size below 200 µm). The powder was bleached in NaOCl solution (10 times dilution, 0.26% active chlorine) for 5 h and washed with milli-Q water several times until no trace of NaOCl was left. This treatment allows removing organic exogenous or endogenous contaminants that can be entrapped in the highly porous skeleton (Figure 3.1), while keeping intact the most tightly bound skeletal matrix components [162]. The extraction was performed according to a published procedure [81].
Section 3.3 Materials and Methods

Figure 3.1: SEM images from skeleton macro and microstructure. (A) Radial corallites arrangement with separate walls and internal septa, (B) Closer view into corallites showing different vestigial and complete septa. Polished and EDTA-etched sections from a longitudinal cut: (C) Fibrous microstructure, (D) Close-up showing evidence of crystalline fibers. The skeleton of *Acropora millepora* is composed exclusively of aragonite. Macroscopic views of the skeleton surface reveal cup-like rounded corallites separated by distinct walls and the highly porous coenosteum. The radial corallites are evenly distributed and of approximately equal size, with complete and incomplete septa. In longitudinal sections, a rather uniform microstructure is evident, but with unevenly distributed porosity. At higher magnification, individual needle-like fibers, constituting the basic units of the coral microstructure [8], are observed. These exhibit a discrete size with trabecular orientation (white arrows) [163].

In brief, the dried powder put in suspension in cold water was decalcified overnight in acetic acid (10% vol/vol) at 4 °C with an electronic burette (Titronic Universal, Schott, Mainz, Germany). The solution was centrifuged: the organic pellet (acid-insoluble matrix, AIM) was rinsed several times with Milli-Q water and freeze-dried. The supernatant (acid-soluble matrix, ASM) was treated on an ultrafiltration cell (Amicon 400 mL, 10 kDa cutoff membrane) for volume reduction, then once concentrated, dialyzed several days against water at 4 °C and freeze-dried.

63
The extraction was performed in duplicate (2 x 30g of skeletal powder under the same NaOCl treatment) in order to check the reproducibility of the results.

### 3.3.3 ASM/AIM analysis on 1D and 2D gel electrophoresis

The skeletal matrix – both ASM and AIM fractions - was analyzed on 1D electrophoresis and stained with silver nitrate [109]. The ASM was directly denatured with Laemmli buffer [108] according to standard conditions while the AIM was only partly solubilized by Laemmli buffer. The solubilized fraction is defined as the Laemmli-soluble/acetic acid insoluble fraction (LS/AIM). Electrophoresis was performed on discontinuous 12% acrylamide mini-gels at 100 V for 15 min and 150 V for 1 h. For the 2-D electrophoresis, IPG strips (ReadyStrip™, BioRad) were loaded with 180 μg of ASM dissolved in 180 μL of re-hydratation buffer (6 M urea, 2 M thiourea, 4% (w/v) Chaps, 20 mM dithiothreitol, 0.1% ampholytes, 0.001% bromophenol blue) and re-hydrated overnight at 50 V (25 °C) in a PROTEAN® IEF cell (BioRad). Focusing was carried out at 250 V for 15 min, followed by 4000 V for 2 h and 4000 V until 10000 Vh. Subsequently the IPG strips were equilibrated by transfer for 10 min into 1 mL of equilibration buffer I (6 M urea, 2% SDS, 375 mM Tris/HCl pH 8.8, 20% glycerol) with 2% (w/vol) dithiothreitol, followed by 10 min in equilibration buffer II, with 2.5% (w/vol) iodoacetamide instead. Strips were rinsed in 25 mM Tris, 192 mM glycine and 0.1% SDS (TGS) and placed on the top of precast NuPAGE(R) BisTris Novex SDS-polyacrylamide gels (4-10%) (Invitrogen, Carlsbad, CA, USA) with an overlay solution of 0.5% agarose/TGS (w/v).
Electrophoresis was performed at 200 V for 35 min. Both 1D and 2D gels were stained with silver nitrate [109].

### 3.3.4 Proteomic analysis

The AIM and ASM were prepared for proteomic analysis according to a routine procedure including reduction, alkylation, trypsin digestion, drying and solubilization in TFA [49, 50]. MS analyses were performed in duplicate on a LTQ Velos (ThermoScientific, France) instrument in the positive ion mode. The ion source was equipped with a picoTip emitter as nanospray needle (FS360-75-30-CE-5-C10.5, NewObjective, USA) operating at 1.5 kV. The acquisition was done with the Excalibur 2.1 software (ThermoScientific). Typically two scan events were used: at first, m/z 400-1600 survey scan MS with enhanced resolution; secondly, data dependent scans MS/MS on the twenty most intense ions from the previous event. The spectra were recorded using dynamic exclusion of previously analyzed ions for 0.6 min. The MS/MS normalized collision energy was set to 35eV. LC was performed on an Ultimate 3000 nano-LC system (Dionex, Voisins Le Bretonneux, France). Chromatographic separation of peptides was obtained with a C\textsubscript{18} PepMap micro-precolumn (5 µm; 0.3 mm x 5 mm) for a 3 min desalting and a C\textsubscript{18} PepMap nano-column (3 µm; 100 Å; 75 µm x 150 mm) with a gradient elution at a flow rate of 300 nL/min. Eluent A was a mixture of 95% H\textsubscript{2}O, 5% CH\textsubscript{3}CN and 0.1% formic acid (vol/vol). Eluent B was a mixture of 20% H\textsubscript{2}O, 80% CH\textsubscript{3}CN and 0.1% formic acid (vol/vol). The gradient program was from 0% B to 50% B over 120 min and 100% B for 10 min.
3.3.5 *In silico* analysis of the SOMPs

The data obtained with the LTQ-velos system was used for protein identification with the MASCOT search engine (version 2.1, Matrix Science, London, UK). MS/MS raw data were searched with carbamidomethylation as fixed modification, and methionine oxidation, asparagine and glutamine de-amidation as variable modifications. The tolerance of the precursor and fragment masses was set to 0.4 Da. Proteins identified with at least two distinct peptides were considered valid assignments. The corresponding peptide sequences were validated by manually inspection of the MS/MS spectra (Appendix A). The subject of MASCOT searches was a pooled protein database comprising the 6-frame translated nucleotide sequences from different publicly available sources in April 2012: *Acropora millepora* sequences were downloaded from NCBI including 101,380 plus 15,389 sequences from the nucleotide and EST databases, respectively. EST (7.964) and nucleotide (5.112) data from the genus *Symbiodinium* were also included together with the assembled EST of two *Symbiodinium* strains (Mf104b - clade B (76,284 EST) and KB8 - clade A (72,152 EST)) [164] from http://medinalab.org/zoox. The proteins identified through MASCOT were used in similarity searches against the UniprotKB/Swissprot database. Subsequently protein sequences were analyzed for the presence of signal peptides with Signal IP 4.0 (http://www.cbs.dtu.dk/services/SignalP/), transmembrane domains with TMHMM v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) and further characterized for homologous domains and regions using the InterproScan platform (http://www.ebi.ac.uk/Tools/pfa/iprscan/). In addition, the proteins with transmembrane domains were also analyzed for potential cleavage sites cleaved by
proteases with PeptideCutter (http://web.expasy.org/peptide_cutter/).

### 3.3.6 Homology analysis and protein comparisons at the domain level

Homology analysis were performed using the 36 *A. millepora* sequences involved in biomineralization (transcripts and proteins) against the predicted coding genes of *Acropora digitifera* (http://marinegenomics.oist.jp/genomes/gallery), *Nematostella vectensis* (http://www.uniprot.org/) and *Hydra magnipapillata* (ftp://ftp.jgi-psf.org/pub/JGI_data/Hydra_magnipapillata/). First, searches with a local BLAST (version 2.2.25+) [165] were performed using the transcripts (and protein sequences) of the 36 SOMP from *A. millepora* against:

1. the coding genes of *Acropora digitifera* (TBLASTN), with default parameters;
2. the coding genes from *N. vectensis* and *H. magnipapillata* (TBLASTX), with default parameters;
3. the predicted proteins of the 3 cnidarian genomes (BLASTP), with and without SEG (i.e. low complexity filter on query sequence) [166].

The best hit for each SOMP (E value threshold < $10^{-4}$) was selected and the sequences globally aligned using Needleman-Wunsch algorithm for pairwise alignment (cutoff: 30% identity, expect for mosaic proteins) (Appendix B, Table 3). The best matches were also manually compared for their domain architecture. Second, the Neighborhood Correlation (NC) method
(http://www.neighborhoodcorrelation.org/) [167] was implemented for homology identification between the 3 genomes (A. digitifera vs. N. vectensis and A. digitifera vs. H. magnipapillata). Multidomain proteins with high confidence NC coefficients (> 0.8) were considered strictly homologues [168]. Orthologues of SOMPs in A. digitifera genome were selected and subsequently analyzed for the presence of the homologues in N. vectensis and H. magnipapillata. The SOMPs and corresponding best matches in the 3 genomes that were not complete sequences but showed significant blast scores (E value < 10^{-4}), were not assessed for their homology and just considered similar. High-throughput proteomic datasets obtained from OMs occluded in CaCO₃ structures were collected from the literature and their primary sequences analyzed with InterProScan: Stronglylocentrotus purpuratus (tooth [169], spicules [170], test and spine [171]), Gallus gallus (eggshell [172,173]), Lottia gigantea (shell [174]), Crassostrea gigas (shell [175]) and Pinctada (shell [54]). The complete set of signatures obtained in these proteomes (InterPro entries - domains and repeats) was compared with those present in the SOMPs.

3.4 Results and Discussion

3.4.1 Analysis of the matrix on gel

Coral skeleton was treated with bleach twice to remove contaminating tissue, symbionts and bacteria, that can be entrapped in the highly porous skeleton (Figure 3.1), the second treatment being conducted after the skeleton had been reduced to fine powder (granulometry < 200 µm). The residual powder was then extracted
with 10% acetic acid, yielding acid-soluble (ASM) and acid-insoluble (AIM) matrix fractions, assumed to be contaminant-free and closely associated with the aragonite skeleton. This extraction process yielded approximately 0.034±0.01% of ASM and 0.23±0.03% (w/w) of AIM relative to the skeleton weight. 1D-gel electrophoresis of the ASM extract (Figure 3.2 A) showed 3 main discrete bands at around 120, 90 and 64 kDa embedded in a diffuse background of more dispersed ‘smearing’ macromolecules, while the AIM was characterized by a smear lacking discrete bands. On the 2D gel (Figure 3.2 B), the 90 and 64 kDa fractions, and, to a lesser extent, the 120 kDa fraction, in the ASM are characterized by large spots circa pH 3, suggesting that these macromolecules are very acidic.
3.4.2 Identification and characterization of SOMP

Raw data generated by LC-MS/MS on the ASM and AIM were analyzed with the software MASCOT to search against a pooled database consisting of transcripts and predicted proteins of *Acropora millepora* and the genus *Symbiodinium*. This procedure was initially made on the AIM and ASM prepared after only a single (whole skeleton) bleach treatment (‘first batch’), but was then repeated on organic matrix fractions (‘second batch’) after the second bleaching step (i.e. on skeleton reduced to fine powder) outlined above. This enabled to analyze the effect of extended bleaching on removal of contaminants (Figure 3.3). Whereas the proteomic analyses of the ‘first batch’ revealed intracellular coral proteins and proteins of symbiont origin (Chapter 4) the second batch was free of such
contaminants, enabling the detection of 43 unique *Acropora millepora* proteins likely to constitute the SOMP repertoire *i.e.*, proteins that are strongly associated to the skeleton.

![Venn diagram showing the comparison of proteins identified by proteomics on the acid-soluble (ASM) and acid-insoluble matrices (AIM) in two different conditions, batch 1 and batch 2.](image)

Figure 3.3: **Comparison of the proteins identified by proteomics on the acid-soluble (ASM) and acid-insoluble matrices (AIM) in two different conditions, batch 1 and batch 2.** Batch 1 consisted of treating the skeletal fragments with sodium hypochlorite once, while batch 2 consisted of batch 1 followed by a subsequent similar treatment on the skeletal sieved powder (< 200 µm). Extracts from batch 1 showed evidence of contamination with the identification of specific intracellular proteins from *A. millepora* (tubulins, histones, ATP-synthase) and proteins from zooxanthellae: 2 contaminants of 23 identifications in the ASM and 28 contaminants of 38 identifications in the AIM. In contrast no contaminants were identified in batch 2, indicating that a second bleach treatment on powder is effective in removing potential sources of contamination and is required for obtaining exclusively SOMPs. The 22 SOMPs identified in batch 1 were also present in batch 2.

Of these 43 proteins, 36 assignments could be made with high confidence (*i.e.* with more than one unique peptide matching the sequence, see Appendix B - Table 1), while 7 sequences each with only a single peptide match were dropped from the list (Appendix B - Table 2) despite their properties being generally consistent with those of the high-confidence dataset. Figure 3.4 provides a schematic overview of the domain structure and general properties of the *Acropora* SOMPs that were
classified based on the in silico sequence analyses. From the 36 SOMP s identified, 22 proteins were common to both ASM and AIM. Similar ASM and AIM protein repertoires have also been observed in the case of mollusk shell matrices [111]. On the other hand, 12 SOMP s were exclusively associated with the insoluble organic fraction of the skeleton, and only 2 with the soluble matrix. Possible causes for this apparent bias (Figure 3.3) between matrices include the over/under representation of certain peptides that is inherent to the proteomic approach, and the fact that some proteins from the acid-soluble fraction may bind acid-insoluble components.
Figure 3.4: List of SOMPs identified by proteomics in the acid soluble and insoluble matrices extracted from *Acropora millepora* skeleton. Proteins were named according to the characterization of their primary structure.
Acidic proteins

In both ASM and AIM the most abundant proteins (high emPAI values [176], Appendix B - Table1) were acidic (Figure 3.2 B), i.e., with $pI < 4.5$ [92] and rich in negatively charged residues (Asp, Glu).

Asp and Glu-rich proteins are supposed to interact directly with calcium carbonate crystals promoting crystal nucleation, determining the growth axes and inhibiting the crystal growth [123,177]. Due to their polyanionic character, Asp and Glu-rich proteins usually have high-capacity, but low-affinity, calcium binding properties [93]. Among the A. millepora SOMP s, 6 acidic proteins were identified (Figure 3.4), including two similar (48.8% identity; Figure 3.5) proteins each containing ~20% of Asp residues and named as SAARP1 and 2 (Secreted Acidic Asp-Rich Protein).

Figure 3.5: Pairwise sequence alignment of the secreted acidic Asp-rich proteins: SAARPs 1 and 2 (UniprotKB Ac. Nos.: B3EWY6, B3EWY8). Identical residues are dyed in blue. Sequence alignments were performed and visualized with Jalview [178].

These proteins are likely to belong to a family of highly acidic proteins conserved across scleractinians. Indeed two Asp-rich proteins, named coral acid-rich proteins (CARPs), were recently identified in Stylophora pistillata skeleton [156]. Similarly
to CARPs, SAARPs contain two Asp-rich regions intercalated by two non-acidic regions (Figure 3.4). In particular CARP 4 exhibits the highest identity percentage with the two SAARPs, 60.6 and 45.2%, for SAARP 1 and 2 respectively. However, the predicted protein sequences of CARP4 and CARP5 are shorter than SAARPs and lack a putative peptide signal required for targeting these proteins to the secretory pathway. Interestingly the 4 acidic proteins also show significant identity, in the non-acidic regions, with the protein fragment P27 identified in the skeleton of \textit{S. pistillata} [156] (Figure 3.6), however this similarity is not understood from a functional viewpoint.

The 2 SAARPs together with a third protein, Acidic SOMP, corresponded to the identifications with the highest emPAI scores (Appendix B – Table 1). The Acidic SOMP has lower Asp content (9.9%) and its sequence does not contain two Asp-rich regions as in SAARPs (Figure 3.4). Still, it shows significant similarities in the non–acidic region with both SAARPs (Figure 3.7).
Chapter 3 The Skeletal Proteome of the Coral *Acropora millepora*: the Evolution of Calcification by Co-option and Domain Shuffling

The three other skeletal proteins included the secreted acidic proteins Amil-SAP1, Amil-SAP2 (19 and 21.6% of Asp and Glu, respectively) and one Glu-rich protein. Amil-SAP1 and Amil-SAP2 are the counterparts of Adi-SAP1 (Figure 3.8) and Adi-SAP2, respectively, which have been hypothesized to have roles in calcification along with seven other proteins, Adi-SAP3 to -SAP9, in *Acropora digitifera*, on the basis of their high Asp plus Glu content [143]. Of these acidic proteins, only SAP1 and 2 exhibit a predicted signal peptide, and are confirmed here to be true SOMPs. BLASTP searches with the 6 acidic SOMPs retrieved significant matches (E value < 10⁻⁴) with proteins in the UniprotKB database. Interestingly, among the best matches for Amil-SAP1 and the SAARPs were shell matrix proteins (aspeins) from *Pinctada fucata*, *Pinctada maxima* and *Isognomon perna*. However, in these cases, similarity scores might be largely misleading due to the relatively low complexity of the sequences, dominated by acidic residues; note also that none of these SOMPs had significant similarity to annotated domains from the InterPro database. Moreover, these data confirm the direct involvement of acidic proteins in skeletal formation in *Acropora millepora*.  

---

**Figure 3.7:** Multiple sequence alignment of the C-terminal from the Secreted Acidic Asp-rich Proteins, SAARP1 and 2 with two related sequences, Uncharacterized SOMP-4 and Acidic SOMP (UniprotKB Ac. Nos.: B8UU74, B3EWY7). Conservation of residues is dyed by shades of blue: the darker the color, the more conserved the residue. Sequence alignments were performed with MUSCLE [179] and visualized with Jalview [178].

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAARP1/381-386</td>
<td>HLLACGIL-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 302</td>
</tr>
<tr>
<td>Uncharacterized_SOMP_4/116-224</td>
<td>HMMCGAMD-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 310</td>
</tr>
<tr>
<td>SAARP2/365-192</td>
<td>HMMCGAMD-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 358</td>
</tr>
<tr>
<td>Acidic_SOMP/271-359</td>
<td>HMMCGAMD-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 359</td>
</tr>
<tr>
<td>Uncharacterized_SOMP_4/116-224</td>
<td>HMMCGAMD-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 310</td>
</tr>
<tr>
<td>SAARP1/381-386</td>
<td>HLLACGIL-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 302</td>
</tr>
<tr>
<td>Acidic_SOMP/271-359</td>
<td>HMMCGAMD-NCOT IAMFLEPTED-CERKRPFFKDNNHIDSUYFDG 359</td>
</tr>
</tbody>
</table>

The three other skeletal proteins included the secreted acidic proteins Amil-SAP1, Amil-SAP2 (19 and 21.6% of Asp and Glu, respectively) and one Glu-rich protein. Amil-SAP1 and Amil-SAP2 are the counterparts of Adi-SAP1 (Figure 3.8) and Adi-SAP2, respectively, which have been hypothesized to have roles in calcification along with seven other proteins, Adi-SAP3 to -SAP9, in *Acropora digitifera*, on the basis of their high Asp plus Glu content [143]. Of these acidic proteins, only SAP1 and 2 exhibit a predicted signal peptide, and are confirmed here to be true SOMPs. BLASTP searches with the 6 acidic SOMPs retrieved significant matches (E value < 10⁻⁴) with proteins in the UniprotKB database. Interestingly, among the best matches for Amil-SAP1 and the SAARPs were shell matrix proteins (aspeins) from *Pinctada fucata*, *Pinctada maxima* and *Isognomon perna*. However, in these cases, similarity scores might be largely misleading due to the relatively low complexity of the sequences, dominated by acidic residues; note also that none of these SOMPs had significant similarity to annotated domains from the InterPro database. Moreover, these data confirm the direct involvement of acidic proteins in skeletal formation in *Acropora millepora*.
Figure 3.8: Mapping of the two protein fragments from Amil-sap1 (UniprotKB Ac. Nos.: B3EWZ0 (N-terminal) and B3EWZ1 (C-terminal)) with corresponding ortholog from A. digitifera (Adi-sap1). Mapping was visualized with Jalview [178]. Identical residues are shaded in blue.

Unique uncharacterized proteins

High-throughput approaches such as proteomics and transcriptomics have often revealed completely novel proteins that lack conserved domains or significant database matches that would allow to classify them into families and to hypothesize about their functions [72]. Galaxin exemplifies the case of such ‘orphans’ since its function remains unknown [180]. In A. millepora skeleton we identified two distinct galaxins along with further 8 uncharacterized ‘orphan’ SOMPs (Figure 3.4) referred here as USOMPs. These proteins do not constitute a single group and, with the exception of three proteins, exhibit no significant similarities. Two of these exceptions are galaxin proteins, and the third is provided by USOMP-4, with a region near the C-terminus that has similarity to three of the acidic proteins discussed above (Figure 3.7). The presence of multiple galaxin-related genes has previously been reported [143,180]; one of the galaxins identified here corresponds to GenBank ADI50283.1, a previously known gene from A. millepora [180]. The second sequence (Galaxin 2) is orthologous with Galaxin 2.
from *A. digitifera* [143]. The two *A. millepora* galaxins have 31.5% identity overall, the di-Cys repeat region being most similar. Note that the di-Cys pattern might indicate a role in the assembly and support of the macromolecular framework [56]. Indeed a multiple alignment with 5 galaxins and similar sequences from *Nematostella vectensis* (GIs: 156374951, 156377965) and *Hydra magnipapillata* (GI: 221103149) shows the conservation of at least nine double Cys motifs (Figure 3.9). Similarly in *S. pistillata* skeleton, three proteins were identified, P12, P16 and P22 [156], which can fit in our description of uncharacterized proteins. From these, P16 and P22 are completely unique while P12 shows similarities with USOMP8 (Appendix B - Table 5).
Figure 3.9: Multiple sequence alignment showing the sequence similarities between galaxins from Acropora millepora (Amil), Acropora digitifera (Adi), Galaxea fascicularis (Gfas) and the detected homologues from Nemastella vectensis (Nemve) and Hydra magnipapillata (Hmag). Amil-Galaxin (Uniprot Ac. No.: D9IQ16), and Gfas-Galaxin (Uniprot Ac. No.: Q8I6S1) correspond to the same form of galaxin and are grouped together with Adi-Galaxin1 (predicted from Adi transcriptome: EST_assem_14006). Subsequently the Amil-Galaxin 2 (Uniprot Ac. No.: B8UU51) and Adi-Galaxin 2 (predicted from Adi transcriptome: EST_assem_8935) are shorter lacking the Cys residues and the presence of other served substitutions are indicated (A). Regions with high frequency of semi-conserved substitutions are also indicated (A). Sequence alignments were performed with MUSCLE [179] and visualized with Espript 2.2 [107].
Extracellular matrix-like proteins

Other SOMPs, although in some cases lacking clear overall matches, contain conserved regions or domains. The largest group comprises those with one or more domains usually associated with extracellular matrix (ECM) or cell-adhesion proteins from other metazoans. Amongst the fifteen ECM-related proteins identified in the skeletal organic matrix of *A. millepora* most contained multiple domains (Figure 3.4). Several of these ECM-related SOMPs resembled human cell-adhesion proteins such as Mucin-4 (Mucin-like SOMP, B3EY9) and Hemicentin-1 (Coadhesin, B3EZW3), which are heavily glycosylated and rich in cysteine residues, thereby forming disulfide bridges involved in modulating their adhesive properties. SOMPs that exhibit mucin-like or coadhesin signatures have previously been reported in the mineralizing matrices of mollusks [51,88,125], such as thrombospondin type 1 repeats, von Willebrand factor type A and epidermal growth factor-like domains, which are typical of ECM proteins involved in cell-cell and cell-substrate adhesion and in the binding of other macromolecules [181,182]. Moreover, many ECM domains present in *Acropora*’s skeletal proteome have been detected in previous high-throughput proteomic studies on the OM of mollusks, sea urchin, chicken eggshell and coral (Appendix B - Table 4) but only the Polycystic Kidney Disease 1 (PKD1)-related protein (B8UU59) contains domains (lipooxygenase, egg-jelly receptor and PKD/chitinase) not previously found in proteomic analyses of other skeletal organic matrices. On the other hand some of the datasets used in these comparisons consist of broader lists of proteins, including some of intracellular origin that may be derived from cell debris and are most likely not involved in biomineralization [170,175]. Thus comparisons were interpreted carefully and taking into account other evidences such as the expression in skeleton secretting-tissues [54,175,183],
Section 3.4 Results and Discussion

the direct interaction with CaCO$_3$ or the detection in the calcifying mediums [184] (Appendix B – Table 4).

From the fifteen ECM-related SOMPs, 8 have homology domains with proteins from the *S. pistillata* skeletal proteome (Appendix B - Table 4 and Table 5): Mucin-like, Coadhesin, Ectin (B3EWZ8), EGF and Laminin G domain-containing protein (B8UU78), MAM and LDL-receptor domain-containing protein 1 (B3EWZ5) and 2 (B3EWZ6), Zona pellucida domain-containing protein (G8HTB6) and Protocadherin-like (B8V7Q1). Still, apart from Protocadherin-like, we could not confirm candidate orthologous for these SOMPs in *S. pistillata* since most sequences correspond to fragments. Four ECM-related SOMPs were exclusively detected in the AIM, two of them having homologues usually found in the central nervous system in metazoans [185,186] - Protocadherin-like and Neuroglian-like protein (B8VIW9), also the EGF and Laminin G domain-containing protein and finally Collagen (B8V7R6). The latter is homologous to the human Alpha-2 type I Collagen (P08123) which is found in bone [187,188]. Fibrils of collagen are generally in close contact with the mineral phase and remain as an insoluble component of the organic fraction throughout the extraction process. While collagen has been detected in the spicules of gorgonian [189,190] and scleractinian corals [156] this is the first report of the occurrence of collagen type I in the OM of a coral skeleton.

**Enzymes**

Four of the SOMPs identified correspond to three types of enzymes (Figure 3.4). The first enzyme corresponds to a 148 AA fragment (B8V7P3) of an Alpha-type Carbonic anhydrase (CA), a zinc metalloenzyme that catalyzes the reversible hydration of carbon dioxide. CA activity has been detected in the skeletal organic
matrix of both zooxanthellate (Acropora hebes [191]) and azooxanthellate (Tubastrea aurea [192]) scleractinians.

In the case of the CA identified here, its best BLAST hit in UniprotKB is the α-CA from Stylophora pistillata, STPCA-2 (C0IX24). This enzyme was previously shown to be highly active compared to other CAs and was localized in the oral endoderm and aboral tissue [193]. Despite its high similarity with the cytosolic human Carbonic anhydrase II, STPCA-2 has unusual features such as a putative signal peptide and an insert of 30 amino acids between positions 212 and 243. Moreover the same protein was also identified in the skeleton of S. pistillata by proteomics [156]. These results, together with the data presented here for Acropora, directly link one specific CA to skeleton formation. Similarly to a mechanism proposed for some mollusks [53], the CA found here may display its enzymatic function in the extracellular calcifying medium, and be subsequently occluded in the growing skeleton.

The second SOMP (B3EWZ9), Hephaestin-like, is a secreted member of the copper oxidase family with 1114 AA length, containing a copper-binding site and a cupredoxin domain. Hephaestin proteins function as copper-dependent ferroxidases, mainly involved in iron and copper metabolism at membranes [194,195]. Although to date there is no experimental evidence of activity of this type in mineralizing matrices, it is reasonable to propose that the Hephaestin-like protein catalyzes the oxidation of iron (from Fe^{2+} to Fe^{3+}) in the compartment where aragonite precipitation occurs. It was previously suggested that corals may entrap iron in the skeleton as a detoxification mechanism when high concentrations of the metal are present in the reef [196]. Alternatively, the incorporation of iron in the skeleton of A. millepora may be a normal process, regulated by Hephaestin.
Interestingly, cupredoxin domains have also been identified in the shell of the mollusk *Crassostrea gigas* (in the L-ascorbate oxidase, K1PLZ9; Laccase-15, K1QQA2 and Laccase-18, K1PMS4) and in the tooth and spicules of the echinoderm *S. purpuratus* (in a protein similar to Hephaestin, H3JMP5) (Appendix B - Table 4). Finally, the third enzymatic function assigned is observed in two SOMP (B8V7S0, B8VIV4), exclusive to the AIM. It corresponds to proteases containing peptidase S1/S6, chymotrypsin/Hap and CUB domains.

Even though they are present at relatively lower abundance (Appendix B – Table 1), the enzymes detected here may play crucial roles in the supply of bicarbonate ions at the site of calcification (CA), in the regulation of iron and copper metabolism (hephaestin), and in the assembly/cleavage of the organic matrix (proteases).

**Toxin-like**

One SOMP (B7W114) corresponds to a secreted protein that has high similarity (50%) with the N-terminus of a toxin (of 1052 aa long) from the cephalopod *Sepia esculenta* (SE-cephalotoxin) (B2DCR8; BLASTP E value = 8^{-43}). While a number of toxins have been characterized from cubozoans (box jellyfish) [197] and anthozoans (sea anemones) [198]. However very little is known about toxins in scleractinian corals, though studies have demonstrated antibacterial activity of the mucus [199,200]. The level of identity between B7W114 and the cephalotoxin (Figure 3.10) is suggestive of conserved function. If the toxic character of the Cephalotoxin-like SOMP is later confirmed, we may hypothesize that this protein acts in the coral skeleton in a similar way to the lysozyme in the chicken eggshell [184], which due to its well-known anti-microbial properties was suggested to have a protective function in the eggshell in addition to its ability to interact with calcium carbonate.
3.4.3 Proteins with transmembrane domains

The occurrence of transmembrane proteins associated to calcium carbonate biominerals has been explained as contamination by soft tissues [175]. However, with the double-bleaching procedure used in the present case, this hypothesis is very unlikely and we suggest another explanation. Eleven SOMPs identified among the different groups described above contain putative transmembrane (TM) domains. Ten of these SOMPs have a single TM domain located at their C-terminus (Figure 3.4) the exception being the PKD1-related protein, which is predicted to contain eleven TM helices, four of which are part of the polycystin cation channel domain. SOMPs with TM domains represent about one half of the ECM group, one protein (Hephaestin-like) in the enzyme group, one protein (Uncharacterized SOMP-3) of the ‘orphan’ group and two in the acidic group (SAARP 2 and SAP-1). We propose that the TM proteins detected here are cleaved in their extracellular region, becoming subsequently occluded in the forming biomineral.
Three lines of evidence support this hypothesis. First, without exception, the 75 peptides detected by LC-MS/MS that correspond to the eleven TM domain-containing SOMPs belong to the extracellular region located at the N-terminal side of the proteins (Appendix B - Table 1). Second two proteases were detected amongst the SOMPs having chymotrypsin/Hap domains that may constitute the required tool for cleaving the extracellular region of the TM domain-containing SOMPs. Third, the eleven TM-containing SOMPs possess predicted chymotrypsin cleavage sites in their extracellular regions (Appendix B - Table 1). Specific cleavage sites for chymotrypsin and TM domains are illustrated in Figure 3.4 for two SOMPs: SAARP2 and Zona pellucida domain-containing protein. These lines of evidence are congruent with the recent data on the spicule matrix proteins of the sea urchin S. purpuratus [170], so the mechanism proposed here may apply more generally.

If this hypothesis is later confirmed by other approaches, this may call for a complete change of paradigm in the biomineralization field, whatever the biological model studied. Until now, the ‘molecular tools’ controlling the mineralization process were all supposed to be secreted outside the calcifying epithelial cells in order to interact with the inorganic precursors of mineralization. The data presented here suggest that membrane-bound proteins may also contribute to the process via their extracellular domains, which are subsequently incorporated in the biomineral after being cleaved by peptidases. Furthermore, it suggests that there is a true connection between the epithelium and the mineral front, and that the mineral deposition is accomplished under the direct guidance of the cell surfaces, rather than remotely, as ‘classical’ views tend to show, in particular in mollusk shell models [18].
Figure 3.11: Primary structures of transmembrane SAARP2 and Zona pellucida domain-containing protein, including: putative peptide signals (underlined), codon stop (*), TM domains (rectangle), peptides identified by MSMS (red) and chymotrypsin high specificity cleavage sites (residues [FYW] not before P highlighted in green) with more than 80% of cleavage probability.

3.4.4 SOMP s in early stages of calcification affected by high CO₂

The availability of transcriptome data derived from Acropora millepora primary polyps [144] permitted investigation of the impact of elevated CO₂ on expression of the SOMP s identified here. Comparisons with the datasets submitted to the NCBI Gene Expression Omnibus database ( GEO) under the accession number GSE33016 confirmed that genes corresponding to all of the SOMP s are expressed during the early stages of calcification and at least 26 of these were up regulated while 4 were down regulated in polyps exposed to higher concentrations of CO₂ (Table 3.1). SOMP s are therefore appropriate molecular markers for the investigation of the impacts of elevated CO₂ on Acropora calcification.
Table 3.1: Differential expression of genes involved in *Acropora millepora* bio mineralization according to a previous experiment on primary polyps [144]: up-regulated genes (green), down-regulated genes (red), not available (-). Fold-changes (P value > 0.05) were obtained through the analysis of the count data available on the NCBI Gene Expression Omnibus database (GEO) under accession number GSE33016, using the edgeR package [201]. Transcript levels were originated from *Acropora millepora* primary polyps at 380 (control), 750 and 1000 ppm CO$_2$ after 3 days exposure.

<table>
<thead>
<tr>
<th>Protein Groups</th>
<th>Cluster</th>
<th>NCBI Ac. No.</th>
<th>Protein Names</th>
<th>CO2 750 ppm (vs. control)</th>
<th>CO2 1000 ppm (vs. control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidic</td>
<td>Cluster009205 JR01945</td>
<td>SAARP 1</td>
<td>1.41</td>
<td>2.32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster008253 JR972076</td>
<td>Acidic SOMP</td>
<td>-</td>
<td>-1.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster017392 JR991407</td>
<td>SAARP2</td>
<td>-</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster018838 JTD06291</td>
<td>SAPI</td>
<td>2.01</td>
<td>3.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster037255 JTD18094</td>
<td></td>
<td>3.77</td>
<td>5.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster02029 JR983041</td>
<td>SAP2</td>
<td>1.81</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster014254 JR93175</td>
<td>Glu-rich protein</td>
<td>1.93</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>Extracellular matrix proteins</td>
<td>Cluster011118 JTD11118</td>
<td>MAM and LDL-receptor domain containing protein 1</td>
<td>3.81</td>
<td>3.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster000006 JR994474</td>
<td>MAM and LDL-receptor domain containing protein 2</td>
<td>2.29</td>
<td>2.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster018354 JTD18096</td>
<td>THP-rich protein</td>
<td>1.87</td>
<td>1.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster010849 JR978035</td>
<td>EGF-like</td>
<td>-</td>
<td>3.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster012957 JTD13217</td>
<td>MAM and fibronectin-containing protein</td>
<td>6.70</td>
<td>8.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster007429 JTD16410</td>
<td>MAM and fibronectin containing protein (isoform)</td>
<td>-</td>
<td>1.77</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster000133 JR991141</td>
<td>PKD1-related protein</td>
<td>1.98</td>
<td>2.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster011245 JR973492 (JN631095)</td>
<td>Zona pellucida domain containing protein</td>
<td>-1.42</td>
<td>-1.86</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster001085 JR980881</td>
<td>EGF and laminin G domain containing protein</td>
<td>-</td>
<td>1.44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster000035 JTD11093</td>
<td>Proteasome-like</td>
<td>1.62</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster005041 JR991083</td>
<td>Collagen</td>
<td>8.07</td>
<td>11.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster005565m JR991827</td>
<td>Neurogin-like</td>
<td>2.60</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster015162 JR989025</td>
<td>CUB domain-containing protein</td>
<td>1.87</td>
<td>2.87</td>
<td></td>
</tr>
<tr>
<td>Enzymes</td>
<td>Cluster002345 JTD09483</td>
<td>Hephaestin-like</td>
<td>2.23</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster020494 JR998014</td>
<td>Carbonic anhydrase</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster023283 JR970990</td>
<td>CUB and Ser protease domain-containing protein 1</td>
<td>7.24</td>
<td>9.54</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.1 (cont.)

<table>
<thead>
<tr>
<th>Cluster005989</th>
<th>JT008002</th>
<th>CUB and Ser protease domain-containing protein 2</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster013356</td>
<td>JR933671 (HM162215)</td>
<td>Galaxin</td>
<td>1.69</td>
<td>2.81</td>
</tr>
<tr>
<td>Cluster013531</td>
<td>JR936600</td>
<td>Galaxin 2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster013623</td>
<td>JR927770</td>
<td>USOMP-1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster014180</td>
<td>JR973746</td>
<td>USOMP-1</td>
<td>1.01</td>
<td>1.95</td>
</tr>
<tr>
<td>Cluster013072</td>
<td>JR997000</td>
<td>USOMP-2</td>
<td>1.53</td>
<td>1.96</td>
</tr>
<tr>
<td>Cluster026802</td>
<td>JT044948</td>
<td>USOMP-4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster026853</td>
<td>JR927717</td>
<td>USOMP-5</td>
<td>4.56</td>
<td>4.03</td>
</tr>
<tr>
<td>Cluster012833</td>
<td>JR971508</td>
<td>USOMP-6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster014464</td>
<td>JR982600</td>
<td>USOMP-7</td>
<td>4.16</td>
<td>6.67</td>
</tr>
<tr>
<td>Cluster001446p</td>
<td>JR998260</td>
<td>USOMP-8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cluster006260</td>
<td>JT014391</td>
<td>USOMP-9</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Uncharacterized proteins

Cluster005989 | JT008002 | - | - |

Toxin Cluster001924 | JR986059 | Protein similar to cephalotoxin | - | - |

3.4.5 Homology comparison between Acropora, Nematostella vectensis and Hydra magnipapillata

In an attempt to unravel the origins of the A. millepora ‘skeletome’, homologues of the 36 biomineralization proteins were searched amongst the predicted proteins (using BLASTP) and corresponding transcripts (BLASTN, TBLASTX) from the three cnidarians for which whole genome sequence data are presently available - Acropora digitifera, a scleractinian coral which diverged from A. millepora in the Mio-Pliocene [202], the sea anemone Nematostella vectensis, and Hydra magnipapillata. Sea anemones are classified as a distinct Order (Actinaria) from corals (Scleractinia) but in the same sub-class (Zoantharia = Hexacorallia) within Class Anthozoa. Whereas sea anemones are considered to be “close”, but skeletonless, relatives of corals, Hydra is a phylogenetically remote cnidarian belonging to Class Hydrozoa [203–205]. This approach enabled proteins that are involved in Acropora biomineralization but have ancient cnidarian origins (i.e. are present in Hydra as well as Acropora) to be resolved from those that may be anthozoan-
Section 3.4 Results and Discussion

specific but not restricted to calcifying anthozoans (i.e. present in *Nematostella* as well as *Acropora*) and those that are so far unique to stony corals. Clear orthologs of each of the 36 *A. millepora* SOMPs could be identified in *A. digitifera* (Appendix B - Table 1 and Table 3), confirming that all the sequences assigned to *A. millepora* are of coral origin and are not derived from zooxanthellae.

BLAST comparisons against the predicted proteins of *Nematostella vectensis* and *Hydra magnipapillata*, indicate that eight of the SOMPs do not have counterparts in the two non-calcifying cnidarians: five “orphan” proteins (USOMPs), SAP-1, SAP-2 and the SOMP similar to cephalotoxin – (outer circle, Figure 3.12). Indeed SAPs were previously suggested to be restricted to members of the genus *Acropora* [143].

To date a BLAST search against the cnidarian sequence data on NCBI with these proteins only retrieved orthologous sequences in other *Acropora* species (Appendix B - Table 3) and low sequence similarities (E value < 10^{-5}) in *Porites astreoides* (USOMP-2), *Montastrea faveolata* (USOMP-2, USOMP-4), *Aiptasia pallida* (USOMP-1, USOMP-3), *Anemonia viridis* (USOMP-3) and *Clytia hemisphaerica* (Similar to cephalotoxin). In particular, the apparent absence of homologues of the Cephalotoxin-like sequence in other cnidarians is surprising given the similarity between the *Acropora* and cuttlefish (*Sepia*) proteins. Note that the SE-cephalotoxin (Uniprot Ac: B2DCR8) identified in the salivary glands of the cuttlefish *Sepia esculenta* (Cephalopoda, Mollusca) does not exhibit any significant similarity to known proteins [206]. In turn the striking similarity (50%) between the *Acropora* and *Sepia* sequences (Figure 3.10) suggests that the ancestor of these proteins predates the divergence of mollusks and corals (i.e. 550-600 million years ago (MYA) or more) but could have been lost in most of the other eumetazoan lineages. Such a genome restructuring, *i.e.* massive loss, has been recently documented via large-scale comparisons of distantly related genomes: for example,
the gastropod *Lottia* and *Nematostella* share exclusively 89 gene families that are not retrieved in other phyla investigated [207]. Amongst the other SOMPs, we have distinguished similar (white circle area, Figure 3.12) from homologous (green, blue circles, Figure 3.12) proteins taking into account the statistical significance of BLAST searches, the percentage of identity (cutoff 30%) in the pairwise global alignment, the common domain architecture and the fact that many of the SOMPs are mosaic proteins, i.e. proteins with multiple domains [208,209] (Appendix B Table 3). In particular for those proteins from the ECM-like group and enzymes, it is difficult to infer homology merely on the basis of BLAST searches (generally highly significant within a certain domain) together with identity percentages and alignment coverage (generally low). To overcome this difficulty the Neighborhood Correlation method was applied to the 3 genomes. This method has been developed and used to accurately identify homologues in complex multidomain families [167,168], enabling the prediction of similar proteins (due to domain insertion) from homologous proteins (due to common ancestor).

This combined analysis implied that at least three SOMPs have homologues in the anthozoan but not in *H. magnipapillata* (middle circle, Figure 3.12): USOMP-7, Neuroglian-like and Zona pellucida domain-containing protein, while other 9 SOMPs, all from the ECM-like group, have homologues in the three genomes (inner circle, Figure 3.12). Several of these conserved ECM-like SOMPs are also found in vertebrates, such as the Protocadherin-like, Collagen, and PKD1-related proteins. Whether these proteins function in similar ways in the skeletal matrix and in the ECM remains to be clarified. However, these results strongly corroborate the idea of a modern skeletal matrix derived from the recruitment of non-calcifying ECM-components.
Conversely, homologues for 16 of the SOMPS could not be confidently identified in the non-calcifying cnidarians due to specific factors: incompleteness of at least one of the sequences in study (Carbonic anhydrase, Peptidase, Glu-rich, USOMP-5), low identity percentages (cutoff 30%) (Hephaestin-like, Galaxins, SAARPs, Acidic SOMP, USOMP-8) and different combination of domain architectures and neighborhood coefficients above the threshold (Mucin-like, Coadhesin, Threonine-rich, CUB domain-containing protein, Peptidase) (Appendix B - Table 3). In consequence these SOMPs were considered as ‘similar to’ and not confirmed as homologous or non-homologous (white circle area, Figure 3.12). Interestingly, homologues of hephaestin are present in both chordates and the symbiotic sea anemone *Anemonia viridis* whereas only low similarity matches were identified in *N. vectensis* and *H. magnipapillata* (Figure 3.13). Also the SOMP Mucin-like has similar domain architecture in the C-terminus side (extracellular nidogen, AMOP, vWD, EGF) to the human Mucin-4 found in chordates, however the same domain combination was not identified in the genomes of *Nematostella* and *Hydra* (Appendix B - Table 3).
Chapter 3 The Skeletal Proteome of the Coral *Acropora millepora*: the Evolution of Calcification by Co-option and Domain Shuffling

Figure 3.12: Resume of the results of similarity searches with BLAST and homology detection (by global alignment, domain architecture comparison and the Neighborhood Correlation method) using the 36 SOMPs (and corresponding transcripts) from *Acropora millepora* and the genomes of *Acropora digitifera*, *Nematostella vectensis* and *Hydra magnipapillata*. Proteins in the outer circle (red) do not have similarities (and homologues) in the predicted proteins of *Nematostella* or *Hydra*. Proteins in the middle circle (green) have homologues in *Acropora* and *Nematostella*. Proteins from the inner circle (blue) have homologues in *Acropora*, *Nematostella* and *Hydra*. SOMPs in the white region of the circle show considerable similarity with proteins from *Nematostella* and *Hydra* but their homology is not certain. The phylogenetic tree on the upright side represents the relationships previously purposed between Cnidaria [205]. dcp = domain containing protein.
Figure 3.13: Multiple sequence alignment showing the sequence similarities between hphaestin-like proteins from Acropora millepora (Uniprot Ac. No.: B3EWZ9), Aiptasia pallida (NCBI GI: 387005847), Homo sapiens (Uniprot Ac. No.: Q6MZM0), Mus musculus (Uniprot Ac. No.: Q3V1H3) and the cupredoxin-domain containing protein from Hydra magnipapillata (NCBI GI: 221113181). Conservation of residues is dyed by shades of blue: the darker the color, the more conserved the residue among the five species. Sequence alignments were performed with MUSCLE [179] and visualized with Jalview [178].
3.5 Conclusions

The work described here represents the first comprehensive survey of the skeletal molecular toolkit of the scleractinian coral *Acropora millepora*. In total, 36 proteins were identified from the acid soluble and insoluble skeletal organic matrices. The proteins of these two fractions overlapped significantly, only 2 proteins being exclusively associated with the soluble fraction and 12 with the insoluble fraction. It is unclear how complete this survey is; trypsin-resistant skeletal proteins, in particular those that do not possess suitable trypsin cleavage sites, may not have been detected. It is also possible that the calicoblastic cells secrete proteins that guide the calcification process but are not occluded in the skeleton and would therefore not be detectable by the approach used here.

One tenet of our analyses is that the classical view of a coral skeletal matrix composed primarily or exclusively of acidic proteins [210,211] is eclipsed by the idea of a more complex matrix consisting of a diverse range of proteins with a larger spectrum of different potential functions. Whilst acidic proteins remain quantitatively key players in calcification (as aptly illustrated by the 2D gel), other SOMPs, some present in lower amount, are also likely to fulfill important roles in the deposition of the skeleton. The second tenet of our analysis is that the SOMPs are not exclusively secreted. Some SOMPs are membrane-bound proteins, the extracellular domains of which are subsequently cleaved in the mineralizing space and occluded in the growing biomineral.

What kind of information on the macroevolution of calcifying matrices can be inferred from this analysis? The evolutionary picture that emerges from our analysis is multifaceted. Whilst all of the SOMPs identified in *A. millepora* have
homologues encoded in the genome of *Acropora digitifera*, surprising few of these are actually unique to the scleractinians, *i.e.* present in the two acroporids but absent from non-calcifying cnidarians (*N. vectensis, H. magnipapillata*). These coral-restricted proteins may be ‘novelties’ in the calcification repertoire. The timing of this innovation might have occurred deep in the Paleozoic when scleractinian corals emerged (ca. 450 MYA) [212]. In contrast, a considerable proportion of the ECM-like SOMP identified here have homologous proteins in *N. vectensis* and *H. magnipapillata* suggesting that their origins predate the divergence of the classes Anthozoa and Hydrozoa [204], which occurred in the Cambrian or earlier. Each of these proteins, and a number of others subsequently lost from *N. vectensis* and *H. magnipapillata* (typified by the proteins similar to Cephalotoxin, Hephaestin-like and Mucin-like), was presumably co-opted to function in calcification when this trait arose in the scleractinian lineage. Broadening the comparison to include other OM proteomes from metazoans revealed that most of the domains represented in the *A. millepora* skeletal proteome also occur in other OM (Appendix B – Table 4), pleading for the concept of a pool of ‘shared domains’ potentially involved in biomineralization. However, only few of the proteins containing these common domains are predicted to be homologous, suggesting that they have been independently recruited to roles in calcification at different times across the various taxa with this trait. Beyond macro-evolutionary considerations, the dataset presented here constitutes an important tool for understanding the molecular bases of calcification in stony corals, and for quantifying the impact of environmental changes (such as OA), on this process.