Amelogenin form and function by computational analysis

ABSTRACT

The mechanism for enamel mineralization by the X-chromosomal Amelogenin gene product (AMELX) remains a mystery. I piece together the existing relevant evidence from the literature to describe constraints for modeling the tertiary protein structure, dissect AMELX into functional regions, and apply predicted functional regions as peptides in a gamut of mineral interaction experiments**.** To understand relative position of functional regions, I implement knowledge-based computational biology approaches to protein structure prediction. To ascertain quantitative importance of each amino acid to AMELX function, I assess evolutionary conservation and distinction using evolutionary trees built with multiple sequence alignments for each residue. To model hydroxyapatite interaction regions, I apply spatial and electrostatic assessments of calcium ion binding sites to putative AMELX, LRAP-4, and LRAP+4 protein structures. I decipher structural importance for each amino acid by applying knowledge-based scoring functions to virtual mutant structures. Novel functional regions emerge within the hydrophobic region, including two regions predicted for higher functional importance than any other AMELX region, termed the proximal HΦ and central HΦ regions. An interspersed region with the highest effect on structural stability of the putative AMELX hexameric model appears to induce specific spatial proximity of the proximal HΦ and central HΦ regions. The resulting construct suggests a novel mineralization functionality. The distal HΦ region is predicted to contribute non-specific hydrophobic forces inducing AMELX-AMELX interactions. Previously identified functional regions are subdivided for functionality, including the YINFSYE motif from the A domain. Experimental assessment by others verifies all predictions made for the proximal HΦ and central HΦ regions: derived peptides bind hydroxyapatite, and a peptide comprised by both catalyzes the formation of hydroxyapatite in a nonprecipitating solution equivalently to the full length protein. As well, an additional functionality emerges from the YINFSYE motif as nucleating calcium phosphate intermediates, described by our hexamer model to coincide with oligomerization. The analyses of functional importance and the relative positioning of functional regions within the predicted quaternary structure concur with established data and substantially expand the understanding of AMELX function by region and residue. These findings elaborate the functional significance of differentially conserved residues in mammalian versus non-mammalian amniote AMELX. The consensus of multiple orthogonal prediction methods, each based on large high quality data, indicates that I have found some if not all unique functionalities hidden within the puzzle of AMELX. The model describes dynamic hexamer assembly bringing together the YINFSYE motif to nucleate calcium phosphate intermediates, then matured by the continuously exposed proximal HΦ and central HΦ regions.

1. Introduction

Food acquisition and masticatory function in amniotes arise in part from the structural properties of enamel. To construct enamel in regenerative medicine, it is imperative to understand its formation as guided by AMELX, the most abundant protein in enamel during all stages of development. Mysteries remain for the mechanistic basis of AMELX function, including the tertiary structure, as well as the identity of the regions responsible for interactions with the forming hydroxyapatite surface and AMELX-AMELX interactions.

AMELX bears critical roles in enamel formation. The various expression products of the AMELX gene comprise 95% by weight of total protein in the developing enamel matrix (Bartlett et al., 2006; Sasaki et al., 1997; Termine et al., 1980; Veis, 2003). The diverse functions proposed for AMELX include seeding calcium phosphate formation (Tarasevich et al., 2007), guiding mineralization by directly limiting hydroxyapatite growth (Robinson et al., 1979), intercellular signaling (Boabaid et al., 2004; Lacerda-Pinheiro et al., 2006; Le et al., 2007; Tompkins and Veis, 2002; Tompkins et al., 2005), adhesion of the enamel matrix to the ameloblasts Tomes' process (Kirkham et al., 2006), and rod-interrod luting (Nanci et al., 1996).

Mineral constituency in AMELX-related Amelogenesis Imperfecta (AI) is equivalent to wild type (wt) enamel (Kida et al., 2007), which implicates the unique function of AMELX in mineral organization rather than crystal seeding. AMELX gene knock-out mice demonstrate an absence of prismatic enamel crystal or rod formation (Gibson et al., 2001; Kim et al., 2004; Prakash et al., 2005). During the formation of hydroxyapatite crystal clusters, the growing apatite crystals adhere to each other through the molecular self-association of interacting AMELX molecules (Moradian-Oldak et al., 1998).

AMELX removal is necessary for enamel maturation, and occurs through catabolic digestion by enamel-specific proteases kallikrein-4 (KLK4) and matrix metalloproteinase 20 (MMP-20; enamelysin). MMP-20 acts at specific sites, producing functional cleavage products (Simmer and Hu, 2002), possibly divided as functional regions.

The elusive structure of AMELX and yet unidentified mineralization regions instigate this study. The unique physical properties of this protein which enable production of the hardest vertebrate tissue do not lend themselves to traditional approaches to determine protein structure (Moradian-Oldak, 2007). However, progress may be acheived through computational approaches.

Data from in vitro, in vivo, and Amelogenesis Imperfecta human phenotype (AI) experiments are reviewed and compiled with a computational model of AMELX structure and function by region.

This analysis focuses on the mature full length functional human AMELX protein "H175" (the expression product of exons 2,3,5,6ABCD,7), LRAP-4 (exons 2,3,5,6D,7) and LRAP+4 (exons 2,3,4,5,6D,7).

1.2. Computational Biology Approaches

Understanding proteins at a mechanistic level requires identification of functional regions and residues. Computational genomics offers a variety of approaches for assigning function to residues and regions of proteins (Oren et al., 2007; Wang et al., 2008). The overlapping predictions of these methods cross-substantiates their reliability, as these methods use different data types for their analyses. I apply these methods to elucidate the distribution of functions throughout AMELX.

I derive patterns from large experimental data sets to predict protein structure (chapter 5, Liu et al., 2009; Samudrala et al., 1999; Xia et al., 2000), and assess importance of individual amino acids to protein structure stabilization (chapter 3, Cheng et al., 2005), catalytic function (chapter 2, Wang and Samudrala, 2005; Wang et al., 2008), ion binding sites (chapter 2, Cheng et al., in preparation), and hydroxypatite interactions (chapter 4). These tools enable study of human molecules in parallel to those of model animals, enabling translation to medicine and tissue regeneration.

Knowledge of protein tertiary structure is necessary to understand function, due to the creation of active sites by three-dimensional (3D) proximity of chemical functional groups. The accuracy of protein structure prediction has been greatly advanced in recent years, as documented by the CASP experiments (Samudrala and Levitt, 2002; Schueler-Furman et al., 2005; Zhang, 2007), such that protein structures and active site topology can be predicted computationally. Only a few structures are known for proteins that interact with hydroxyapatite, and none directly image the protein-hydroxyapatite interface. Nonetheless, with almost fifty thousand protein structures in the PDB (Berman et al., 2000) to apply into threading for development of decoys, predictions can take advantage of all available atomic resolution structural information created under the same evolutionary constraints. This structural study also forms a compliment to experiments assessing the secondary structure of AMELX in solvated conditions highly dissimilar to the enamel matrix.

The novel method of accurately predicting metal ion-binding pockets using 3D space and charge distributions across solvent-exposed areas on a protein structure (SOAK; (Cheng et al., in preparation)) enables prediction of calcium ion binding sites, independent of sequence information. The spatial grid coordinate system simplifies the geometry of valence electron orbitals, bonds, and side chains. The ratio of the area covered by an atom cluster to the total area of the contacting shell measures the compression of ions around specific chemical groups and simulates a partial covalent effect. This 3D method, found at <http://protinfo.compbio.washington.edu/soak>, performs an exhaustive grid-based search around the

entire structure, and applies a geometric knowledge-based scoring function to reveal the location of the best pocket (Cheng et al., in preparation). Calcium ion binding is used as a model for the protein-hydroxyapatite interface in the determination of the AMELX region(s) interacting with hydroxyapatite.

The sequence analysis techniques employed within MFS find residues evolutionarily conserved for either chemical function or fold stabilization. Functional residues can be sieved out through subtraction of fold stabilizing residues. Structural residues are found by creating amino acid mutations and determining the effect on tertiary structure using biophysics-based modeling and assessment (Cheng et al., 2005; Wang et al., 2008). If the naturally occurring amino acid has a worse stability score than the mutants, it does not aid structural integrity, and thus if it is evolutionarily conserved, it likely contributes to chemical function. Conversely, a loss of stability with mutation is evidence for a structurally important residue (Cheng et al., 2005; Wang et al., 2008).

In chapter 2 we demonstrated that training the meta-functional signature sequence analysis platform (MFS) for a specific enough function, in that case calcium ion binding, can greatly improve predictive ability. This finding represents a conceptual shift from our previous findings, wherein neither training MFS to find catalytic residues or ligand binding residues specifically were as accurate for the trained task, as when training on both (Wang *et al.*, 2008). Of course the functional sites in each of these sets were highly diverse. So either the data were not robust enough to allow saturation of any training method, the training method was not capable of simultaneously learning multiple modes of function, or the sequence evolutionary signal exploited by the original MFS method truly is the most salient feature. Per the demonstration of the ability for MFS to learn a highly specified function in chapter 2, i.e. mfsCa, and weights given to the original functions in the MFSCa compilation, the case appears to be a combination of the latter two. Thus we ask whether the new algorithms, methods, and heuristics presented in chapters 2 and 3 might work to improve predictions for a different specific application, phosphate ion binding residue prediction, and the same general purpose application of the original MFS.

Each separate method within MFS generates scores via conceptually simple and easy-to-implement algorithms, while their combined use outperforms sophisticated algorithms performing only one task (Wang et al., 2008). Similarly, each of the computational methods employed in this study cannot be held as empirical evidence, but the overlap of these predictions bears strong probability for function prediction.

2. Methods

2.1. Protein structure prediction

Contemporary protein structure prediction methods use known structures as variable length templates for construction assembly. I developed an initial set of incomplete structure alignments in the form of meta-predictions using 3D-Jury (<http://bioinfo.pl/meta>; Ginalski et al., 2003). I incorporated these as templates for the development of many full length models using the Protinfo protein structure prediction protocol (<http://protinfo.compbio.washington.edu/protinfo_abcmfr>; Hung et al., 2005; RAMP: A suite of programs to aid in the modelling of protein structure and function <http://compbio.washington.edu/ramp/>). In addition I retrieved full length models from the I-TASSER server <http://zhang.bioinformatics.ku.edu/I-TASSER> (Wu et al., 2007). These models were iteratively refined using template-, geometry- and physics-based methods included within the RAMP suite (Hung et al., 2005; Samudrala et al., 1999, Liu et al., 2009). From all decoys created, I used an all-atom conditional probability discriminatory function (RAPDF; Samudrala and Moult, 1998), to select the top models. I then found the best combination of stable fragments using a graph-theoretic clique finding approach, in the creation of an optimized conformation model (chapter 5, Liu et al., 2008). This resulted in twenty five models for H175, LRAP+4, and LRAP-4.

Model selection using experimental annotation - I incorporated all available experimental evidence contributing to atomic level characterization of AMELX structure, such as indication of surface presentation. I concatenated these data to guide an atomic level representation of this protein (table 1). This representation informed the final selection of the H175, LRAP+4, and LRAP-4 tertiary models.

Multimer construction - As there are no oligomer templates with even remote similarity to AMELX, I cannot use template-based modeling (Kittichotirat et al., 2009). Modeling by prescribed symmetry would be highly useful (André et al., 2007), but the core contacts of AMELX are not tight enough to allow use of this tool. Therefore I model the oligomer by hand, placing each monomer in relation to each other as guided by the delineated protein interaction regions discussed above. I iterate reorienting to maximize shape complementarity and potential energy minimization to tighten the model, adding monomers to improve symmetry.

2.2. Residue function prediction

SOAK ion-binding pockets - I performed an exhaustive grid-based search placing ions into all solvent exposed sites in the selected model. The sites are scored with a grid-based exhaustive sampling algorithm, clustered using a variant of an unsupervised hierarchy clustering algorithm, and refined with Monte Carlo minimization. The scoring function was developed from geometrical parameters observed

for inorganic single atom ions coordinating small molecule organics found in many crystallography structures publicly available through the PDB (Berman et al., 2000). This scoring function was verified by ranking binding affinities between proteins and docked ions versus the native binding site <http://protinfo.compbio.washington.edu/soak> (Cheng et al., in preparation).

Meta-functional signature - We developed a new general meta-functional signature with the updated methods developed in chapters 2 and 3. The general MFS compilation was applied to estimate the importance of function for each residue by training the combination of updated algorithms on the same two databases of functionally important residues as in the original MFS (Chelliah et al., 2004; Porter et al., 2004). The result is a new function, mfs2. The analysis described in the text of both the Results and Discussions sections use the original MFS scores.

mfsPO4 - We applied the automated MFS2 retraining tool built for calcium ion binding (chapter 2) and deleterious nonsynonymous SNP (chapter 3) predictions to learn to predict 835 phosphate binding residues in 190 PDB proteins with less than 35% sequence identity. We focus on proteins binding to soluble calcium and phosphate ions in PDB structures as a model for interactions with these molecules when integrated by ionic bonds into the hydroxyapatite mineral.

Stability - I calculated the importance of each residue to structural stability by evaluating the spread of pseudoenergies (RAPDF; Samudrala and Moult, 1998) of serial virtual mutations to the 19 alternate amino acids.

Enamel pellicle similarity - We applied the pellitrix method to the H175 sequence as described in chapter 4.

3. Results

3.1. Full length AMELX

The structure prediction protocol produced a mixture of linear and globular models for the human full length variant (including exons 2,3,5,6ABCD,7: H175) matching distributions seen in the literature (Margolis et al., 2006; Oobatake et al., 2006). The linear models did not produce regular folds, but the carboxy-terminal 40 residues loop back towards the amino-terminal for all linear models. The only consistent contacts occur between the carboxy-terminal residues and the beginning of the intermittent hydrophobic region (residues 45 to 61; backbone shown as purple tube in figure 1c-f)

The predicted globular structures do not exhibit significant regions of typical secondary structure motifs (figure 1a). This is consistent with NMR and CD data for full length AMELX (Le et al., 2006). These models were carried forward for further analysis as candidates for the functional full length AMELX form.

In each human globular full length model, the A domain folds upon itself (backbone shown as yellow tube in figure 1c-f), exhibiting independent stability. A stable fold for the A domain is supported by the ability of a construct composed by only the A domain to produce interactions with full length AMELX (Paine and Snead, 1997), as well as the measurements directly indicating separate folding behavior for this region (Goto et al., 1993; Matsushima et al., 1998). Contact maps for the B domain in the globular models were not consistent. The only other consistent contacts were seen between parts within the second hydrophobic region, specifically between residues 67 to 73 and 98 to 103 (backbone shown respectively as purple and pink tube in figure 1c-f).

No interaction was observed between the glutamine carbonylamides and the main chain carbonyls of proline. This interaction has been postulated to stabilize β-turns or β-spirals via hydrogen bonds and explain the high levels of these amino acids in AMELX (Kelly et al., 2001; Stapley and Creamer, 1999). It is possible that this discrepancy is a failure of the model. It is also possible that some of these carbonylamides instead interact with forming hydroxyapatite.

3.2. LRAP

LRAP-4 is predicted by all structure prediction methods used here to be globular, uniformly presenting with two α-helices occurring for residues 16 to 31 and 51 to 59. These helices match NMR and CD data showing regular helical secondary structure for LRAP (Le et al., 2006). It is interesting to note that the refined model (figure 2a) developed contacts between side chains of the two α-helices, not seen in the original top models. An average diameter of 14.8 Ångstroms was found across the high scoring predictions.

The majority of glutamic acid side chains in LRAP-4 are predicted to extend outward from the protein main chain much in the same way as the α-carboxy glutamic acids of osteocalcin. The parallel orientation of these four glutamic acids away from the high MFS-scoring region and their proximity to alanine-45, previously shown to be close to the hydroxyapatite surface (Shaw et al., 2004) indicates that these residues may function in binding hydroxyapatite.

Part of the first α-helix observed in LRAP-4 is maintained in the LRAP+4 models, directly after the exon 4 insertion, for residues 37 to 45. In all top models, this helix is present and flanked by residues 17 to 22 and 70 to 73. In the consensus model (figure 2b) the residues separated by exon 4 are brought together spatially, creating an alternate conformation to the continuous α-helix seen in LRAP-4. This different conformation may be relevant to cellular signaling (Warotayanont et al., 2008; Warotayanont et al., 2009).

The consistent pattern of secondary structure observed in the top models for the LRAP molecules indicates stable folds that enable function in cellular signaling activity within solvated conditions.

3.3. AMELX model selection using experimental annotation

Experimental criteria for full length AMELX protein structure selection after the formal knowledge-based functions are summarized in table 1 and diagrammed in figure 1b. The 3rd highest RAPDF ranked model produced for H175 is the only globular model which fits all criteria from experimental data (figure 1c-f). In particular, very few globular models present residues 16 to 33 along the protein surface. This region contains bonds cleaved by non-specific endoproteases (Moradian-Oldak et al., 2001), AI-related mutation sites (Lench and Winter, 1995) (Kida et al., 2007), a tryptophan shown to freely rotate on the surface (Oobatake et al., 2006), a phosphorylated serine (Fincham et al., 1994), a lysine available for cross-linking (Brookes et al., 2000; Brookes et al., 2006), and is essential for AMELX-AMELX interactions with the B domain (Paine and Snead, 1997). As well, the selected model is the only model to bring many high MFS-value residues into 3D clusters, which occurs across three adjacent strands within the hydrophobic intermittent region. Thus the H175 model 3 is chosen as the putative full length AMELX structure.

In the H175 model residues 1 to 105 adopt a relatively tight fold with loose construction thereafter. The remaining 70 residues are stabilized by hydrophobic interactions until the B domain and carboxy terminal, which adopt stable folds adjacent along the surface to the A domain. The 46.8 Ångstroms diameter for the chosen globular model corresponds to the 40 to 80 nm diameter measurements of the AMELX nanosphere globular subunit observed with SEM imaging (Du et al., 2005). The structural domains match existing data of separated folds for the A domain, the hydrophobic intermittent region, the B domain, and the carboxy-terminal region (Goto et al., 1993; Matsushima et al., 1998).

The relative positioning of established and predicted functional regions along the protein surface in our top globular model produces chiral projections of AMELX-AMELX interaction regions, thereby enabling patterned presentation of other domains along the exterior surfaces of chains of multimeric constructs. If the A domain, the B domain, and the later part of the hydrophobic intermittent region direct oligomer construction (e.g. tetramer, hexamer), they would present the proximal half of the hydrophobic intermittent region along the surface of such an oligomer, adjacent to one or two of the same regions in the oligomer. The polar carboxy terminal region protrudes orthogonally from the protein surface, exactly fitting models previously proposed based on the relative hydrophilicity of side chains in this region (Margolis et al., 2006; Snead, 2003). The external placement of this hydrophilic region may aid solvent interactions. Multiple carboxy-terminal regions may also function together to either direct orientation of AMELX within oligomers through repulsion, or facilitate inter-oligomer and inter-nanosphere attraction through complimentary polar charge distributions for overlapping, oppositely

oriented carboxy-terminal regions. Repulsive forces between the subunits inside nanospheres would limit coacervation in a predictable manner, which is essential to the controlled assembly of specific sized nanospheres. Attractive forces would provide a specific mechanism for nanosphere interactions, which is postulated to occur in a highly governed manner (Du et al., 2005).

I cannot quantitatively measure the accuracy of our full length AMELX molecular structure without more definitive experimental structure assessment data. From experience with protein structure prediction, I estimate this model to be within 10 Ångstroms root mean squared deviation of the actual structure. Nonetheless I assert that the putative model provides a useful tertiary conformation from which conclusions can be drawn and experiments can be designed.

3.4. Hexamer construction

I challenged the hypotheses regarding oligomer form conceptually explored in the previous subsection, by constructing an oligomer with the AMELX H175 model. I induced coupling of monomer (figure 3a) domains A and B to form a dimer (figure 3b), which left vacant space suggesting placement of another monomer to create a trimer (figure 3c,f). Meanwhile, the long axis of this trimer has one large hydrophobic face, facilitating the binding of another trimer to make a hexamer (figure 3d,g,h). The resulting hexamer successfully buries much of the hydrophobicity (figure 3e), and does not leave a broad face nor concavity to support additional monomer grouping. The resulting model describes "3-d" symmetry: two facing plates of three fold radial symmetry (figure 3c-h). This model perfectly matches that of Du et al., 2005 (figure 3i-n) and the triangular orientations of globular forms within their transmission electron microscopy images (figure 3o). Functional relevance is discussed below.

3.5. Meta-Functional Signatures

MFS scores were generated for all major splice isoforms of human AMELX (figure 5). MFS scores were mapped to the AMELX model (figure 1c-f, 3f-h, 5a, 6b,e), LRAP (figure 2a,b), and OCN (figure 2c,d) structure models for indication of active sites functioning in mineralization, cellular signaling, and protein-protein interactions.

Reproduction of AMELX functional data - Naturally occurring AMELX missense mutations occur at residues uniformly conserved within 52 characterized mammalian AMELX sequences (Delgado et al., 2007) and have been shown to cause missense mutations with corresponding developmental defects in enamel formation (see references within tables 1 and 2). These sites are used to measure the ability of MFS to predict functionally important residues within AMELX (table 2). The comparisons are used as positive controls. They also function to gauge the magnitude of importance given by MFS scores for residues of unknown function. As well, MFS scores for the conserved phosphorylated serine-16

(Fincham et al., 1994) and residues within the established functional regions, A, B (Paine and Snead, 1997), and ATMP (Ravindranath et al., 1999), are indicative of the applicability of MFS to AMELX, and aid in the identification of specific functional residues within these regions.

The AI-related missence mutation sites all are scored above the mean for AMELX MFS scores. MFS scores for these residues range from 0.57 (Z = 0.63, Z = (mfs - mean) / standard deviation) for the p.T21I site to 0.76 (Z = 1.75) for the p.H46L site (table 2). A significantly higher MFS score for proline-22, 0.68 (Z = 1.28), as compared to the adjacent mutation site, 0.57 (Z = 0.63), matches the more destructive phenotype for the hypomaturation defect related to the p.T21I mutation, as compared to the smooth hypoplastic phenotype of the p.P22R mutation (Kida et al., 2007).

The phosphorylated serine-16 is assigned an MFS score of 0.84 (Z = 2.22; table 2), the third highest score in AMELX. The other two serines in the A domain, shown to not be phosphorylated (Fincham et al., 1994), are given much lower scores, thus selecting serine-16 for specialized function. The AI-related missense mutation sites have higher MFS scores in human than mouse, although in vivo recapitulation of the p.T21I, and p.P40T mutations in mice maintain the deleterious loss of function seen in human AI (Moradian-Oldak et al., 2000; Paine et al., 2002).

Proline-169 in M180 has been shown to stabilize AMELX-AMELX interactions for constructs limited to the region surrounding the B domain, while point mutations of this residue within a full length construct did not affect AMELX-AMELX interactions (Paine et al., 2003). These data confirm the prediction of a very low MFS score for this residue, 0.28 (Z = -1.07; table 2).

Reproduction of known osteocalcin hydroxyapatite-binding residues - I validated the ability of MFS to detect specific residues interacting with hydroxyapatite using the most clearly defined examples known, given by Xray diffraction models of porcine (figure 2c; PDB identifier 1q8h; Hoang et al., 2003) and fish osteocalcin (figure 2d; PDB identifier 1vzm; Simes et al., 2003) coordinating a sheet-like distribution of calcium or magnesium ions, respectively. After enabling MFS application to this sequence by simplifying the representation of α-carboxy glutamic acids to the unmodified form, MFS selected all residues coordinating metal ions in the structures, within the top 10 scores for both models (figure 2c,d). Although they adopt a very similar global fold, the sequences for these structures are shown to be substantially different by a ClustalW alignment (figure 2e). As the major contribution to MFS scores is given by sequence analysis, the two structures may be used as somewhat independent verification of the ability of MFS to select hydroxyapatite binding residues.

Predictions for the AMELX A domain - Six of the highest 15% MFS scores occur in the first 33 residues of the A domain, with a mean score equivalent to that of the entire protein (0.46; table 3). However, the region from residues 16 to 28 has a much higher mean score of 0.58, indicating that residues in this region are responsible for the principal function of the A domain, presumably AMELX-

AMELX interactions with the B domain but also mineral interaction. Looking across mineralization proteins, a unique trimer pattern emerges with residues successively bearing functional groups: hydroxyl, hydroxyl, carboxyl. For example the SSD repeats of DSSP have been shown to specifically bind hydroxyapatite, a function maintained by mutation of serines to threonines (also contains a hydroxyl group) and the aspartic acids to glutamic acids (also contains a carboxyl group; Yarbrough et al., 2010). Here serine-16 is the lead of such a sequence: SYE (serine and tyrosine side chains both terminate with a hydroxyl; glutamic acid like aspartic acid terminates with a carboxyl).

The proximal portion of the A domain is not evidenced to play a strong functional role. The occurrence of histidine-9 and proline-10 in the top 15% MFS scores, and similarity between the proximal A domain and the HΦ region, indicate similar function for these regions. Possibilities for the type of function include hydrophobic attraction contributing to AMELX-AMELX interactions (intra-oligomer or intra-nanosphere) and hydroxyapatite binding. The former proposition matches implications from non-specific proteolytic cleavage that the proximal A domain may not be solvent exposed (Moradian-Oldak et al., 2001).

ATMP - The average MFS score of the ATMP region is well above that of the entire protein (0.55; table 3), with glutamic acid-39 scoring highest (0.78). This N-acetyl D-glucosamine, cytokeratin-5 and -14 binding site plays a critical role in tethering nanosphere chains to the glycocalyx of the retracting ameloblasts (Ravindranath et al., 1999; Ravindranath et al., 2000; Ravindranath et al., 2001; Ravindranath et al., 2003). Mutation of the tyrosines in this region has been shown to decrease this function (Ravindranath et al., 2000; Ravindranath et al., 2001), which matches the moderately high scores for tyrosine-36 (0.58) and -38 (0.64). Proline-34 has an even higher score (0.65) than the AI-related proline mutation site (0.62), with the 6th highest sequence conservation score of the entire protein, suggesting these prolines not only hold a specific local fold, but are recognized in protein-protein interactions as well.

B domain - The mean MFS score for the B domain is a low 0.48 (table 3). The four top 15% MFS scores within this region indicate that these residues are the functional constituents of the B domain. The highest MFS score is assigned to threonine-159. This residue is replaced with a proline in most mammals, including mouse, pig, and cow, suggesting that this residue enables a unique important interaction for the B domain in humans. This function would not be discovered by traditional sequence similarity approaches.

Although AMELX-AMELX interaction is measurable for murine constructs not including residues after proline-169 of M180 (Paine et al., 2003), function was noticeably higher for those bearing the subsequent ATDK sequence. Serine-165 in H175, correlating to the alanine-171 in this extended region of M180, is assigned a high score (0.77) suggesting an important role for this residue in AMELX-

AMELX interactions. This site has not been previously attributed to the function of either the B domain or the carboxy-terminal region.

Carboxy-terminal region - If the carboxy-terminal region does interact with hydroxyapatite, it is not the only region to do so. A low mean MFS score (0.48) and low maximum score for this region (0.57; table 3) suggest a low functional contribution of the carboxy-terminal region. This finding contraindicates responsibility for the essential AMELX functions in controlling and seeding mineral formation. However, the consolidation of the majority of charged residues in amelogenin indicates a specific function.

HΦ Intermittent region - Strong evolutionary conservation, uniqueness among similar sequences, and observed rate of occurrence in functional sites is found for many residues within the hydrophobic intermittent region (the HΦ region, residues 45 to 149, main chain shown as purple, pink, and red in figure 1c-f), previously thought to bear no function due to presumptions of internalization for this region. In particular, the regions from serine-54 to histidine-68 (main chain shown as purple in figure 1c-f) and from histidine-91 to histidine-99 (main chain shown as pink in figure 1c-f) contain many high MFS scores in both human AMELX (table 3). Here I divide the hydrophobic intermittent region into putative functional regions corresponding to these high MFS score clusters, termed the proximal HΦ region and the central HΦ region, respectively.

The highest mean score for all AMELX regions is found for the central HΦ region (0.58), which is only rivaled by that for the distal A domain, and that for the proximal HΦ region (0.53; table 3). Our top model predicts that nine of the eleven top 15% MFS scoring residues in the proximal and central HΦ regions fall into a tight spatial cluster (figures 1c-f, 3a-h, 5). The combination of high regional mean MFS scores, many of the highest MFS scores in AMELX, and spatial proximity for the highest MFS score residues, creates strong evidence of an important enzymatic site existing in the hydrophobic region of AMELX.

A crucial role for histidines in AMELX function is suggested by the scoring of all ten histidines between histidine-46 and histidine-99 within the top 15% of the entire protein. The MFS amino acid type contribution scores this amino acid high, as the databases with which the method was developed contain many coordinating intermediate states within enzyme active sites. Many of the histidines in the proximal and central HΦ regions occur with adjacent glutamines. The carbonylamide side chains of glutamines have polar distributions viable to support histidines in stabilizing compact positioning of the charged hydroxyapatite constituents, and may represent paired coordination groups matching the polarities of the calcium and phosphate or calcium and hydroxyl groups of hydroxyapatite.

Two of the fourteen histidines in AMELX are scored much lower than the others. These occur in the distal HΦ region, which portrays low importance for this region. The sequence spanning the central

HΦ region and the MMP-20 cleavage site are assigned a mean MFS score of 0.42, with a maximum of 0.62. A smaller sub-region with a higher mean MFS score of 0.46 is found from histidine-121 to proline-135 (table 3). I call this sequence the distal HΦ region.

The MFS distribution for M180 shows a more clear cluster of high scores in the distal HΦ region, from serine-125 to proline-137 having a mean of 0.58. This region may bear a different function in mice, due to non-conserved serines occurring in the distal HΦ region of M180. Thus, the recent experiments exploring the function of this region in mice (Jin et al., 2007) might not be relevant to humans.

The lower importance placed on histidines within the distal HΦ region and the low MFS scores in H175 indicate a separable, non-chemically mediated contribution. Non-specific hydrophobic attraction contributing to AMELX-AMELX interactions is a viable putative function for this region, matching that suggested by interpretation of Y2H data (Paine and Snead, 1997).

LRAP-4 - Residues 12-26 contain all 10 of the highest MFS scores in LRAP-4 and a high mean (0.69) relative to that of the entire protein (0.39; table 3), strongly implicating this as the region recognized in cellular signal transduction. This hypothesis is supported by the majority of this region presenting as an α-helix in all top structural models generated by multiple protein structure prediction methods and secondary structure prediction servers (figure 2a).

LRAP+4 - The highest MFS scoring residues in LRAP+4 are essentially the same amino acids transposed from LRAP-4, after accounting for the exon 4 insertion (table 3). Nonetheless, a more markedly defined hot spot for function is seen for the KWY sequence (residues 38 to 40). These residues fall within an α-helix existing as the only predicted stable secondary structural motif in the protein (figure 2b). The consensus model brings together the residues of this α-helix with the other highest 10 MFS scoring residues, into spatial proximity (figure 2b). The result is a localized surface presentation of all the high MFS scoring residues, substantially different but analogous to the continuous string seen in LRAP-4. This externalized set of highly conserved residues, with most side chains projecting normal to the protein surface, are excellent candidates for signaling function via protein-protein interactions.

Exon 4 produces an established functional alteration in LRAP (Tompkins and Veis, 2002; Tompkins et al., 2005; Veis, 2003), though a mechanism involving either disruption of tertiary structure or novel chemical contribution has not been elucidated. The fourteen residue exon 4 insertion appears vital to the creation of a differentiable signaling region between the LRAP splice isoforms. The structure and function prediction presented here indicate that this insertion may create a new signal motif using the same high MFS scoring residues for both proteins. Fourteen residues is more than long enough for such a loop. Such a simple structural function would explain the high variability for this region across AMELX genes in mammals. Low MFS scores within the exon 4 region indicate a more structural than

chemical contribution to function. It is particularly interesting that in LRAP+4, the HMM scores are above zero for only residues 12 to 18 and 33 to 47, perfectly deleting exon 4. Although it is always possible to align a few sequences to any protein region, the hidden Markov model expunges these data as insignificant. MFS analysis does not support direct functional contribution of exon 4 in LRAP+4; but rather a unique function by disruption of the LRAP-4 functional domain.

3.6. Structural stability scores

The top four most structurally important residues of the putative H175 model are found between valine-78 and methionine-84 of the intermittent hydrophobic region, with many other high scores in the vicinity. This localized, structurally important region occurs between two of the highest MFS scoring regions, the proximal HΦ and central HΦ regions. This is a fascinating find, as the most common contacts conserved within the top twenty structural models occur between these two regions. The putative model presents these high MFS regions with essentially all residues having contacts with each other. I conclude that the structurally important intervening region stabilizes the relation of the two HΦ regions, creating a functional domain.

Other clusters of structurally important residues are found throughout the protein, occurring between the ATMP region and the proximal HΦ region, within the proximal HΦ region, directly after the central HΦ region, within the distal HΦ region, and between the B domain and the carboxy-terminal region. High stabilization intervening between functional regions indicates importance for the relative orientation of functionalities in AMELX.

Another interesting finding in the structural stability scores to AMELX is the placement of residue p.P22 in the top 10% of the structural importance measure. This indicates a crucial contribution of this amino acid at this position to the stability of the domain and the overall protein structure. This finding provides offers explanation for the markedly different phenotype for the AI-related missense mutation at this site as compared to the other three sites, particularly that for the adjacent missense mutation site p.T21 (Kida et al., 2007).

As mentioned above, when it is the carboxy-terminal most residue, keeping the identity of position 169 as proline enables B domain AMELX-AMELX interactions in M180, presumably by stabilizing internal stability (Paine et al., 2003). However, constructs with mutated position 169, but maintaining the four amino acids after this residue, ATDK, produce much stronger interactions. One residue within this ATDK region, aspartic acid-167 (correlating to 172 in M180), is rated as the most important structural site distal to the MMP-20 cleavage site and the 9th highest stability score in AMELX. Thus, it may be aspartic acid-167 that promotes the local fold, enhancing B domain function.

3.7. Specific calcium and phosphate binding predictions

Benchmark of phosphate binding prediction - In figure 4 we demonstrate similar accuracy for predicting phosphate ion binding residues as for calcium ion binding residues in chapter 2. This represents more evidence that sequence-based methods can increase accuracy through specification.

Benchmark of general function prediction - Applying the new protocols to retrain mfs results in more accurate predictions (figure 5), which demonstrates the value of sequence-based structural inferences of function.

Reproduction of known osteocalcin metal ion binding sites - The two known Xray diffraction structures of porcine (with calcium; 1q8h; Hoang et al., 2003) and fish osteocalcin (with magnesium; 1vzm; Simes et al., 2003) were used as positive controls for finding metal ion binding sites in proteins interacting with hydroxyapatite, as these crystal structures demonstrate specific placement of externally bound metal ions thought to mimic hydroxyapatite binding function known for osteocalcin (figure 2c,d). It is particularly appropriate for the case of AMELX, as OCN also maintains a random coil structure in solution when no metal ions are present. The presence of physiologic levels of calcium induces strong α-helix secondary structure (Hauschka and Carr, 1982) and change of secondary structure for LRAP and AMELX (Le et al., 2006).

Two predicted magnesium ion binding sites match crystal structure sites for two magnesium ions in 1vzm. Other calcium and magnesium sites do not match known locality, but are found to occur along the same face of the protein, coordinating the same residues (figure 2c,d). It is important to note that this method is significantly affected by small inaccuracies produced by protein structure prediction methods.

SOAK Comparison - Although the SOAK method is not benchmarked to compare affinity for metal ions between proteins, the scoring function used to select optimal binding sites produces absolute scores. As the grid-based search space used in SOAK is exhaustive, slight differences in scoring arise between proteins of different sizes and numbers of clefts, mainly due to the effects of clustering. Scores can be compared between proteins in an analogous but not exact manner.

The mean SOAK scores for the predicted calcium ion binding sites of each protein are used to predict relative affinity, with lower scores indicating better binding. LRAP-4 has a mean of 4.68, which closely approximates the values of 4.07 and 4.82 for the osteocalcin structures. The higher values of AMELX and LRAP+4, are 8.67 and 11.35 respectively, indicating lower calcium ion affinity for these molecules as compared to LRAP-4 and osteocalcin. This trend matches the six-fold higher affinity demonstrated for LRAP-4 as compared to AMELX (Le et al., 2006).

LRAP - All LRAP-4 predicted calcium ion binding sites occur within the 8.0 Ångstrom constraint to alanine-45 shown by solid state NMR studies (Shaw et al., 2004), or lie in a plane continuous with these other ions (figure 2a). This model supports LRAP-4 binding the hydroxyapatite surface such that the

high MFS score region hypothesized to control cellular signaling function is displayed away from the surface.

The predicted LRAP+4 model does not support interactions with a planar mineral surface. LRAP+4 appears to adopt a more globular structure with all high MFS scoring residues presenting in a continuous surface not likely to bind calcium (figure 2b).

Ion binding scores for LRAP-4 are more favorable than those for LRAP+4. Metal ion binding for the LRAP proteins appears to congregate around glutamic acids, as found for the α-carboxy glutamic acids of the osteocalcin proteins.

AMELX - Eight calcium ion binding sites were predicted by SOAK. Overlap is found between four of the eight calcium ions and half of the highest 15% MFS scores, all occurring within a confined spatial cluster (figure 7a). These four calcium ions occur along a linear arrangement with one at each orifice of a pore-like structure. The two best scoring calcium ions for the putative AMELX structure are found in the center of this chain of ions. All five residues predicted to coordinate the single highest calcium ion site prediction fall within the top ten MFS scores. The superimposition of predictions for high functional importance by MFS, SOAK, and that developed by sequence extrapolation of phage- and cell-display hydroxyapatite binding sequences (Ersin Emre Oren and Mehmet Sarikaya, personal communication; figure 7, main chain in red ribbon) strongly indicates a novel functional region participating in mineralization. I extract this portion of the structure as a consensus functional unit in figure 7b.

As mentioned above, the precise spatial analysis used in the SOAK method is particularly sensitive to the small inaccuracies endemic to structure prediction methods. However, contacts between the proximal and central HΦ regions occurred for all top globular models, indicating a very high probability of spatial proximity for these regions which can be understood to mimic naturally occurring contacts within the elusive AMELX tertiary structure.

Calcium binding residue prediction - mfsCa predicts calcium ion binding by the following residues N14, E18, E39, D157, T159, E161, D167, E172, E173, and D175 (figure 5, green circles), which form three clusters along the sequence and two solo residues. These residue may come together in the tertiary structure to form two separate functional sites. The distribution of residues predicted to bind calcium by sequence match the four extraneous SOAK structure based predictions (figure 7a, green spheres). Overlapping predictions from the original MFS, mfs2, mfsCa, and SOAK suggest four additional calcium interaction sites for AMELX.

Phosphate binding residue prediction - mfsPO4 predicts many phosphate binding residues in AMELX (figure 5, purple diamonds). Eleven of fourteen histidines in the protein are predicted to specifically bind phosphate. As the unique function of this protein is to guide the fusion of solvated calcium and phosphate into solid hydroxyapatite, many of these predictions may be accurate. Many hydroxide

bearing residues are also predicted to bind phosphate. The predictions cluster with calcium binding residue predictions in the A domain YINFSYE sequence noted for high MFS scores above, discussed further below. The B domain DLTLE sequence also clusters high MFS, mfsCa, and mfsPO4 predictions, as does the carboxy terminal KREEVD sequence.

Enamel pellicle peptide similarity - pellitrix predicts four enamel interaction regions and two specific noninteraction regions (figure 5, blue line). Dissection of the favored interaction regions suggests the forming enamel interactions arise from the following segments: 1-10, 46-68, 90-127, and weaker interaction for 151-163. The first three of these interaction regions present strong phosphate binding predictions, and are predominantly comprised of histidine, glutamine, and proline. The 46-68 region is enriched with three hydroxyl bearing residues, and similarly the first five residues of the 90-127 region has two hydroxyl bearing residues. The 46-68 region spans two groups of high MFS and mfs2 score predictions, while high function scores could parse the 90-127 region into a functional region (91-99), and a support region (100-127).

The two regions predicted lower than random expectation match the FPM enamelysin cleavage site and the YINFSYE region described in the paragraph above. The FPM outer residues are perfectly conserved in mammals, as this is the first step in degradation of AMELX - a necessary part of its function. It follows logically that within the forming hydroxyapatite environment, the component enabling the first step of breakdown would evolve to stay away from hydroxyapatite. Positive prediction for important function, calcium ion interactions, and phosphate interactions, but negative prediction for interaction with hydroxyapatite surface is interpreted as phase catalysis for the YINFSYE region, i.e. nucleation.

3.8. Composite function from hexamer model

I mapped the top 15% MFS scores to the hexamer model and docked calcium ions with SOAK (figures 3f,g). The principle consensus region maps to the solvent exposed periphery. Upon application of molecular dynamics energy minimization to further stabilize the structure, the channel of calcium ions opens, exposing the joint 46-68 and 91-99 predicted functional site (figure 3f).

At the center of the hexamer construct a new functionality emerges (figures 3f,g, 6). This site displays a convergence of high MFS, mfs2, mfsCa, and mfsPO4 scoring residues from the A domain and B domain, including the YINFSYE region discussed in the section immediately above. The SOAK predictions increase for this site, resulting in as many calcium binding sites as for the rest of the construct (figures 3f,g, 6b).

The predicted central functional site motivated a comparison of the symmetry of our hexameric model to the symmetry of the hydroxyapatite unit cell (figure 8a to 6b, 6e to 6f). The 30 degree offset

facing units with three fold symmetry ("3-d") matches perfectly. Thus I rebuilt the YINFSYE motif and surrounding residues around the hydroxyapatite unit cell (figure 8c,d). All functional groups align, supporting coordination of this central site as a nucleator.

On the outset I anticipated finding some cleft or internal functional site that would nucleate hydroxyapatite (or a seed calcium phosphate intermediate such as octacalcium phosphate). This was not found with the monomer model, but instead emerges from the functional hexamer (figure 8). Incorporating the transiency of AMELX oligomer formation (figure 3i-p; Du et al., 2005) describes our predicted kinetic model for hydroxyapatite formation: oligomer assembly catalyzes calcium phosphate nucleation at the central site, while the peripheral site that does not require oligomeric assembly matures the mineral into hydroxyapatite.

3.9. Correlations with bench results

Peptide design - Peptides were designed using the murine amelogenin sequence by Ersin Emre Oren, Ram Samudrala, Mehmet Sarikaya, and the author. Phage- and cell-display methods (Ersin Emre Oren and Mehmet Sarikaya, personal communication) were the most important, and produced ADP1, ADP2, ADP4, ADP5, and ADP6. The MFS and SOAK calculations strengthened the positive predictions for ADP1, ADP2, and ADP4. The literature review described here, in face of the low functional predictions by the three methods for the carboxy-terminus, inspired ADP3. The spatial concordance of the predictions of these three methods on the predicted structure as described here designed ADP7.

Briefly, the following regions were predicted to bind hydroxyapatite: ADP1, ADP2, ADP4, ADP7; and the others were predicted to not bind hydroxyapatite. The portion of the ADP5 region arising from the A domain (the first 8 residues) was predicted by MFS to be functionally important, and by SOAK to bind calcium ions. Creation of the corresponding ADP5 peptide predated hexamer modeling and nucleation predictions.

Personal communication - Gungormous, Wilson, So, Hnilova, Chang, Fong, Oren, Tamerler, Sarikaya, and others of GEMSEC tested these peptides for binding hydroxyapatite, catalyzing mineralization, and other specific interactions with calcium. Access to data courtesy of Mehmet Sarikaya and GEMSEC.

Hydroxyapatite binding - Surface plasmon resonance and quartz crystal microbalance experiments were performed as described in Tamerler et al. (2006). The ADP1, ADP2, ADP4, ADP7 peptides bound Conversely, the ADP5 peptide seems to egress actively from the hydroxyapatite surface, as the measured value for this peptide is less than that of the weakest control binder. This finding actually fits logically with the model of ADP5 catalyzing nucleation of calcium phosphate crystals, as this should occur far enough away from the existing mineral surface for the protein to exit.

Effects on a precipitating solution - Stoichiometrically equivalent amounts of each peptide were added to a precipitating solution of calcium phosphate and alkaline phosphatase. Precipitate and supernatant were retrieved from each condition at time points ranging from 30 min to 24 hours. Imaging was performed with scanning electron microscopy, loss of calcium measured by spectrometric absorbance, and remaining calcium measured by spectrometry of color changing chelator.

ADP7 and M180 (the full length AMELX construct) precipitated calcium by highly similar slow kinetics relative to the other conditions and form crystal morphologies separable from the others. Discernable plate-like morphology is observed for these two constructs by one hour. At roughly three hours enamel-like elongated crystal bundles self organize for both. These data suggest that ADP7 contains the long sought unique mineralization region of AMELX.

ADP5 catalyzed calcium precipitation, whereas addition of any other peptide maintains calcium in solution relative to the peptide control. The morphology of precipitates are unique with rings roughly five times smaller than the spherical forms produced by any other peptide or unguided aggregation. Clearly this peptide is carrying out a unique function.

Hydroxyapatite formation - Stoichiometrically equivalent amounts of each peptide were added to nonprecipitating calcium phosphate solution.

The ADP7 construct, and not any constituent peptide (ADP1, ADP2, ADP4), induces the formation of a precipitate with Xray diffraction patterns matching that of the prototypical hydroxyapatite diffraction pattern, equivalent to that formed in the presence of full length AMELX. The ADP3 peptide that was predicted to not catalyze the formation of hydroxyapatite, despite many papers assuming this region to carry out the unique functionality of AMELX based on the accumulation of charged residues, does not catalyze the formation of hydroxyapatite in these conditions. Intriguingly, the ADP5 peptide does not catalyze hydroxyapatite formation in these conditions either. Thus it is likely that the underlying function is to knock out mineral seeds from solution, rather than produce hydroxyapatite directly.

4. Discussion

4.1. Benchmark

The function prediction methods MFS and SOAK are shown to be accurate in selecting residues that interact with hydroxyapatite and finding external metal ion binding sites representing hydroxyapatite calcium ions (figure 2) for two proteins known to have individual amino acid resolution for hydroxyapatite interactions: porcine and fish Osteocalcin (Hoang et al., 2003; Simes et al., 2003). All six amino acids coordinating metal ions in each of the crystal structures for these proteins were found

within the top ten MFS scores, and the ion sites were closely matched by SOAK predictions using calcium and magnesium ions.

Modeled structure and constituent function - For the first time, an atomic level resolution structure model of a folded, functional AMELX molecule is posited. The structure model portrays spatial clustering of known and predicted functionally important residues, forming relatively isolated constructs for the A and B domains (Paine and Snead, 1997), a protruding carboxy-terminal region, solvent accessible presentation for specific and non-specific proteolytic cleavage sites (Moradian-Oldak et al., 2001; Ryu et al., 1999), tryptophans known to freely rotate (Oobatake et al., 2006), as well as lysines known to be available for binding intermolecular cross-linkers (Brookes et al., 2000; Brookes et al., 2006). The possibility of the co-existence of unfolded and folded AMELX protein is supported by the concept that many proteins exist transiently as folded during function and relatively unfolded when not bound to the physiologic partner protein or metabolic substrate (Shoemaker et al., 2000).

I elucidate the functional and structural contribution of each amino acid in H175. The MFS predictions depict a model of functional importance that fits all existing data delineating structure and function by region or individual amino acid, including positive identification of AI-related mutation sites (Collier et al., 1997; Hart et al., 2002; Kida et al., 2007; Lench and Winter, 1995), the phosphorylated serine (Fincham et al., 1994), the A and B domains (Paine and Snead, 1997), and the ATMP region (Ravindranath et al., 1999; Ravindranath et al., 2000; Ravindranath et al., 2001; Ravindranath et al., 2003), as well as discrimination of a proline known to not be important for function (Paine et al., 2003).

4.2. Protein interaction

Analysis of the Y2H data first demonstrating AMELX-AMELX interactions reveals a contribution distributed throughout the region from residues 110-157 (Paine and Snead, 1997). Our findings here concur that this latter portion of the hydrophobic region, termed the distal HΦ region, facilitates AMELX-AMELX interactions. The uniqueness of the sequence suggests a unique function: framing the architecture of mammalian enamel. I hypothesize that in additional protein-protein interaction experiments, a construct including the distal HΦ region and the B domain would reproduce AMELX-AMELX interactions with the A domain seen for the full length construct. As well, the hexameric construct portrays this region connecting to A and B domains.

4.3. Unique mineralization function

Substantial efforts have been devoted to understanding AMELX structure, function, and interactions. Yet the characteristic function of this protein is not understood. For example the A and B domains regions are clearly assigned function in AMELX-AMELX interactions (Paine and Snead,

1997). Yet the 106 intervening residues have only been correlated to hydrophobic interactions. Many of these residues are conserved across all mammals, but none in reptiles (Delgado et al., 2007), which do not exhibit higher level rod-interrod architecture. I refute the possibility that the intermittent region could be isolated from the protein surface, particularly as there are more residues in this region than the entire rest of the protein. As some parts of this region must be presented on the monomer or oligomer surface, the region is in a position to interact with the forming hydroxyapatite. Few experimental studies have assessed this region.

The intermittent hydrophobic region is conserved across mammals (Delgado et al., 2007), precisely framed by MMP-20 cleavage sites (Ryu et al., 1999), and yet no sequence has been observed with higher than random similarity; if there were no particular function it would have long ago been lost to antiquity. Residues 45-109 are the only ones for which there is no evidence for function in maintaining higher order aggregation; thus this region sticks out as being potentially responsible for guiding hydroxyapatite mineralization.

I find such a novel mineralization functionality in this portion of the intermittent hydrophobic region through the mutual interaction of two regions I term the proximal HΦ region and the central HΦ region. These regions present higher levels of functional importance than any other in AMELX, span the region of highest structural importance, are predicted to coordinate a channel-like configuration of calcium ions, score the highest on the GEMSEC hydroxyapatite binder matrix method, the pellitrix method (described in chapter 4), and are exposed on the hexameric model. In the full hexamer model the predicted mineralization region is not only preserved away from the monomer interfaces, but presents on the surface, exposed to forming hydroxyapatite.

Interestingly, in a very early study the hydrophobic intermittent region was postulated for mineralization function, although this was for a more distal segment (Renugopalakrishnan et al., 1989), including the specific prediction of a calcium channel, not dissimilar from our own current predictions.

4.4. Refuting a previous hypothesis

It has long been thought that the carboxy-terminal region is principally responsible for controlling mineral formation (Snead, 2003). However, a construct with the thirteen residue carboxy-terminal peptide truncated is still able to direct crystal growth (Beniash et al., 2005), and our ADP3 peptide made from the sequence of the carboxy region does not affect mineralization nor interact with mineral differently than negative controls. The newer versions of MFS: mfs2, mfsCa, and mfsPO4, all predict high functionality for this region, whereas the original version did not. This trend arises mostly from the exponential growth of the protein sequence databases between 2006 and 2010. For example the hidden Markov model of the alignment produced by searching the 2006 UniProt database does not include the carboxy ten residues, whereas these same residues receive high scores in the current version.

The new analysis presented in figure 6 predicts significant function. The AMELX protein displays abnormal sensitivity to pH, temperature, solute concentration, etc., so it is possible that the conditions for the function of this region have not yet been reproduced experimentally. However, the ability of MFS to find catalytic sites indicates that the carboxy-terminal region is not the most critical functionality in AMELX. Furthermore, if this region were responsible for guiding hydroxyapatite mineralization and not inter-oligomer interactions or hydroxyapatite binding, the LRAP molecules would presumably be able to similarly guide hydroxyapatite mineralization.

4.5. LRAP±4

Predictions of structure, functional importance by individual amino acid, and calcium ion binding are also made for LRAP-4 and LRAP+4. I find a specific string of amino acids to maintain the dominant function in LRAP-4, and that these residues are similarly involved in function for LRAP+4, but with a completely different structural presentation. The two helices found in all top LRAP-4 models induce a planar presentation of side chains, particularly for glutamic acid, as seen in the Osteocalcin crystal structures wherein this enables adhesion to a flat hydroxyapatite surface (Hoang et al., 2003; Simes et al., 2003). Calcium ion prediction sites match the locality shown for alanine-45 in LRAP-4 (Shaw et al., 2004), and form a plane-like distribution analogous to a hydroxyapatite surface. Calcium ion interactions are predicted to be principally mitigated by glutamic acids, concurring with primary and tertiary structure analogies to Osteocalcin.

Comparison of calcium ion binding scores between LRAP-4 and AMELX match observations of higher calcium ion chelation for LRAP-4 (Le et al., 2006). These values for LRAP-4 are similar to those seen for both Osteocalcin structures, but comparison with those of LRAP+4 indicates less favorable calcium ion and hydroxyapatite interactions for LRAP+4. Remembering that LRAP is completely represented in AMELX, the markedly different ability of LRAP to strongly bind calcium indicates a significant conformational difference for the calcium ion binding amino acids, which is represented by the putative models for each.

4.6. Concluding statement

The model I developed has found many if not all unique functionalities hidden within the puzzle of AMELX. The model describes dynamic hexamer assembly bringing together the YINFSYE region to nucleate calcium phosphate intermediates, which are then matured by the continuously exposed proximal and central HΦ regions (figure 9).