Molecular Dynamics Simulations of the Intrinsically Disordered Protein Amelogenin

Authors:

Alessandra Apicella,[a] Matteo Marascio,[a,b*] Vincenzo Colangelo,[b] Monica Soncini,[b] Alfonso Gautieri,[b] Christopher J.G. Plummer[a]*

a. Laboratoire de Technologie des Composites et Polymères (LTC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 12, CH-1015 Lausanne, Switzerland

b. Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

** These authors contributed equally to this work.*

Corresponding author:

Monica Soncini

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

monica.soncini@polimi.it

The Version of Record of this manuscript has been published and is available in "Journal of Biomolecular Structure and Dynamics" 01 July 2016 http://www.tandfonline.com/10.1080/07391102.2016.1196151

Abstract

Amelogenin refers to a class of intrinsically disordered proteins that are the major constituents of enamel matrix derivative (EMD), an extract of porcine fetal teeth used in regenerative periodontal therapy. Modifications in molecular conformation induced by external stresses such as changes in temperature or pH, are known to reduce the effectiveness of EMD. However, detailed descriptions of the conformational behavior of native amelogenin are lacking in the open literature. In the present work, a molecular model for the secondary and tertiary structure of the full-length major porcine amelogenin P173 was constructed from its primary sequence by replica exchange molecular dynamics (REMD) simulations. The REMD results for isolated amelogenin molecules at different temperatures were shown to be consistent with the available spectroscopic data. They therefore represent an important first step towards the simulation of the intra- and intermolecular interactions that mediate self-organization in amelogenin and its behavior in the presence of other EMD components under conditions representative of its therapeutic application.

Keywords: amelogenin, molecular dynamics, protein structure prediction, thermal stress.

Introduction

The regenerative process initiated by the precipitation of enamel matrix derivative (EMD) proteins at the site of periodontal defects has been the subject of various *in vitro* and *in vivo* studies. However the precise role of the individual components of the EMD in its conformational response to conditions representative of its therapeutic application remains poorly understood (Lindskog, 1982; Slavkin et al., 1989). It is nevertheless generally recognized that maintaining the native conformation of the EMD and its ability to self-organize are crucial to its biological effectiveness, as well as the stability of EMD-based products intended for dental use against premature precipitation during long-term storage. Indeed, environmental stress associated with the manufacture, heat treatment and storage of protein-based therapeutics is generally associated with denaturation and loss of function, so that it is of considerable practical importance to understand the underlying mechanisms for these processes (Kiefhaber, Rudolph, Kohler, & Buchner, 1991; Lumry & Eyring, 1954; Wiedemann-Bidlack, Beniash, Yamakoshi, Simmer, & Margolis, 2007). Amelogenin, which accounts for about 90 % of the total mass of EMD and is also the most widely investigated EMD protein, is known to play a key role in the formation of enamel tissue, and its conformational behavior is assumed to dominate that of EMD.

Amelogenin is classified as an intrinsic disordered protein (IDP). Contrary to classical proteins, IDPs or structured proteins with intrinsically disordered regions (IDRs), adopt a spectrum of states, ranging from unstructured to partially structured conformations, that are responsible for protein function regulation in several biological processes (Banerjee, Chakraborty, & De, n.d.; Banerjee & De, 2015; Das & Pappu, 2013; Espinoza-Fonseca & Kelekar, 2015; Leonova & Galzitskaya, 2015; Lindorff-Larsen, Trbovic, Maragakis, Piana, & Shaw, 2012; Yadav, Rai, Hosur, & Varma, 2015). Although amelogenin is considered to be an IDP, its secondary structure is locally well-defined (Delak et al., 2009), and it has been shown that at least one variant of porcine amelogenin, consisting of 173 residues, self-assembles to form relatively ordered structures in aqueous media as the pH is raised (Beniash, Simmer, & Margolis, 2012). Amelogenin self-assembly is known to be sensitive to solvent conditions, “nanospheres” of 5 to 200 nm in diameter having been detected by dynamic light scattering depending on pH, temperature, and ionic strength (J Moradian-Oldak, Leung, & Fincham, 1998). Moreover, large-scale aggregation is thought to originate from hydrophobic interactions between these nanospheres subsequent to thermal stress, which are in turn associated with specific amelogenin domains (Wen, Moradian-Oldak, Leung, Jr., & Fincham, 1999). However, while considerable insight into the conformational behavior of amelogenin and its biological function has been gained from spectroscopic and controlled proteolysis, for example (J. Moradian-Oldak, Jimenez, Maltby, & Fincham, 2001; Paine & Snead, 1997), the experimental study of individual amelogenin molecules is complicated both by their intrinsic disorder and their tendency to aggregate (B. Aichmayer et al., 2005; Barbara Aichmayer et al., 2010; Zhang, Ramirez, Liao, & Diekwisch, 2011).

In an effort to gain further understanding of the response of EMD to its environment, we therefore carried out a computational study of amelogenin at pH 5 at *T* from 4 to 80 °C, which is the regime of immediate practical interest. A model of the full-length native major porcine amelogenin P173 was constructed from its primary sequence by means of replica exchange molecular dynamics simulations (REMD), a widely used computational tool for the prediction of protein folding (Sugita & Okamoto, 1999). Previous studies successfully performed Monte Carlo simulations of IDPs up to 59 residues (Das, Crick, & Pappu, 2012; Das & Pappu, 2013). On the other hand, molecular dynamics (MD) simulations explored the possibility to investigate large IDPs up to 94 (Colson, Thompson, Espinoza-Fonseca, & Thomas, 2016) residues and for up to 200 ms (Lindorff-Larsen et al., 2012), also focusing on interesting aspects such as different outcomes obtained using different available force fields (Zhou, 2007) and folding of protein segments (Cino, Choy, & Karttunen, 2012; Espinoza-Fonseca & Kelekar, 2015) However, MD simulations of IDPs made of hundreds amino acids is still computationally challenging. REMD represents then a useful tool for the assessment of the protein folding of full size proteins, giving the possibility to explore the structure of IDPs as large as amelogenin. An important rationale for the use of REMD simulations is that it allows extended sampling of the free energy space beyond regimes that are directly accessible to nuclear magnetic resonance (NMR) studies, which require dissolution of the protein molecules (Espinoza-Fonseca, 2009). Even so, the available NMR data provide a convenient yardstick for assessing the results of the simulations under specific conditions, as will be discussed further in what follows.

Materials and methods

Replica Exchange Molecular Dynamics simulations

The GROMACS code (Van Der Spoel et al., 2005) and the CHARMM22/CMAP force-field were used for all the simulations (Mackerell, 2004). The amelogenin P173 primary sequence was obtained from the Uniprot data bank (entry number P45561). The initial structure has been assembled using PyMOL, creating the initial simulation structure on the basis of the primary structure available on Uniprot (amelogenin P173, Sus Scrofa). Because REMD simulations of the whole amelogenin protein were prohibitively expensive in terms of computing time, its primary sequence was divided into five shorter subsystems (Fig. 1):

Met1-Tyr17 and Gly8-His32 (encompassing the N-terminus domain), Met29-Pro87 and Pro84-Ser151 (encompassing the central domain), and Met149-Asp173 (the C-terminus domain). The total length of each subsystem was chosen (i) to avoid discontinuities within known putative secondary structures (Zhang et al., 2011) (e.g. α-helices) and (ii) to give a reasonable compromise between computational cost and the number of subsystems. Overlapping sequences were included in adjacent subsystems for the purposes of the merging procedure. In a preliminary step, each of the five subsystems was energy minimized for 50'000 steps using the conjugate gradient algorithm and equilibrated with 1 ns MD simulations at $T = 4$ °C using the NPT ensemble. Non-bonding interactions were computed using a neighbor cut-off list at 1.35 nm, with a switching function between 1.0 and 1.2 nm. Folding was subsequently modelled using REMD simulations. The number of replicas for the successive subsystems was set to 13, 16, 21, 23 and 12, respectively, with T ranging from 4 to 273 °C. The set of temperature were obtained with the temperature generator for REMD-simulations provided by http://folding.bmc.uu.se/remd/ and using an exchange probability of 0.3 (Patriksson & van der Spoel, 2008). Replica exchanges were attempted between neighboring pairs every 5000 integration steps (10 ps) and the simulation time was 100 ns. All MD simulations were carried out in a generalized Born implicit solvent with constraints on the vibration of bonds involving hydrogen atoms, thus allowing a 2 fs time step. A Berendsen thermostat was used to maintain each system at the reference T.

Clustering

A clustering technique based on a root mean square deviation (RMSD) cut-off was applied to frames of the last 50 ns of REMD simulations using the Clustering plug-in available in the Visual Molecular Dynamics (VMD) package (Mackerell, 2004). Briefly, the clusters are defined by an arbitrary RMSD cut-off of the structures at the same temperature, which was chosen according to the structure distribution and varied from 0.3 to 0.4 nm. We then determined the representative structures for the most populated clusters based on the average RMSD within the cluster. In particular, the center of the cluster was determined as the average of all the structures belonging to the given cluster. Among these structures, the structure closer to the center of the cluster was then chosen to be the most representative of that given cluster. These representative structures define the putative conformation of each amelogenin subsystem at a specific value of temperature. The resulting conformations were displayed using VMD and secondary structures analyzed with the Sequence Viewer extension of VMD (based on the STRIDE algorithm) (Frishman & Argos, 1995).

Full-length P173 amelogenin models

The full-length amelogenin structure was built (using HyperChem Pro 6.0) based on the assembly of the most populated clusters of the five amelogenin subsystems obtained at the end of the REMD simulations. Structural alignment of the overlapping sequences of the subsystems was performed. This procedure led to two different initial models of amelogenin (P173a, P173b). Moreover, because the histidine-rich domain (Leu46-Pro87) showed two equiprobable structural clusters (see the results section), a third initial model was generated using in the structure of model P173a the structure associated with the secondary cluster (P173c). The three initial models were minimized and equilibrated for 100 ns using a classical MD simulation in explicit water (SPC water model). The RMSD of the structure was monitored to guarantee its stability. Finally, the full-length amelogenin structures were analyzed with the STRIDE algorithm in order to verify the stability of the secondary structures predicted by the REMD simulations.

Radius of gyration and solvent accessible surface area from the full-length amelogenin structural model

The radius of gyration was calculated from the MD simulation trajectories by means of rgyr GROMACS algorithm, which calculates the RMSD values between the amelogenin center of gravity and the amelogenin residues. The Solvent Accessible Surface Area (SASA) algorithm, also available in GROMACS, was used to assess the solvent accessible surface (SAS) for each amino acid residue.

H-bonds and salt bridges

The presence of H-bonds (defined by donor-acceptor distance of 3Å, and an angle < 20°) was investigated at the beginning (first frame) and at the end (last 50 ns) of the MD simulation in order to compare the possible effects induced by the use of implicit and explicit water models used for the REMD and MD simulations, respectively. This analysis provides information about possible changes occurred in the protein structure due to the use of different mediums around the protein; indeed, the presence of implicit or explicit water may affect the H-bonds formation among residues of the protein and thus the overall structure. We finally identified and compared the most stable H-bonds (occupancy > 50% during the last 50 ns of MD simulation) among the different full-length models (P173a, P173b, P173c) in order to highlight the protein domains mainly involved in intra-molecular contacts. We also analyzed and compared all the salt bridges formed in the different full-length models by using a VMD algorithm (with a 3.2 Å cut-off) in order to assess possible differences in the folding of the three putative structures.

Results

Simulated secondary structures of amelogenin P173

An overview of the amelogenin secondary structures obtained from the REMD simulations at 4, 25 and 80 °C is given in Fig. 2, which shows the dominant primary clusters (i.e. the most likely configurations). Clusters with a similar structure were obtained by grouping the frames generated by the REMD simulations. A relatively highly-populated cluster therefore corresponded to a relatively stable structure (one that is close to the global energy minimum). Certain subsystems showed two or three nearly equivalent clusters, indicating loose folding. The secondary clusters and tertiary clusters, where present, are shown in the Supplementary Information (Fig. S1).

The structure of the amelogenin N-terminus (Met1-Trp45) may be deduced from the REMD results for the first three subsystems. At 4 °C the N-terminus showed strong folding, as reflected by the presence of a single highly-populated cluster (75 %), although the sequences linking it to the central domain showed two equally probable conformations, implying a more loosely folded region. Secondary structure analysis indicated the presence of a disordered β-turn sequence (Met1-His9) followed by distinct α-helices. A first helix (Pro10-Tyr17) was observed at the end of the first subsystem and at the beginning of the second subsystem, where it was split into two sequences (Pro10-Asp14, Ser16-Val19). An additional α-helix (Pro22-Ile30) was also present in the second subsystem, while region Met29-Trp45 showed two putative conformations: a single α-helix (Pro33-Tyr37, primary cluster, Fig. 2) and two $\alpha_{3\text{-}10}$ helices (Tyr34-Ser36 and Met42-Gly44, secondary cluster, Fig. S1). At 25 °C, two α-helices (Pro10-Val19 and Pro22-Ile30) were followed by three putative conformations: unstructured coil/β-turn (principal cluster, Fig. 2), α-helix (Pro33-Tyr37, secondary cluster, Fig. S1), and α-helix/$\alpha_{3\text{-}10}$ helix (Tyr34-Tyr37 and Met42-Gly44, tertiary cluster, Fig. S1). At 80 °C, the secondary structure was similar to that at 25 °C, with the exception of the α-helix Pro41-Ile50 in the secondary cluster (Fig. S1).

The central domains (Leu46-Pro157) were covered by the third and the fourth amelogenin subsystems. At 4 °C, the simulations did not indicate well-defined folding, but a number of structural clusters with limited populations (≈ 20% to 40%) were present. The first part of the central domain (Leu46-Met86), covered by the third subsystem, showed two principal clusters with almost identical probabilities (≈ 44 and ≈ 40 %), implying it to be weakly structured. The first of these clusters (chosen as the principal cluster, Fig. 2) showed a β-turn structure, while the second showed a heterogeneous series of mixed helices (secondary cluster, Fig. S1), i.e. three short α-helices (Leu46-Ile50, Pro52-Gln56, His 68-Met72) and a limited $\alpha_{3\text{-}10}$ helix (Ser61-Ala63). At 25 °C, three clusters were observed, with populations of

17.5%, 16.5% and 16%, indicating further loosening of the folding with increasing *T*. The first cluster was mainly unstructured showing only the $\alpha_{3\text{-}10}$ helix (Ser61-Ala63, Fig. 2). The second was completely unstructured (cf. the first cluster obtained at 4 °C), but the third showed two α-helices (Leu46-Ile50 and His 68-Met72) and a $\alpha_{3\text{-}10}$ helix (Ser61-Ala63), although it lacked the α-helix seen at 4 °C for Pro52-Gln56 (Fig. S1). At 80 °C, the first part of the domain was essentially unstructured, the simulations resulting in two clusters with population densities of only 15% and 13%.

The second part of the central domain (Pro87-Pro157), covered by the fourth amelogenin subsystem, was again found to be loosely folded, producing three sparsely populated clusters ($\approx$ 26%, 24% and 23%) at 4 °C. All of these clusters were characterized by a single α-helix (Pro145-Gln150), while Pro87-Pro144 segment was mostly unstructured, with the exception of a short α-helix (Pro96-His99) in the second cluster and a $\alpha_{3\text{-}10}$ helix (Thr97-His99) in the third cluster (Fig. S1). At 25 °C there was no dominant structure, but the α-helix (Pro145-Gln150) seen at 4 °C was present in most of the sparsely populated clusters. The second part of the central domain was again unstructured at 80 °C, showing four clusters with similar populations, in which β-turns and β-bridges dominated the secondary structure.

The C-terminus (Leu158-Asp173) covered by the fifth amelogenin subsystem showed the largest principal cluster population of 87% at 4 °C (Fig. 2). Secondary structure analysis indicated it to be characterized by a single $\alpha_{3\text{-}10}$ helix (Leu158-Ala160), followed by a longer α-helix (Pro162-Lys168). Similar structures were observed at 25 °C, but at 80 °C the C-terminus was fully unstructured.

Simulated tertiary structure of full-length P173 amelogenin

Full-length P173 amelogenin models were assembled from the structures of the five subsystems obtained with the REMD simulations at 4 °C in order to identify its tertiary structure. The overall stability of the secondary structure was also assessed from classical MD simulations using the full-length models: P173a, P173b and P173c. In each case, the secondary structure tended to be conserved in the N- and C-termini during the MD simulations, while the central domains remained relatively unstructured and did not converge to a unique conformation. Fig. 3 shows a snapshot of model P173a after 100 ns of MD simulation. RMSD calculation from the trajectory of the three simulations was extracted to investigate models stability after switching from implicit solvent (used in REMD simulations) to explicit solvent of MD long-lasting simulations (Fig.4). All the three models showed to reach a stable conformation after 50 ns of MD simulation. In addition, it is seen from the timeline of the secondary structures (Fig. 5) that the second part of the central domain retained its initially unstructured configuration, while the first part of the central domain showed dynamic formation of helices, suggesting an oscillation between unstructured and structured conformations. Such behavior was not observed for model P173c, where helices were initially present in the central domains and persisted over the entire simulation. We conclude that the overall secondary structure is kept similar in the three models, without any significant effect introduced by the construction of the P173 model from the REMD clusters obtained in implicit solvent.

All three full-length amelogenin models showed similar radii of gyration in the range 2.04 to 2.45 nm. The radii of gyration showed a stable trend around its average values (Fig. 6) in the three different amelogenin models. Models P173a and P173c showed very similar and stable radius of gyration values during time (P173a=1.94±0.03 and P173c=1.93±0.02, mean±SD of the last 10 ns of MD simulation). The P173b model, due to a transitory instability (as seen also in the RMSD, Fig. 4), showed the radius decreasing from 2.45±0.11 nm (first 10 ns of the MD simulation) to a stable value of 2.15±0.12 nm, (last 10 ns of the MD simulation) thus falling in the observed experimental range. Comparable values (2.2 nm) have been obtained by dynamic light scattering, and they are also consistent with observations by transmission and scanning electron microscopy, and atomic force microscopy (Humphrey, Dalke, & Schulten, 1996). The solvent accessible surface (SAS) was in each case 126 ± 2 nm^2. Moreover, the SAS computed for each domain (normalized with respect to the number of residues), indicated the N- and C-termini to be more exposed to the solvent than the central domains (Table 1). This is attributed to stronger hydrophobic interactions in these latter, as suggested in previous work, the central domains containing a significantly higher proportion of hydrophobic residues than the termini (Du, Falini, & Fermani, 2005). We note that although the used force field is better suited for globular proteins (with limited validation for IDPs), it showed to well represent P173 behavior observed experimentally in literature.

In our work the use of implicit solvent REMD simulation was mandatory for treating a protein of 173 aminoacids due computational costs. Hence, Our full-length models were extensively simulated in explicit solvent (100 ns) to overcome possible limitations due to implicit solvent approximation. The effect of switching from implicit to explicit solvent can be monitored comparing the number of H-bonds at the beginning of the MD simulation (i.e., the structure coming from implicit solvent REMD simulations) and during the last 50 ns of explicit solvent MD. We observe a limited variation in the number of H-bonds (initial P173a: 18, final P173a 16±5; initial P173b: 16, final P173b 18±5; initial P173c: 10, final P173c 16±5) showing that the use of explicit solvent induces some rearrangement of H-bonds, although limited. This outcome further confirms the necessity of our long-lasting explicit solvent MD simulations to draw robust conclusions on large IDPs such as amelogenin models. The analysis of the most stable H-bonds (occurring for at least 50% of the time during the last 50 ns) showed that there are some very stable H-bonds occurring mainly within the central domain of the protein (residues from 45 to 158) that are common to all models (in particular ALA108-ASN103 and SER129-GLN123) (Table 2). Similarly, salt bridges were analyzed during the last 50ns of the MD simulation, comparing the amino acids involved in bridge formation among the different models P173a, P173b, P173c (Table 3). The calculated salt bridges are mainly located in the C- and N-terminals of the protein. Salt bridges ASP165-LYS166 and GLU170-LYS166 are common to all the three models. Salt bridges involving residues from 165 to 171 are involved in maintaining the secondary structure of the C-term. ARG22 of the N-term forms salt bridges with residues ASP 165, GLU159, GLU170 and GLU171 of the C-term contributing in maintaining the folding of the protein. The same trend is observed for ASP26, which interact with LYS166 and LYS168 of the N-term. Based on these observations we can conclude that the intramolecular interactions involved in the amelogenin folding are characterized by the presence of some H-bonds in the central region and by ionic interactions occurring between charged residues mainly located in the protein terminal domains.

Discussion

Comparison with previous results from the literature

Zhang *et al.* (Zhang et al., 2011) have proposed a putative secondary structure for the full-length murine amelogenin M180 at 25 °C and pH 5.5 based on NMR spectroscopy. While this is a different type of amelogenin, the high inter-species similarity (about 87 %) justifies comparison with the results of the present REMD simulations (Fig. S2). According to Zhang *et al.* (Zhang et al., 2011)*,* two α-helices (Ser9-Val19, Lys24-IIe30) are present in the N-terminal region of M180, separated by a disordered pair of amino acids. This secondary structure shows high similarity (about 81 %) with that predicted here for P173. However, modifications to the primary structure may result in significant changes in the energy landscape and hence in the local conformations. It follows that differences in secondary structure were observed where a basic polar amino acid was substituted by a polar amino acid in the primary structure (Ser9/M180 to His9/P173 and Gln46/M180 to His32/P173) or where steric hindrance due to an aromatic ring (Leu15/M180 to Phe15/P173) resulted in the disruption of the helical structure (Petukhov, Uegaki, Yumoto, & Serrano, 2002). The present model predicted two separate helices to be present in the region Pro10-Val19 (Pro10-Asp14, Ser16-Val19), possibly owing to Ser9-to-His9 and Leu15-to-Phe15 substitutions with respect to M180. A further helix was also predicted for P173 (Pro22-Ile30), followed by a single helix (Pro33-Tyr37, principal cluster) or two $\alpha_{3\text{-}10}$ helices (Tyr34-Ser36 and Met42-Gly44, secondary cluster). Pro35-to-Thr35 substitution may stabilize the α-helix in the region Pro33-Tyr37, given that tryptophan is more favorable to α-helix formation than proline (Frishman & Argos, 1995).

According to the NMR results, the central domain of M180 (Leu46-Ala160) is dominated by a coiled region interrupted by short α-helices (Val53-Gln56 α-helix and Pro74-Gln76 α_{3-10} helix and Ser107-Gln109 α_{3-10} helix), with polyproline type II (PPII) helical structure in the second part of the central domain. However, although the methods used in the present work were not suited to the identification of PPII structure, the REMD simulations indicated sparsely-populated clusters containing short helices and unstructured regions to be present in the corresponding domains in P173, and the MD simulations based on the full-length P173 amelogenin models were characterized by transient helix formation. The apparent flexibility and lack of stability in the central domains (i.e. the implied co-existence of unstable conformations, oscillating between flexible β-turn structures and α_{3-10} helices), as compared with the N- and C-termini, reflect their intrinsic disorder, a characteristic that is argued to be essential for the proper self-assembly of amelogenin (B. Aichmayer et al., 2005; Margolis, Beniash, & Fowler, 2006). The presence of unstable secondary structures has been also confirmed by previous observations based on Fourier transform infrared spectroscopy (FTIR), circular dichroism and NMR (Goto, Kogure, Takagi, Aimoto, & Aoba, 1993; Matsushima, Izumi, & Aoba, 1998; Noguchi, Hayakawa, & Ebata, 1957; Renugopalakrishnan, 2002).

In the case of the C-terminus, the well-defined secondary structure (α_{3-10} helix Leu158-Ala160 and α-helix Pro162-Lys168) implied by the REMD model for P173 was apparently inconsistent with the NMR results for M180, which suggest this domain to be relatively unstructured (Zhang et al., 2011)(Fig. S2). As indicated by both the present SAS analysis and literature data (Renugopalakrishnan, 2002) the C-terminus is the domain most exposed to the solvent, reflecting its relatively high degree of hydrophilicity (Table 1). Under these conditions the formation of α-helices is expected to be favored over other types of secondary structure. The inconsistencies between the NMR and FT-IR results may therefore be due to the difficulties inherent in the use of such techniques to characterize the structure of floating domains, such as the termini.

A further study of recombinant P172 by Beniash *et al.* (Beniash et al., 2012) has also combined NMR and CD with numerical methods. In this work the authors considered conditions under which pure amelogenin is not expected to show self-assembly (10 °C, pH 4), and provided less information on the secondary structure than Zhang *et al.* (Zhang et al., 2011),so that there was insufficient data for detailed comparison with the results of the present model. They associated PPII-type conformations with the Ile70-Pro89 and Pro102-

Pro145 regions, i.e. 22 % and 37 % of the total secondary structure respectively. Comparison with the REMD simulations showed correspondence to be limited to the presence of a single α-helix (Val19-Pro33 in rP172 and Pro22-Ile30 in P173 at 4, 25 and 80 °C).

The effect of thermal stress

The results of the simulations described in the previous sections may be summarized as follows. The hydrophilic terminal regions (Table 1) showed a globular conformation (Fig. 3) and were most exposed to the solvent. Analysis of the secondary structure confirmed the presence of hydrophilic alpha helices (Pro10-Asn14, Ser16-Val19, Pro22-Ile30, Pro33-Tyr37, Pro162-Lys168) and a weaker $\alpha_{3\text{-}10}$ helix (Leu158-Ala160) at 4 °C. The inner domains, which were least exposed to the solvent, did not show well-defined secondary structure, only a single helix being generated close to the C terminus (Met146-Gin150). Thermal stress, induced by increasing the temperature from 4 to 25 and 80 °C led to a progressive loss of secondary structure in the C-terminus, which became fully disordered at 80 °C, and an increase in globularity of the N-terminus. At the same time, β-bridges increased in prominence at 25 °C and 80 °C, especially in the central domains, where they induced a putative fold, exposing an $\alpha_{3\text{-}10}$ helix to the solvent at 25 °C to 80 °C, and a long helix at 80 °C (Trp45-Gln57) that was not present at 4 °C.

It has been shown experimentally that the C-terminus is often associated with self-assembly in amelogenin (for example, the formation of nanospheres) (Janet Moradian-Oldak, Bouropoulos, Wang, & Gharakhanian, 2002), and its interactions with calcium and phosphate ions (Khan, Li, & Habelitz, 2012). Similarly, β-bridges are typically associated with interactions between proteins. The solvent accessible surface analysis (Table 1) confirmed both the N- and C-termini to be substantially exposed to the solvent. The presence of a thermally labile secondary structure in the C-terminus and the increased solvent exposure of β-bridges in the central domains with increasing temperature, therefore supports the hypothesis that thermal stress may provide a trigger for changes in conformation that lead to significant modifications in how amelogenin molecules interact with each other, and with their environment.

Conclusions

To our knowledge, this is the first time that the conformational behavior of full-length native porcine amelogenin P173 has been predicted from its primary sequence. The initial simulation results for the secondary and tertiary structures at pH 5 and 4 °C were in good agreement with literature data for native murine amelogenin obtained under similar conditions (Zhang et al., 2011) and the full-length model was shown to be suitable for the investigation of conformational changes induced by thermal stress. In future work, it is intended to use the present techniques to simulate systems comprising two or more amelogenin proteins and/or other molecules present in the EMD pool, with the aim of modeling amelogenin self-assembly mechanisms, identifying specific protein domains associated with the aggregation process, and investigating the influence of environmental conditions.

Acknowledgements

The authors wish to thank the Competence Center for Materials Science and Technology (CCMX, www.ccmx.ch), the Institut Straumann AG and Fondazione Cariplo (grant n° 2013-0766) for financing this work.

References

Aichmayer, B., Margolis, H. C., Sigel, R., Yamakoshi, Y., Simmer, J. P., & Fratzl, P. (2005). The onset of amelogenin nanosphere aggregation studied by small-angle X-ray scattering and dynamic light scattering. *Journal of Structural Biology*, *151*(3), 239–249. doi:10.1016/j.jsb.2005.06.007

Aichmayer, B., Wiedemann-Bidlack, F. B., Gilow, C., Simmer, J. P., Yamakoshi, Y., Emmerling, F., … Fratzl, P. (2010). Amelogenin nanoparticles in suspension: Deviations from spherical shape and pH-dependent aggregation. *Biomacromolecules*, *11*(2), 369–376. doi:10.1021/bm900983b

Banerjee, S., Chakraborty, S., & De, R. K. (n.d.). Deciphering the cause of evolutionary variance within intrinsically disordered regions in human proteins. *Journal of Biomolecular Structure and Dynamics*, *0*(0), 1–17. doi:10.1080/07391102.2016.1143877

Banerjee, S., & De, R. K. (2015). Structural Disorder: A tool for housekeeping proteins performing tissue-specific interactions. *Journal of Biomolecular Structure & Dynamics*, *1102*(April), 1–68. doi:10.1080/07391102.2015.1095115

Beniash, E., Simmer, J. P., & Margolis, H. C. (2012). Structural changes in amelogenin upon self-assembly and mineral interactions. *Journal of Dental Research*, *91*(10), 967–72. doi:10.1177/0022034512457371

Cino, E. A., Choy, W. Y., & Karttunen, M. (2012). Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations. *Journal of Chemical Theory and Computation*, *8*(8), 2725–2740. doi:10.1021/ct300323g

Colson, B. A., Thompson, A. R., Espinoza-Fonseca, L. M., & Thomas, D. D. (2016). Site-directed spectroscopy of cardiac myosin-binding protein C reveals effects of phosphorylation on protein structural dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(12), 3233–3238. doi:10.1073/pnas.1521281113

Das, R. K., Crick, S. L., & Pappu, R. V. (2012). N-terminal segments modulate the ??-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *Journal of Molecular Biology*, *416*(2), 287–299. doi:10.1016/j.jmb.2011.12.043

Das, R. K., & Pappu, R. V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(33), 13392–13397. doi:10.1073/pnas.1304749110

Delak, K., Harcup, C., Lakshminarayanan, R., Sun, Z., Fan, Y., Moradian-Oldak, J., & Evans, J. S. (2009). The tooth enamel protein, porcine amelogenin, is an intrinsically disordered protein with an extended molecular configuration in the monomeric form. *Biochemistry*, *48*(10), 2272–2281. doi:10.1021/bi802175a

Du, C., Falini, G., & Fermani, S. (2005). Supramolecular assembly of amelogenin nanospheres into birefringent microribbons. *Science (New York, N.Y.)*, *307*(5714), 1450–4. doi:10.1126/science.1105675

Espinoza-Fonseca, L. M. (2009). Leucine-rich hydrophobic clusters promote folding of the N-terminus of the intrinsically disordered transactivation domain of p53. *FEBS Letters*, *583*(3), 556–560. doi:10.1016/j.febslet.2008.12.060

Espinoza-Fonseca, L. M., & Kelekar, A. (2015). High-resolution structural characterization of Noxa, an intrinsically disordered protein, by microsecond molecular dynamics simulations. *Molecular bioSystems*, *11*(7), 1850–6. doi:10.1039/c5mb00170f

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function and Genetics*, *23*(4), 566–579. doi:10.1002/prot.340230412

Goto, Y., Kogure, E., Takagi, T., Aimoto, S., & Aoba, T. (1993). Molecular conformation of porcine amelogenin in solution: three folding units at the N-terminal, central, and C-terminal regions. *Journal of Biochemistry*, *113*(1), 55–60. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8454575

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. doi:10.1016/0263-7855(96)00018-5

Khan, F., Li, W., & Habelitz, S. (2012). Biophysical characterization of synthetic amelogenin C-terminal peptides. *European Journal of Oral Sciences*, *120*(2), 113–122. doi:10.1111/j.1600-0722.2012.00941.x

Kiefhaber, T., Rudolph, R., Kohler, H. H., & Buchner, J. (1991). Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. *Bio/technology (Nature Publishing Company)*, *9*(9), 825–829. doi:10.1038/nbt0991-825

Leonova, E. I., & Galzitskaya, O. V. (2015). Cell communication using intrinsically disordered proteins: what can syndecans say? *Journal of Biomolecular Structure & Dynamics*, *33*(5), 1037–1050. doi:10.1080/07391102.2014.926256

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., & Shaw, D. E. (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, *134*(8), 3787–3791. doi:10.1021/ja209931w

Lindskog, S. (1982). Formation of intermediate cementum. II: a scanning electron microscopic study of the epithelial root sheath of Hertwig in monkey. *Journal of Craniofacial Genetics and Developmental Biology*, *2*(2), 161—169. Retrieved from http://europepmc.org/abstract/MED/6184381

Lumry, R., & Eyring, H. (1954). Conformation Changes of Proteins. *The Journal of Physical Chemistry*, *58*(2), 110–120. doi:10.1021/j150512a005

Mackerell, A. D. (2004). Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*. doi:10.1002/jcc.20082

Margolis, H. C., Beniash, E., & Fowler, C. E. (2006). Role of macromolecular assembly of enamel matrix proteins in enamel formation. *Journal of Dental Research*, *85*(9), 775–793. doi:10.1177/154405910608500902

Matsushima, N., Izumi, Y., & Aoba, T. (1998). Small-angle X-ray scattering and computer-aided molecular modeling studies of 20 kDa fragment of porcine amelogenin: does amelogenin adopt an elongated bundle structure? *Journal of Biochemistry*, *123*(1), 150–6. doi:10.1093/oxfordjournals.jbchem.a021902

Moradian-Oldak, J., Bouropoulos, N., Wang, L., & Gharakhanian, N. (2002). Analysis of self-assembly and apatite binding properties of amelogenin proteins lacking the hydrophilic C-terminal. *Matrix Biology*, *21*(2), 197–205. doi:10.1016/S0945-053X(01)00190-1

Moradian-Oldak, J., Jimenez, I., Maltby, D., & Fincham, A. G. (2001). Controlled proteolysis of amelogenins reveals exposure of both carboxy- and amino-terminal regions. *Biopolymers*, *58*(7), 606–616. doi:10.1002/1097-0282(200106)58:7<606::AID-BIP1034>3.0.CO;2-8

Moradian-Oldak, J., Leung, W., & Fincham, a G. (1998). Temperature and pH-dependent supramolecular self-assembly of amelogenin molecules: a dynamic light-scattering analysis. *Journal of Structural Biology*, *122*(3), 320–7. doi:10.1006/jsbi.1998.4008

Noguchi, J., Hayakawa, T., & Ebata, M. (1957). Reversible heat coagulation of some watersoluble amino acid copolymers. *Journal of Polymer Science*, *23*(104), 843–849. doi:10.1002/pol.1957.1202310428

Paine, M. L., & Snead, M. L. (1997). Protein interactions during assembly of the enamel organic extracellular matrix. *Journal of Bone and Mineral Research : The Official Journal of the American Society for Bone and Mineral Research*, *12*(2), 221–7. doi:10.1359/jbmr.1997.12.2.221

Patriksson, A., & van der Spoel, D. (2008). A temperature predictor for parallel tempering simulations. *Physical Chemistry Chemical Physics*, *10*(15), 2073. doi:10.1039/b716554d

Petukhov, M., Uegaki, K., Yumoto, N., & Serrano, L. (2002). Amino acid intrinsic alpha-helical propensities III: positional dependence at several positions of C terminus. *Protein Science : A Publication of the Protein Society*, *11*(4), 766–777. doi:10.1110/ps.2610102

Renugopalakrishnan, V. (2002). A 27-mer tandem repeat polypeptide in bovine amelogenin: Synthesis and CD spectra. *Journal of Peptide Science*, *8*(4), 139–143. doi:10.1002/psc.378

Slavkin, H. C., Bessem, C., Fincham, A. G., Bringas, P., Santos, V., Snead, M. L., & Zeichner-David, M. (1989). Human and mouse cementum proteins immunologically related to enamel proteins. *BBA - General Subjects*, *991*(1), 12–18. doi:10.1016/0304-4165(89)90021-4

Sugita, Y., & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, *314*(1–2), 141–151. doi:10.1016/S0009-2614(99)01123-9

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*. doi:10.1002/jcc.20291

Wen, H. B., Moradian-Oldak, J., Leung, W., Jr., P. B., & Fincham, A. G. (1999). Microstructures of an Amelogenin Gel Matrix. *Journal of Structural Biology*, *126*(1), 42–51. doi:http://dx.doi.org/10.1006/jsbi.1999.4086

Wiedemann-Bidlack, F. B., Beniash, E., Yamakoshi, Y., Simmer, J. P., & Margolis, H. C. (2007). pH triggered self-assembly of native and recombinant amelogenins under physiological pH and temperature in vitro. *Journal of Structural Biology*, *160*(1), 57–69. doi:10.1016/j.jsb.2007.06.007

Yadav, L. R., Rai, S., Hosur, M. V, & Varma, A. K. (2015). Functional assessment of intrinsic disorder central domains of BRCA1. *Journal of Biomolecular Structure and Dynamics*, *33*(11), 2469–2478. doi:10.1080/07391102.2014.1000973

Zhang, X., Ramirez, B. E., Liao, X., & Diekwisch, T. G. H. (2011). Amelogenin supramolecular assembly in nanospheres defined by a complex Helix-Coil-PPII helix 3D-Structure. *PLoS ONE*, *6*(10). doi:10.1371/journal.pone.0024952

Zhou, R. (2007). Replica exchange molecular dynamics method for protein folding simulation. *Methods in Molecular Biology (Clifton, N.J.)*, *350*(November), 205–223. doi:10.1016/S0009-2614(99)01123-9

Domain	**SAS/residue [nm^2]**	**Total residues**	**Hydrophobic residues**	**Hydrophylic residues**
N-terminus	0.89	45	20 (45%)	25 (55%)
Histidine rich	0.64	42	23 (55%)	19 (45%)
Polyproline repeat	0.65	63	36 (58%)	27 (42%)
C-terminus	1.05	25	11 (48%)	13 (52%)

Table 1: Predicted solvent accessible surface for the different P173 amelogenin domains, number of total residues, and the number of hydrophobic and hydrophilic residues (and their percentage with respect to the total number).

P173a		**P173b**		**P173c**	
H-bond Residues	*occupancy*	H-bond Residues	*occupancy*	H-bond Residues	*occupancy*
		ARG22-ASP26	*58.20%*		
		GLN83-ILE80	*50.60%*		
				ASN103-THR97	*53.64%*
ALA108-ASN103	*59.00%*	ALA108-ASN103	*55.60%*	ALA108-ASN103	*54.97%*
GLN115-GLN49	*57.60%*				
SER129-GLN123	*68.40%*	SER129-GLN123	*64.40%*	SER129-GLN123	*50.33%*
SER151-PRO89	*51.60%*				
				LEU137-PRO130	*50.99%*

Table 2: H-bonds formed between two amelogenin residues in the three models (P173a, b, c) as calculated in the last 50 ns of each MD simulation. H-bonds occurring for more than 50% of the simulated time are reported. H-bonds common to all the amelogenin models are highlighted in green, while H-bonds occurring only in one model are in red. Four H-bonds occurred for in all the models. H-bonds between ALA108-ASN103 and SER129-GLN123 are common to all the three models. The majority of the H-bonds occur within the central domain of the protein (residues from 45 to 158).

P173a	P173b	P173c
ASP26-ARG22	ASP26-ARG22	
	GLU40-HIS122	
ASP26-LYS166		
		ASP26-LYS168
	ASP155-HIS122	
	GLU159-LYS168	GLU159-LYS168
GLU159-ARG22	GLU159-ARG22	
GLU159-LYS168		
	GLU159-LYS24	
	ASP165-ARG22	ASP165-ARG22
ASP165-LYS166	ASP165-LYS166	ASP165-LYS166
ASP165-ARG22		
ASP165-LYS66		
GLU170-LYS166	GLU170-LYS166	GLU170-LYS166
GLU170-ARG22		GLU170-ARG22
GLU171-ARG22	GLU171-ARG22	
GLU171-LYS166		
		GLU171-LYS168

Table 3: Salt bridges formed between two amelogenin residues in the three models (P173a, b, c) occurring during the entire trajectory of each MD simulation. Salt bridges common to all the three amelogenin models are highlighted in green, salt bridges found in two models are highlighted in blue, while salt bridges occurring only in one model are in red. 11 salt bridges are found for model P173a, 10 for P173b and 8 for P173c.

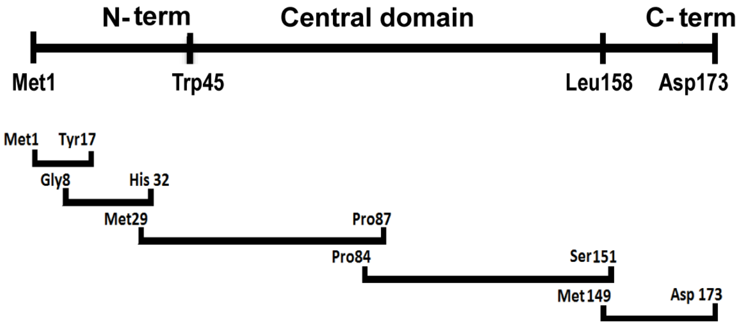

FIGURE 1: Schematic representation of the Amelogenin P173 N-terminus (Met1-Trp45), central domain (Leu46-Pro157) and C-terminus (Leu158-Asp173), along with the sequences of the five subsystems used in the REMD simulations: Met1-Tyr17 and Gly8-His32 (mainly encompassing the N-terminus), Met29-Pro87 and Pro84-Ser151 (mainly encompassing the central domain), Met149-Asp173 (encompassing the C-terminus).

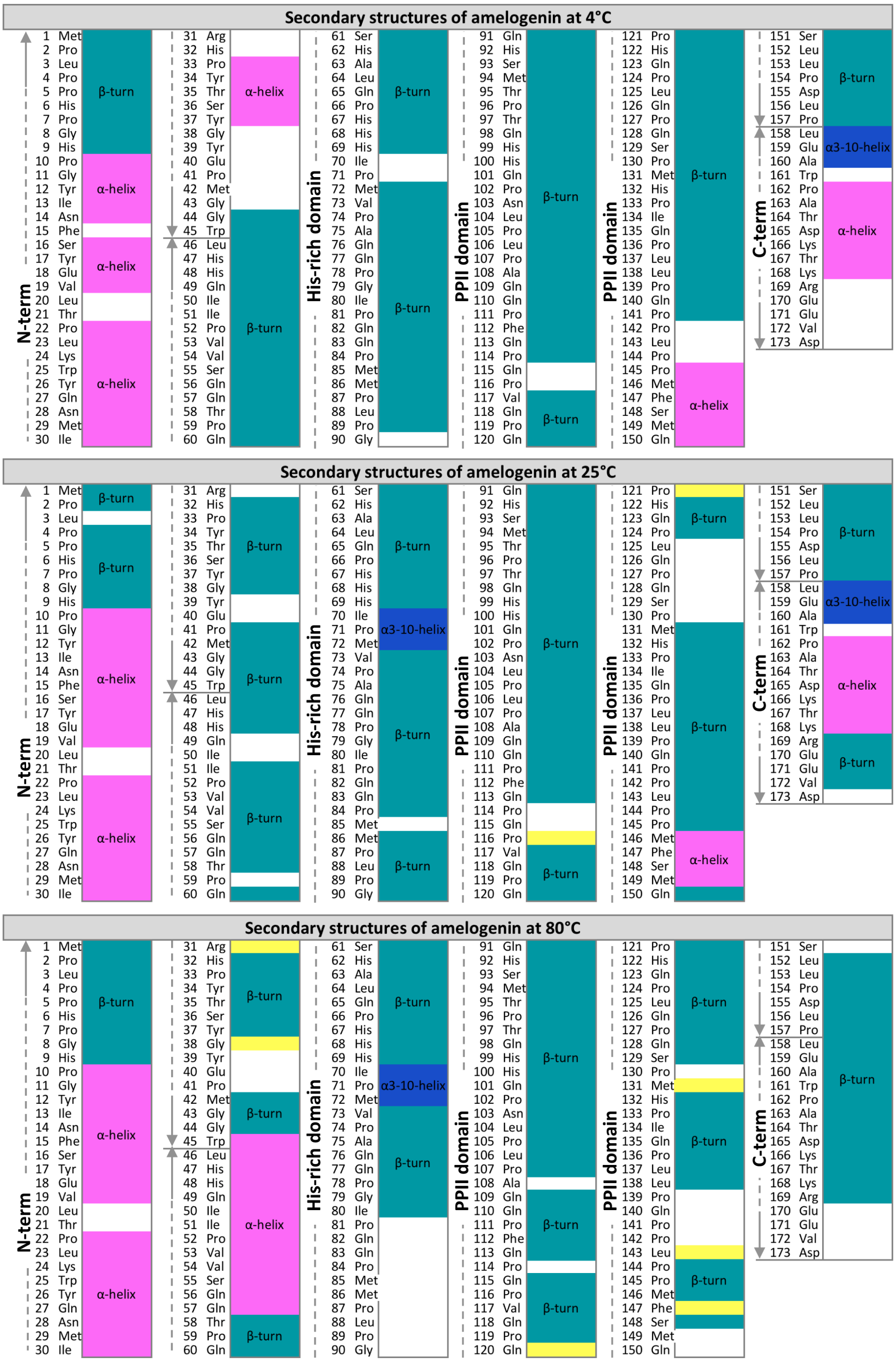

FIGURE 2: Amelogenin P173 secondary structures predicted by replica exchange MD simulations at 4, 25 and 80 °C. The structures shown correspond to the most populated clusters at each *T* (β-turns in green, α-helices in magenta, $\alpha_{3\text{-}10}$ helices in blue, β-bridges in yellow, and coils in white). Secondary structures from the secondary and tertiary clusters of the REMD simulations are given in the Supplementary Material.

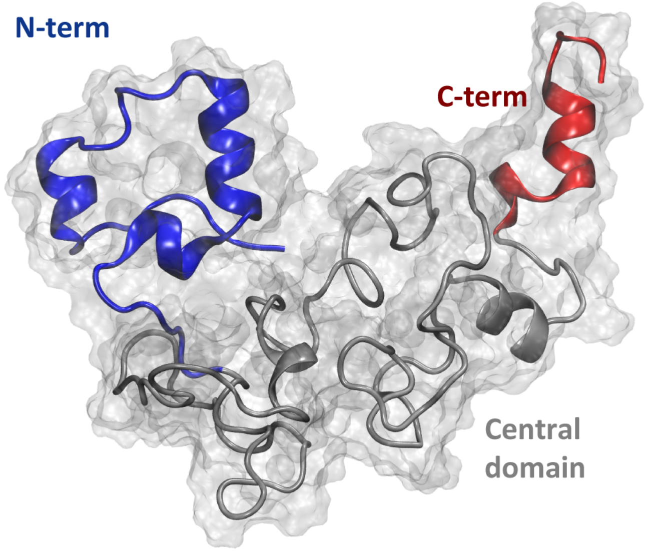

FIGURE 3: Full-length model of Amelogenin P173 at 4 °C after 100 ns of MD simulation. The full-length model was obtained by merging the secondary structures predicted from the primary REMD clusters. The protein domains are highlighted as follows: N-terminus in blue, central domain in dark grey, and C-terminus in red.

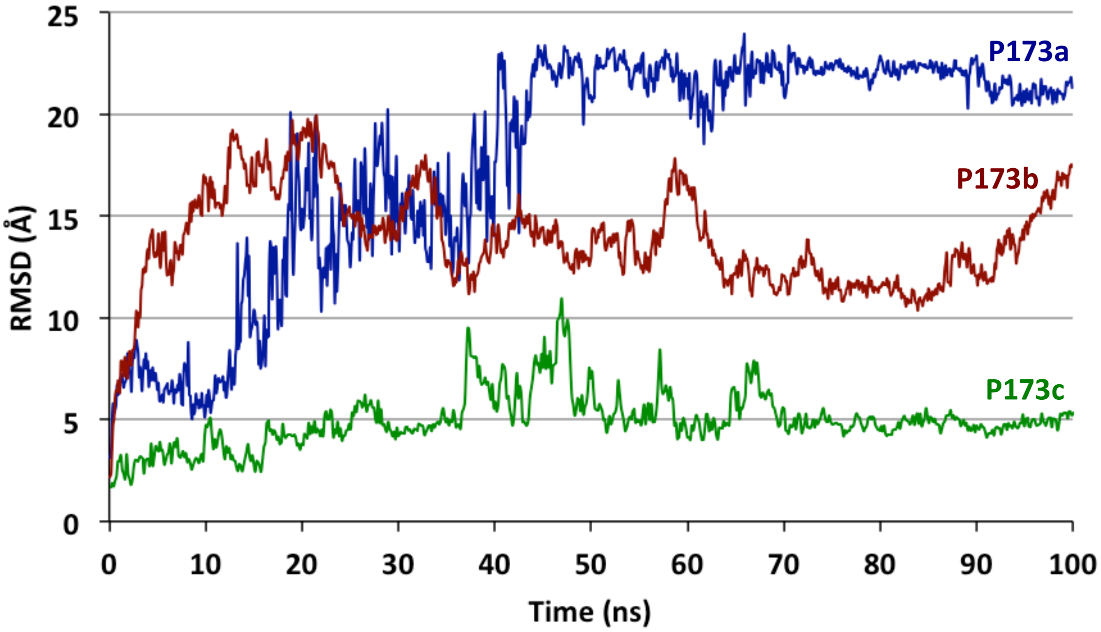

FIGURE 4: Root Mean Square Deviation evolution during the MD simulation in explicit solvent of the model P173a (blue) P173b (red) and P173c (green). The three model achive a stable RMSD value after about 50 ns. Model P173b shows some oscillations around a stable value models (P173a, b, c).

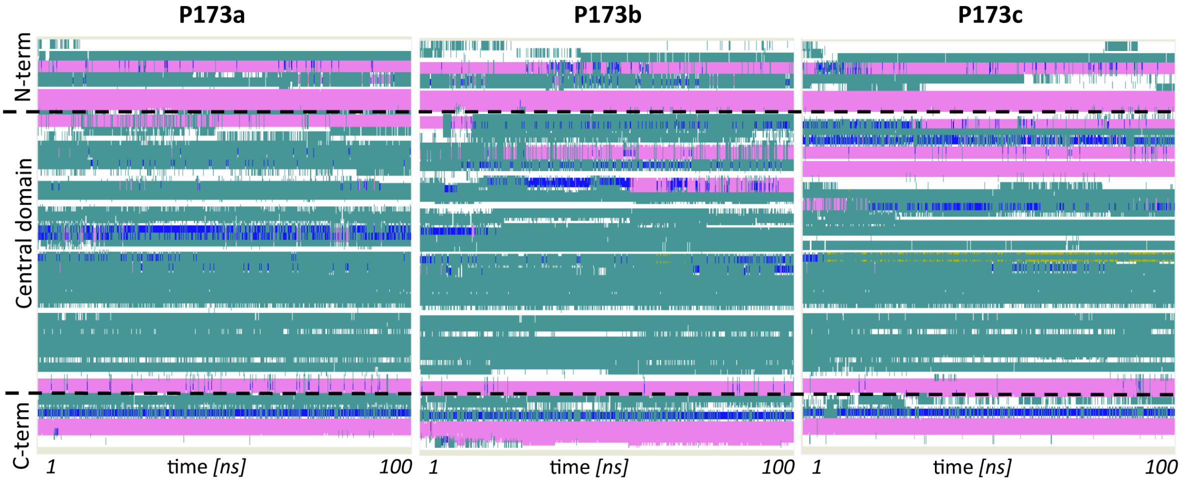

FIGURE 5: Timeline of the secondary structures in the full-length amelogenin models obtained from the trajectories of the 100 ns MD simulations: P173a, reference model constructed from the secondary structures predicted from the primary clusters; P173b constructed using a different assembly procedure to that used for the reference model; P173c constructed using the secondary REMD cluster for the first part of the central domain (β-turns in green, α-helices in magenta, $\alpha_{3\text{-}10}$ helices in blue, β-bridges in yellow, and coils in white).

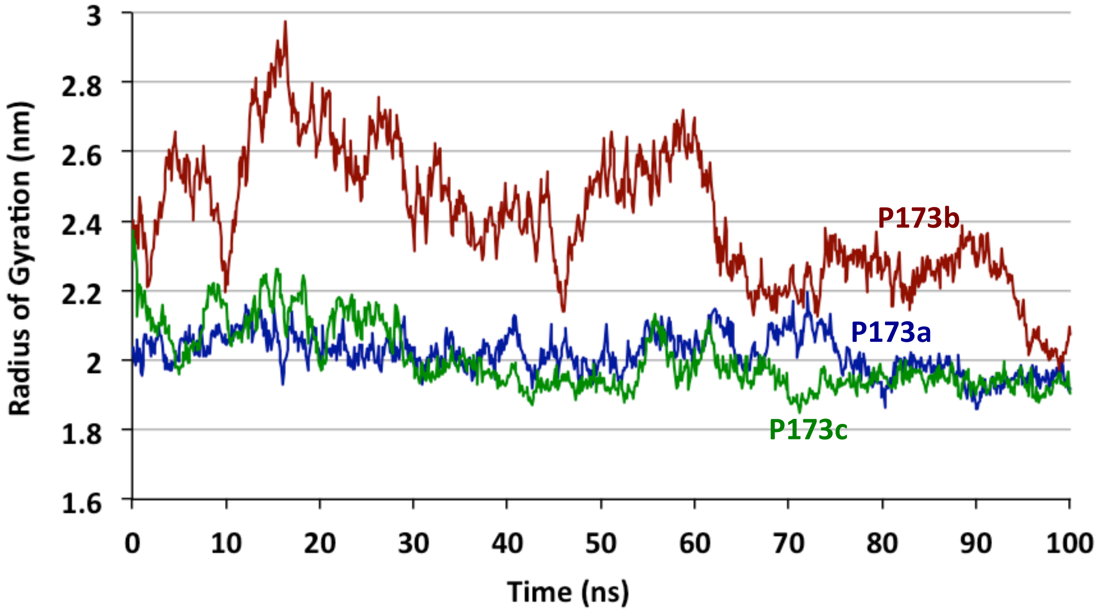

FIGURE 6: Radius of gyration calculated during the MD simulation for the three putative models P173a (blue), P173b (red) and P173c (green). Models P173a and P173c show very similar values, while model P173b shows a trajectory with higher values and some fluctuation until 70 ns.