



REVIEW

# A review of protein structure and gene organisation for proteins associated with mineralised tissue and calcium phosphate stabilisation encoded on human chromosome 4

N. Laila Huq, Keith J. Cross, Men Ung, Eric C. Reynolds\*

Cooperative Research Centre for Oral Health Science, School of Dental Science,  
The University of Melbourne, 711 Elizabeth Street, Melbourne, Vic. 3010, Australia

Accepted 23 December 2004

## KEYWORDS

Multiple phosphoseryl-  
containing proteins;  
Conserved  
chromosomal  
synteny;  
Codon usage bias;  
Gene duplication

**Summary** Several proteins associated with mineralised tissue (teeth and bone) or involved in calcium phosphate stabilisation in the body fluids, milk and saliva have been mapped to the q arm of human chromosome 4. These include the dentine/bone proteins dentine sialophosphoprotein (DSPP), dentine matrix protein 1 (DMP1), bone sialoprotein (BSP), matrix extracellular phosphoglycoprotein, osteopontin (OPN), enamel, ameloblastin, milk caseins, salivary statherin, and proline-rich proteins. The proposed function of those that are multiphosphorylated is: (i) the stabilisation of calcium phosphate in solution (e.g. casein, statherin) preventing spontaneous precipitation and seeded-crystal growth or (ii) promoting biomineralisation (e.g. the phosphophoryn domain of DSPP), where the protein described as a template macromolecule, is proposed to act as a nucleator/promoter of crystal growth. The genes of these proteins have been subjected to conserved chromosomal synteny during mammalian evolution. The multiphosphorylated proteins statherin, caseins, phosphophoryn, BSP and OPN have been characterised as intrinsically disordered. The codon usage patterns for the amino acid serine reveal a bias for AGC and AGT codons within the human genes *dspp*, *dmp1* and *bsp*, mouse *dspp* and *dmp1* but not significantly for statherin or caseins. This pattern was also observed in the gene encoding hen phosvitin that also contains stretches of multiphosphorylated serines and in the *dmp1* gene sequences of mammalian, reptilian and avian classes. In conclusion, these intrinsically disordered multiphosphorylated proteins are the translation products of genes displaying examples of codon usage bias, internal repeats and conserved chromosomal synteny within the mammalian class.

© 2005 Elsevier Ltd. All rights reserved.

\* Corresponding author. Tel.: +61 3 9341 0270; fax: +61 3 9341 0236.  
E-mail address: [e.reynolds@unimelb.edu.au](mailto:e.reynolds@unimelb.edu.au) (E.C. Reynolds).

## Contents

Introduction . . . . .	600
Dentine and bone extracellular proteins (the SIBLING family). . . . .	600
Enamel matrix proteins. . . . .	600
Calcium-sensitive caseins. . . . .	603
Salivary proteins . . . . .	603
Conserved chromosomal synteny of DSPP, OPN, BSP, CSN, AMBN, and STATH genes . . . . .	604
Inversions and translocations of genes . . . . .	604
Prevalence of intrinsic disorder of proteins . . . . .	604
Multifunctionality of proteins . . . . .	604
Amino acid composition . . . . .	605
High incidence of repeats . . . . .	605
Codon bias . . . . .	606
Conclusion . . . . .	607
Acknowledgements . . . . .	607
References . . . . .	607

## Introduction

Several proteins involved in the stabilisation of calcium phosphate in body fluids and/or associated with mineralised tissue (teeth and bone) are encoded on chromosome 4 in the human genome (Table 1 and Fig. 1). These include the phosphophoryn domain and dentine sialoprotein domains of the dentine sialophosphoprotein (DSPP),<sup>1</sup> dentine matrix protein 1 (DMP1),<sup>2</sup> bone sialoprotein (BSP),<sup>3</sup> matrix extracellular phosphoglycoprotein (MEPE)<sup>4</sup> and osteopontin (OPN)<sup>5</sup> generally found in hard tissues; enamelin (ENAM)<sup>6</sup> and ameloblastin (AMBN)<sup>7</sup> found in enamel; statherin (STATH),<sup>8</sup> histatin (HTN)<sup>9</sup> and proline-rich proteins (PROL)<sup>10</sup> found in saliva<sup>11</sup>; and the caseins (CSN)<sup>12</sup> found in milk. Most of the proteins are post-translationally modified being phosphorylated and/or glycosylated. Many of those phosphorylated proteins contain multiple phosphoserine residues in clusters. The proposed functions of the multiphosphorylated proteins are: (i) the stabilisation of calcium phosphate in solution (e.g. casein, statherin) preventing spontaneous precipitation and seeded-crystal growth and/or (ii) biomineralisation (e.g. phosphophoryn), where the protein, described as a template macromolecule, is proposed to act as a nucleator/promoter of crystal growth. The tissue distribution, post-translational modifications and proposed functions of these proteins are summarised in Table 1. Some of the proteins are proposed to have multifunctions and these are discussed in more detail later.

## Dentine and bone extracellular proteins (the SIBLING family)

Within a 375 kb region on human chromosome 4q21 (Fig. 1), there is a cluster of genes coding for the proteins DSPP, DMP1, BSP, MEPE and OPN that have similar exon structures. Exon 1 is non-coding. Exon 2 encodes for the leader sequence plus the first two residues of the mature protein. Exons 3 and 5 often contain the consensus sequence for casein kinase II phosphorylation (SSEE). Exon 4 is usually proline-rich. The last one or two exons encode the vast majority of the protein and usually contain the integrin-binding tri-peptide Arg–Gly–Asp (RGD). Since BSP and OPN (both ~300 aa) and DMP1 (513–657 aa) are small molecules, these proteins are known as the Small Integrin-Binding Ligand N-linked Glycoprotein (SIBLING) family.<sup>13,14</sup> This terminology has limitations as “OPN typically may not be N-glycosylated”<sup>15</sup> and the sizes of the DSPP protein can vary with the species as shown in Fig. 2. It has been suggested that the genes coding for the proteins DSPP, DMP1, OPN, BSP and MEPE have arisen from a common ancestor.<sup>7,14</sup>

## Enamel matrix proteins

Two enamel matrix proteins ENAM and AMBN are coded by genes located on human chromosome 4q13.3. The human AMBN gene has 13 exons coding for a small protein of 447 residues. The human ENAM gene has eight exons coding for a protein that is rich in prolines. The third protein believed to be involved

**Table 1** Proteins associated with mineralised tissue or calcium phosphate stabilisation in body fluids.

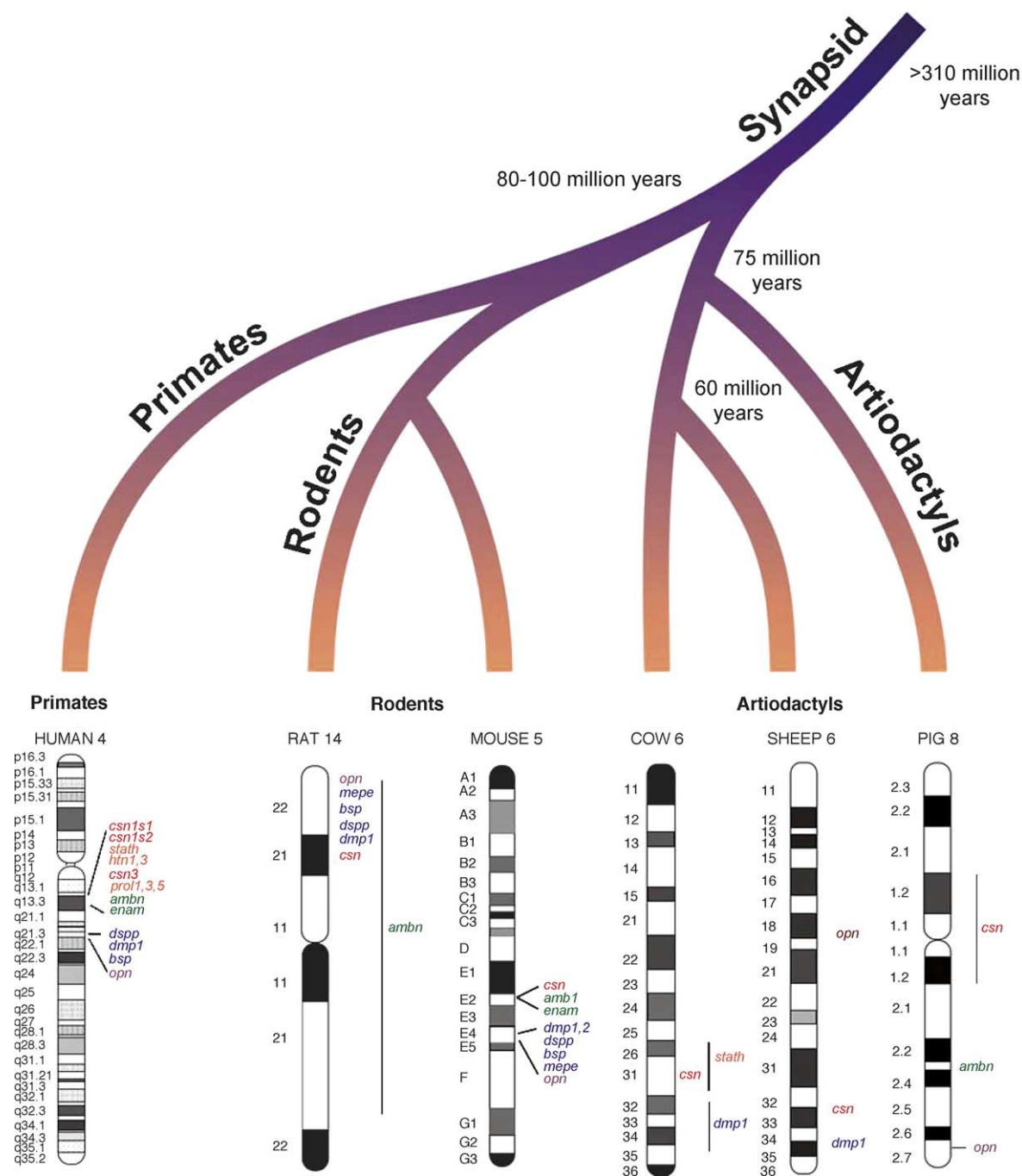
Group	Protein	Abbreviation	Gene	Proposed function	Predominant tissue distribution	Post-translational modification	Chromosomal location	References
SIBLING proteins	Dentine phosphophoryn/dentine phosphoprotein	DPP	<i>dspp</i>	Te	Dentine/bone	Ph <sup>a</sup>	4q21.3	62
	Dentine sialoprotein	DSP	<i>dspp</i>	Un	Dentine/bone	Ph, Gl	4q21.3	62
	Dentine matrix protein 1	DMP1	<i>dmp1</i>	Mu	Dentine/bone	Ph <sup>b</sup>	4q21.3	3,36,37,63
	Bone sialoprotein	BSP	<i>bsp</i>	Mu	Bone, dentin	Ph <sup>c</sup> , Gl	4q21.3	3,37
	Osteopontin	OPN	<i>opn/spp1</i>		Bone/milk	Ph <sup>c</sup> , Gl	4q21.3	3,15,37
	Matrix extracellular phosphoglycoprotein	MEPE	<i>mepe</i>	Mu	Dentine/bone	Ph		4
Enamel matrix proteins	Ameloblastin	AMBN	<i>ambn</i>	Mu	Enamel		4q13.3	16,64,65
	Enamelin	ENAM	<i>enam</i>	Mu	Enamel	Ph, Gl	4q13.3	6
Calcium-sensitive caseins	$\alpha_{s1}$ -Casein	CSN1S1	<i>csn1s1</i>	Mu	Milk	Ph <sup>b</sup>	4q13.3	21,38
	$\alpha_{s2}$ -Casein	CSN1S2	<i>csn1s2</i>	Mu	Milk	Ph <sup>b</sup>	4q13.3	21,38
	$\beta$ -Casein	CSN2	<i>csn2</i>	Mu	Milk	Ph <sup>b</sup>	4q13.3	21,38
Salivary proteins	Statherin	STATH	<i>stath</i>	Mu	Saliva	Ph <sup>c</sup>	4q13.3	8,10
	Proline-rich proteins	PROL	<i>prol</i>	Mu	Saliva	Ph	4q13.3	8,10
	Histatin	HTN	<i>htn1, htn3</i>	Mu	Saliva	Ph	4q13.3	9

Te represents template proteins; Un represents unknown function; Mu represents multiple functions; Ph represents phosphorylation; Gl represents glycosylation.

<sup>a</sup> Contains clusters of DSS, where S is phosphorylated.

<sup>b</sup> Contains SSSEE, where S is phosphorylated.

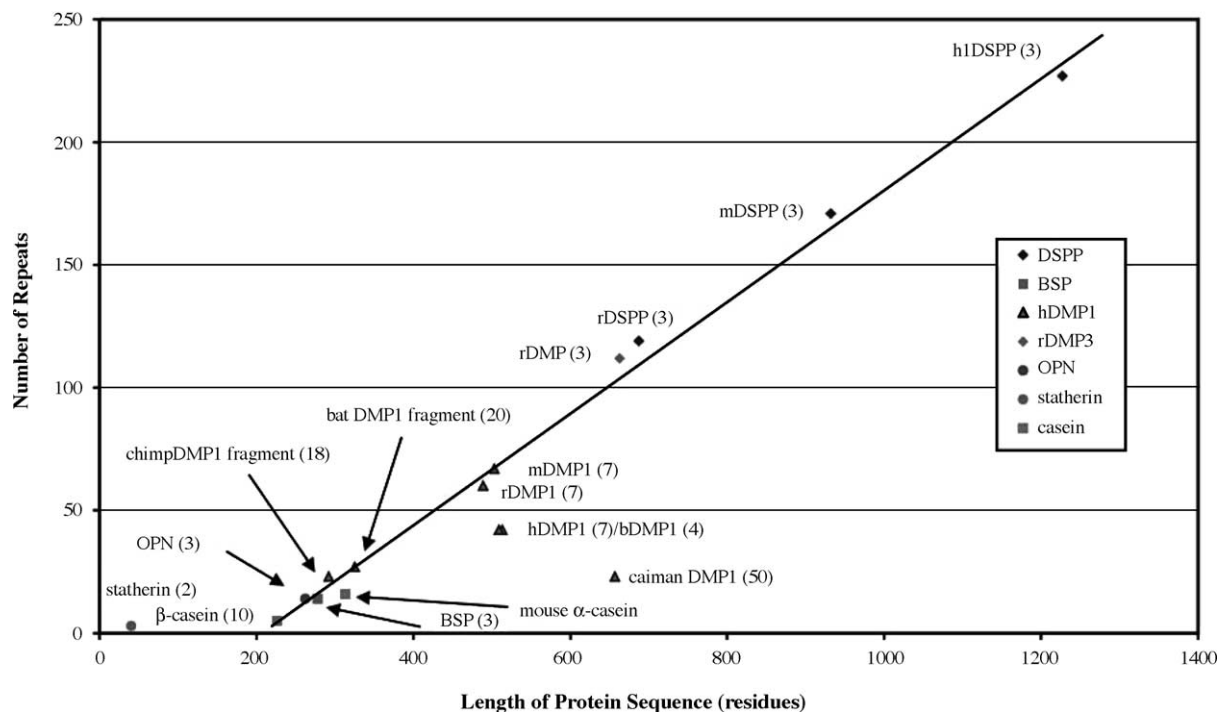
<sup>c</sup> Contains SSEE, where S is phosphorylated.



**Figure 1** Conserved chromosomal synteny observed within the primates (human), artiodactyls (cow, sheep and pig) and rodents (mouse and rat) for the genes encoding the following proteins: milk casein (*csn*), osteopontin (*opn*), bone sialoprotein (*bsp*), dentine matrix protein 1 (*dmp1*), dentine sialophosphoprotein (*dspp*) and ameloblastin (*ambn*) and statherin (*stath*).

in the crystallisation of hydroxyapatite (HA) in enamel is amelogenin (AMEL)<sup>16</sup> which is coded by genes located on both X and Y chromosomes. For all three genes, the introns are phase class 0, i.e. no coding triplet is interrupted by exon boundaries. Due to the structural similarities of the genes, it has been suggested that *ENAM*, *AMBN* and *AMEL* genes have arisen from an ancestral “enamel matrix pro-

tein gene”.<sup>7</sup> Possibly, the first gene duplication generated the *AMEL* gene that was translocated subsequently and the precursor of the *AMBN/ENAM* genes remained on the same chromosome.<sup>7</sup> On the basis of the structure and chromosomal localisation of the *ENAM* gene, the ENAM protein may be considered as a distantly related member of the SIBLING protein family.<sup>14</sup>



**Figure 2** Correlation between the number of repeats as measured by the Internal Repeat Finder and total length of protein. The value in parenthesis indicates the number of residues within a repeat.

### Calcium-sensitive caseins

The caseins, located on human chromosome 4q13, are present in a genomic cluster spanning 250–350 kb depending on the species. This gene cluster is very AT-rich (<37% GC) and has a below average repeat content.<sup>17</sup> Comparative analysis of genomic sequences within the gene cluster reveal that the organisation and orientation of the genes is highly conserved. Within the three genes *csn1s1* coding for  $\alpha_{s1}$ -casein, *csn1s2* coding for  $\alpha_{s2}$ -casein and *csn2* coding for  $\beta$ -casein, the homology between species is greater than between the genes (*csn1s1*, *csn2*, *csn1s2*).<sup>17,18</sup> The conservation between species is mainly in the 5' and 3' untranslated regions, the signal peptide and the major phosphorylation sites.<sup>18</sup> Like the enamel matrix proteins, the casein gene introns are phase class 0. The structure of the *csn2* gene is the most conserved between the species. It has been speculated that the *csn2* gene most closely resembles the ancestral gene sequence.<sup>19</sup>

The number of exons in the CSN1S1-like genes is almost identical and the exon sizes are similar. The variation in mRNA and protein size is mainly due to alternative splicing, duplication, insertion/deletion events and nucleotide substitutions. It is believed that the duplication of the  $\alpha_{s2}$ -casein-like ancestor occurred before the radiation of the artiodactyla-containing clade and the human- and rodent-containing clades. It is possible that the ancestral

Ca-sensitive gene was derived from the ancestral “enamel matrix protein gene”.<sup>7</sup>

### Salivary proteins

The genes for the proline-rich proteins, statherin and the histatin family are also located within this 4q13 region, such that it is possible that the statherin/histatin ancestral gene evolved from the caseins.<sup>7,20,21</sup>

In summary, evidence suggests that the genes coding for the enamel matrix proteins (AMEL, AMBN, ENAM), milk caseins and salivary proteins descended from a common ancestor, the ancestral “enamel matrix protein gene”, by tandem gene duplication.<sup>7,14</sup> It has been further hypothesised that during the evolution of the mineralised skeleton, saliva and lactation, all the bone/dentine genes coding for DSPP, DMP1, OPN, BSP and MEPE and the ancestral “enamel matrix protein gene” arose from a common ancestor by gene duplication and diversification.<sup>7,14</sup> Consequently, all these proteins appear to belong to a “secretory calcium-binding phosphoprotein gene cluster”.<sup>7</sup>

The aim of this review therefore was to compare the protein structure and gene organisation for proteins encoded on human chromosome 4 that are associated with mineralised tissue (teeth and bone) and calcium phosphate stabilisation. Within the subset of these proteins that are multipho-

sphorylated, a further objective was to examine the repeat incidence and codon usage patterns to gain insight into their evolutionary relationships.

### Conserved chromosomal synteny of DSPP, OPN, BSP, CSN, AMBN, and STATH genes

The *csn* and *opn* genes are located on human chromosome 4, sheep and bovine chromosome 6, pig chromosome 8, rat chromosome 14 and mouse chromosome 4 as shown in Fig. 1. The gene loci are available in the following databases: <http://bos.cvm.tamu.edu/sheeparkdb.html>, <http://bos.cvm.tamu.edu/bovgbase.html>, <http://www.genome.iastate.edu/pig> and <http://www.ensembl.org/>. The *stath* gene was identified within the DNA sequence Btau00000000.project\_vbaa\_AB (AC134934) from bovine chromosome 6 (<http://hgsc.bcm.tmc.edu/>). However, the exact chromosomal position is not yet known. The conserved chromosomal synteny supports the hypothesis of a “secretory calcium-binding phosphoprotein gene cluster” arising from the duplication and diversification of ancestral genes during the evolution of the mineralised skeleton, saliva and lactation.<sup>7</sup> The conservation of chromosomal synteny allows the prediction of the location of these genes in other species. For example, as bovine *csn* and *dmp1* are located on chromosome 6, then it is reasonable to predict that the bovine *dspp* gene<sup>1</sup> will also be located on chromosome 6.

### Inversions and translocations of genes

During the evolution of the ungulates (cow, sheep and pig) and rodent (rat, mouse) there have been a number of breakpoints and chromosomal rearrangements in the region of genes corresponding to human chromosome 4.<sup>22,23</sup> For example, on human chromosome 4, the gene order is *csn-ambn-enam-dspp-dmp1-bsp-mepe-opn*. In rat, the order is *opn-mepe-bsp-dspp-ambn-csn*. In mouse, the order is *csn-dmp1-dspp-bsp-opn-dmp2-*. In pig, the order of loci on chromosome 8 is *-opn-ambn-csn-*.

### Prevalence of intrinsic disorder of proteins

The dominant view of protein structure–function is that the amino acid sequence specifies a three-dimensional structure that is a pre-requisite for protein function. However, many proteins do not adopt stable folded structures and, furthermore, display functions requiring intrinsic disorder.<sup>24</sup> It has been predicted that there are over 1000 proteins with long unstructured regions in the SWISS-PROT

protein database.<sup>25</sup> The incidence of disordered proteins have been examined in genomes of 22 bacteria, 7 archaea and 5 eukaryotes.<sup>26</sup> The proportion of the genome encoding disordered proteins was found to increase with the complexity of the organism.<sup>26</sup> Recently, it has been predicted that more than 30% of eucaryotic proteins have disordered regions of >50 consecutive residues.<sup>27</sup>

Of the SIBLING proteins, BSP and OPN have been reported to be completely unstructured and flexible in solution based on NMR spectroscopy studies.<sup>13</sup> Dentine phosphophoryn consists of the tandem repeats Asp–(Ser(P))<sub>n</sub>, where *n* = 1–3. Using NMR spectroscopy, we have recently studied the protein dynamics of the bovine phosphophoryn domain encoded by the *dspp* gene and have demonstrated that this large (~150 kDa) protein is a uniformly flexible molecule which is consistent with its relatively featureless sequence in the mammalian gene family.<sup>28,29</sup> Caseins, containing clusters of phosphoserines have also been designated as disordered proteins on the basis of Raman optical activity studies.<sup>30–32</sup> Salivary proline-rich proteins containing two phosphoserines at the N-termini were designated disordered proteins on the basis of circular dichroism and fluorescent spectroscopy studies.<sup>31,33</sup>

The functional repertoires of unstructured proteins have been recently examined.<sup>24</sup> Molecular recognition is a common function involving binding to other proteins and nucleic acid polymers (DNA, RNA). It was suggested that the lack of structure can confer useful properties in molecular recognition. Disordered regions enable high specificity coupled with low affinity because the free energy arising from the contacts of protein with ligand is reduced by the free energy needed to fold the intrinsic disorder. Furthermore, the disorder enables one protein to bind to differently shaped partners by structural accommodations at the binding interfaces. Also, different disordered sequences can fold to bind a common binding site, and very large interaction surfaces can be created by the unstructured molecule as it wraps-up or surrounds its partner. In addition, faster rates of association can be achieved by reducing dependence on orientation factors and by enlarging target sizes and faster rates of dissociation may be achieved by unzipping mechanisms.<sup>24</sup>

### Multifunctionality of proteins

The properties displayed by disordered proteins explain the multifunctionality reported for the SIBLING proteins, caseins, statherin and proline-rich proteins. For example, phosphophoryn is believed to be responsible for nucleation and regulation of the

growth of HA during dentinogenesis and for the reported CD44-mediated differentiation of pulp cells to odontoblast-like cells.<sup>34</sup> Similarly, OPN and BSP are able to rapidly associate with a number of different binding partners, e.g. integrin and factor H, as well as the mineral phase of bones and teeth<sup>35</sup>. The propensity for multifunctionality of OPN may also be reflected in its broad tissue distribution. Similarly, DMP1 was found to be localised and expressed in the non-mineralised tissues including liver, brain, pancreas and kidney.<sup>36</sup> In contrast, BSP was believed to be restricted to mineralised tissues until BSP, OPN and DMP1 were demonstrated to be expressed in salivary glands of mice and humans along with the three different matrix metalloproteinases (MMP-2, MMP-3 and MMP-9, respectively) that are activated by these three SIBLING proteins.<sup>3</sup> This suggested that the SIBLINGs have additional functions involving local activation of MMPs. Some of the functions of the SIBLING proteins are dependent on the degree of post-translational modification including phosphorylation, glycosylation and proteolytic processing as reported in a recent review.<sup>37</sup>

The milk caseins also display multifunctionality by releasing a range of bioactive peptides during proteolysis in the gut as well as providing peptides/amino acids for nutrition.<sup>38</sup> The functions of these bioactive peptides include enhancement of nutrient absorption, calcium phosphate transport, immunomodulation and opioid activity.<sup>38</sup> Salivary proline-rich proteins bind tannins, in addition to their role of inhibiting the secondary precipitation (seeded-crystal growth) of HA.<sup>39</sup> Salivary statherin is also a multifunctional molecule,<sup>8,40</sup> as in addition to inhibiting both the primary and secondary precipitation of HA in saliva, it also has lubricating and hydrating properties for oral surfaces.

### Amino acid composition

The sequence complexity of structured proteins decreases in the order globular proteins > coiled coil > collagen > silk.<sup>41</sup> Analysis of these disordered sequences revealed that they tend to be of low complexity, over-represented in polar or charged residues and conversely, under-represented in hydrophobic residues.<sup>25,31</sup> The disordered proteins are enriched in Arg, Lys, Gly, Gln, Glu, Pro and Ser residues known as disorder-promoting residues.<sup>27,42</sup> The caseins, BSP and OPN and phosphophoryn are rich in polar and charged residues<sup>1</sup> and the salivary proteins are rich in prolines.

Although the coiled coils and fibrous proteins are enriched in disorder-promoting residues, these proteins are also enriched in specific amino acids that

correspond to their repeating motifs, e.g. Ala, Gly and Pro (silk), Gly and Pro (collagen) and Arg, Gln, Glu (coiled coils). It is interesting to note that the Asp residue is considered to be neither order promoting nor disorder promoting.<sup>27</sup> In contrast, sites of chemical modification, e.g. phosphorylation, are found to occur preferentially at regions of disorder.<sup>43</sup> The caseins, BSP and OPN have multiphosphorylated motifs whereas phosphophoryn is a highly phosphorylated molecule comprising of the repeating sequence motif Asp-(Ser(P))<sub>n</sub>, where  $n = 1-3$ .

### High incidence of repeats

Eukaryotic proteins have been reported to be three times more likely to have internal repeats than prokaryotic proteins.<sup>44</sup> A recent analysis of 126 known disordered proteins, revealed that the percentage of proteins with tandemly repeated motifs is higher within disordered proteins (39%) compared with the overall percentage of tandem repeats in either the yeast (18%) or the human genomes (28%).<sup>45</sup> Casein, OPN and BSP show phosphorylated repeat units, e.g. SSSEE, SGSSEE, whereas phosphophoryn has phosphorylated motifs of DSS.

Low complexity and compositionally biased regions have been observed to evolve rapidly by recombinatorial repeat expansion, replication slippage and substitution mutations within the DNA.<sup>46,47</sup> This has led to a recent proposition that disordered proteins evolve by repeat expansion.<sup>45</sup>

The Internal Repeat Finder (<http://www.doe-mbi.ucla.edu/Services/Repeats>) was used to calculate the number of repeats in several species of DSPP, DMP1, casein and statherin.<sup>44</sup> Fig. 2 shows the correlation between numbers of repeats and length of the protein sequence for DSPP, DMP1, casein and statherin. The number in brackets indicate the number of residues involved in the repeat. Phylogenetic analysis suggests that exon 6 of DMP1 is able to tolerate non-frame shifting insertion and deletion events without loss of function.<sup>48,49</sup> Nevertheless, within the avian, reptilian and mammalian classes, the DMP1 protein overall varies in length between 513 and 657 amino acids, therefore displaying a conservation of protein size. Hence there is a lack of correlation between the number of repeats and length of protein sequence for DMP1.

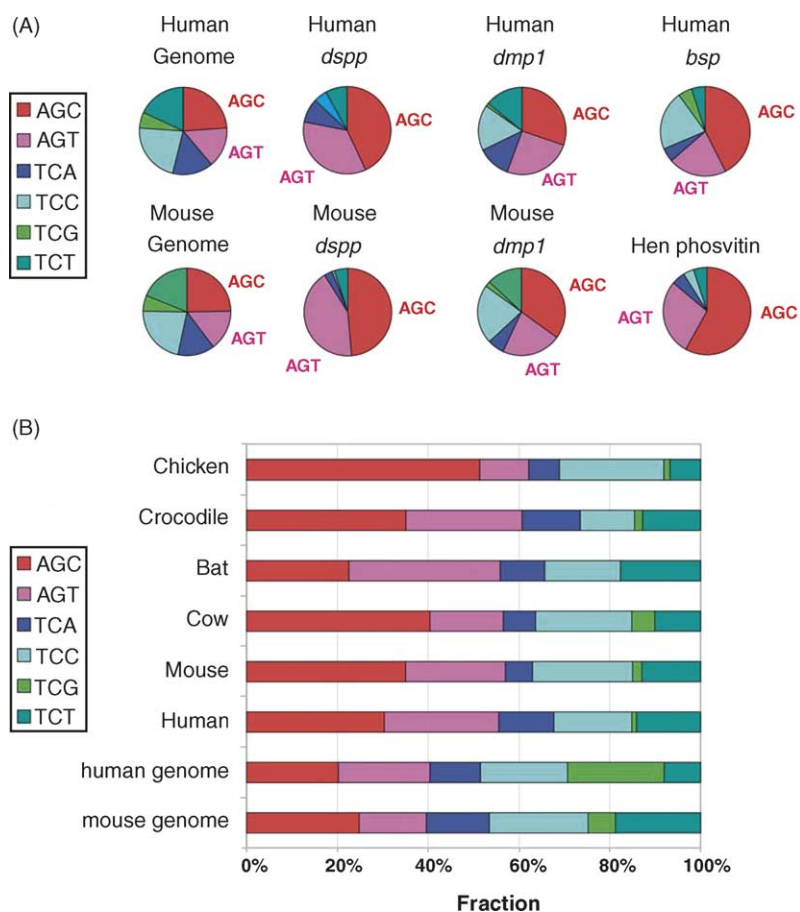
In contrast, DSPP reveals a linear relationship between the number of repeats and total length of sequence. The size of the DSP domain of DSPP is relatively constant within the human, mouse and rat species. Hence, the size variation is provided by the phosphophoryn domain of DSPP. In this case, the protein DSPP supports the recent hypothesis that disordered proteins evolve by repeat expansion.<sup>45</sup>

## Codon bias

Unequal usage of codons is a universal feature of the genomes. The bias in the usage of synonymous codons correlates with the abundance of the corresponding tRNAs and this correlation is particularly strong for highly expressed genes.<sup>50,51</sup> Serine is coded by the six codons AGC, AGT, TCA, TCT, TCC, TCG. However, unequal codon usage is prevalent for serine within the human and mouse genomes. Codon usage was examined using the Graphical Codon Usage Analyser (<http://gcua.schoedl.de/seqoverall.html>). Examination of the coding regions of the proteins listed in Table 1 revealed a bias for AGC and AGT codons within the human genes *dspp*, *dmp1* and *bsp* and in mouse *dspp* and *dmp1* as shown in Fig. 3A. Although it has been suggested that the *dmp1* gene is rapidly evolving and is able to tolerate non-frame shifting insertion and deletion events, strong codon bias was still observed for the coding regions of the *dmp1* gene

within the avian, reptilian and mammalian classes.<sup>48,49</sup> Similarly, codon bias was observed in the phosphovin domain of the chick vitellogenin gene<sup>52</sup> that also contains stretches of multiphosphorylated serines. Significant codon bias was not observed for the serines within statherin or caseins.

Our analyses of the coding regions of these proteins indicate that the codon bias is more pronounced for serines in regions of tandem repeats. This may be explained by the slippage hypothesis that allows elongation and shortening of DNA repeat sequences composed of 1–6 bp long units.<sup>53</sup> DNA polymerase slippage involves the transient dissociation of the replicating DNA strands, followed by misaligned reassociation.<sup>53</sup> The repeat units are inherently flexible and can generate a variety of hairpin, triplex or quadruplex arrangements of DNA strands.<sup>54</sup> Repeat elongation or shortening processes lead to the increase in biological complexity which is considered to be the hallmark of biological



**Figure 3** Codon usage for the amino acid serine in multiphosphorylated proteins. There is a significant bias for the codons AGC and AGT in preference to the codons TCA, TCC, TCG and TCT. (A) Comparison of the human and mouse genomes with the coding regions of the genes *dspp* (dentine sialophosphoprotein), *dmp1* (dentine matrix protein I), *bsp* (bone sialophosphoprotein) and hen (phosvitin). (B) Comparison of the coding regions of the gene *dmp1* (dentine matrix protein 1) within mammalian, reptilian and avian species.

evolution.<sup>55</sup> Hence, repeats that are able to form alternative DNA conformations are expected to be generated more frequently than others.<sup>56</sup>

A survey of sequence repeats consisting of 1–6 bp within several eukaryotic groups has confirmed that the preferred sequence repeat types in exons as well as in other genomic regions are taxon dependent.<sup>56</sup> In addition, repeat abundance and expandability was found to rarely correlate in the case of trinucleotide repeats.<sup>56</sup> Furthermore, trinucleotide repeats in exons exhibited uniformly low expandability. In fact, AGC, coding for serine, is the only trinucleotide motif for which repeats longer than 24 bp can be found in all taxa with expandability values for exon AGC varying between 3% in arthropods and 57% in rodents.<sup>56</sup> A more recent study of tandem repeats in protein coding regions in primate genes has shown that overall, there is no correlation between codon bias and codon propensities in repeat formation.<sup>55</sup>

Nevertheless, our investigations suggest that there is a correlation between codon bias and codon propensities in repeat formation within the genes coding the multiphosphorylated proteins involved in the biomineralisation of hard tissues. This codon bias has survived since the clades split 80–100 million years ago. However, it is unclear what the selection pressure is to maintain the existing codon bias. Perhaps, the codon bias in the ancestral protein reflected translational efficiency, and this bias has been propagated through repeat formation.

Current theories of protein folding proceeding co-translationally, imply a possible influence of mRNA sequence and structural elements on the subsequent protein 3D structure.<sup>57,58</sup> In mammals, synonymous codon families coding for the same residue were reported to have non-random distribution frequencies between protein structure types  $\alpha$ -helix,  $\beta$ -structure and the rest including turns and coil (disordered) regions.<sup>59,60</sup> Furthermore, the relationship between mRNA folding propensities for stem and loop structural elements and the known 3D structures of 105 *H. sapiens* proteins and 88 *E. coli* proteins was investigated to reveal that helices and strands are preferably coded by mRNA stems, whereas the protein coil (disordered) regions are preferably coded by the mRNA loop region.<sup>61</sup> Consequently, the mRNA organisation could have evolved with its codon usage bias to preserve the secondary structures of proteins.

## Conclusion

The genes of the multiphosphorylated proteins statherin, caseins, DSPP, DMP1, BSP and OPN that are mapped to the q arm of human chromosome 4,

have been subjected to conserved chromosomal synteny during mammalian evolution. Except for DMP1, all have been experimentally demonstrated to be intrinsically disordered. The codon usage patterns for the amino acid serine reveal a bias for AGC and AGT codons within the human genes *dspp*, *dmp1* and *bsp*, mouse *dspp* and *dmp1*, and the gene coding for the hen phosvitin protein that also contains stretches of multiphosphorylated serines. Furthermore, there is a correlation between codon usage and codon propensities in repeat formation within the genes coding the multiphosphorylated proteins involved in the biomineralisation of hard tissues.

## Acknowledgment

This work was supported by an Australian National Health and Medical Research Council project grant no. 114202.

## References

1. Huq NL, Cross KJ, Talbo GH, Riley PF, Loganathan A, Crossley MA, et al. N-terminal sequence analysis of bovine dentine phosphophoryn after conversion of phosphoserine to S-propylcysteine residues. *J Dent Res* 2000;**79**:1914–9.
2. George A, Srinivasan R, Thotakura S, Veis A. The phosphophoryn gene family—identical domain structures at the carboxyl end. *Eur J Oral Sci* 1998;**106**:221–6.
3. Ogbureke KU, Fisher LW. Expression of SIBLINGs and their partner MMPs in salivary glands. *J Dent Res* 2004;**83**:664–70.
4. Rowe PS, Kumagai Y, Gutierrez G, Garrett IR, Blacher R, Rosen D, et al. MEPE has the properties of an osteoblastic phosphatonin and inhibitor. *Bone* 2004;**34**:303–19.
5. Sørensen ES, Højrup P, Petersen TE. Post-translational modifications of bovine osteopontin: identification of twenty-eight phosphorylation and three O-glycosylation sites. *Protein Sci* 1995;**4**:2040–9.
6. Hu JC, Yamakoshi Y. Enamelin and autosomal-dominant amelogenesis imperfecta. *Crit Rev Oral Biol Med* 2003;**14**:387–98.
7. Kawasaki K, Weiss KM. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc Natl Acad Sci USA* 2003;**100**:4060–5.
8. Levine MJ. Development of artificial salivas. *Crit Rev Oral Biol Med* 1993;**4**:279–86.
9. Kavanagh K, Dowd S. Histatins: antimicrobial peptides with therapeutic potential. *J Pharm Pharmacol* 2004;**56**:285–9.
10. Van Nieuw Amerongen A, Bolscher JG, Veerman EC. Salivary proteins: protective and diagnostic value in cariology? *Caries Res* 2004;**38**:247–53.
11. Holt C, Van Kemenade MJJM. The interaction of phosphoproteins with calcium phosphate. In: Hukins DWL, editor. *Calcified tissue*. Florida: CRC Press; 1989. p. 175–213.

12. Swaisgood HE, Chemistry of milk protein. Fox PF, editor. *Developments in dairy chemistry*, vol. 1. London: Applied Science Publishers; 1982. p. 1–43.
13. Fisher LW, Torchia DA, Fohr B, Young MF, Fedarko NS. Flexible structures of SIBLING proteins, bone sialoprotein, and osteopontin. *Biochem Biophys Res Commun* 2001;**280**:460–5.
14. Fisher LW, Fedarko NS. Six genes expressed in bones and teeth encode the current members of the SIBLING family of proteins. *Connect Tissue Res* 2003;**44**(Suppl. 1):33–40.
15. Denhardt DT. The third international conference on osteopontin and related proteins. San Antonio, Texas, May 10–12, 2002. *Calcif Tissue Int* 2004;**74**:213–9.
16. Veis A. Amelogenin gene splice products: potential signaling molecules. *Cell Mol Life Sci* 2003;**60**:38–55.
17. Rijnkels M. Multispecies comparison of the casein gene loci and evolution of casein gene family. *J Mammary Gland Biol Neoplasia* 2002;**7**:327–45.
18. Mercier JC, Vilotte JL. Structure and function of milk protein genes. *J Dairy Sci* 1993;**76**:3079–98.
19. Groenen MA, Dijkhof RJ, Verstege AJ, van der Poel JJ. The complete sequence of the gene encoding bovine alpha s2-casein. *Gene* 1993;**123**:187–93.
20. Sabatini LM, Ota T, Azen EA. Nucleotide sequence analysis of the human salivary protein genes HIS1 and HIS2, and evolution of the STATH/HIS gene family. *Mol Biol Evol* 1993;**10**:497–511.
21. Rijnkels M, Elnitski L, Miller W, Rosen JM. Multispecies comparative analysis of a mammalian-specific genomic domain encoding secretory proteins. *Genomics* 2003;**82**:417–32.
22. Lanneluc I, Mulsant P, Saidi-Mehtar N, Elsen JM. Synteny conservation between parts of human chromosome 4q and bovine and ovine chromosomes 6. *Cytogenet Cell Genet* 1996;**72**:212–4.
23. Lord EA, Lumsden JM, Dodds KG, Henry HM, Crawford AM, Ansari HA, et al. The linkage map of sheep chromosome 6 compared with orthologous regions in other species. *Mamm Genome* 1996;**7**:373–6.
24. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry* 2002;**41**:6573–82.
25. Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guillot S, et al. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 1998;**3**:437–48.
26. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;**11**:161–71.
27. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. *J Mol Graph Model* 2001;**19**:26–59.
28. Cross KJ, Huq NL, Loganathan A, Reynolds EC. Structural studies on dentin phosphophoryn. *Aust Dent J* 2003;**53**:4.
29. Cross KJ, Huq NL, Reynolds EC. NMR relaxation studies of bovine dentine phosphophoryn. *Australian and New Zealand Society for Magnetic Resonance Conference, Adelaide, Australia*. 2004.
30. Syme CD, Blanch EW, Holt C, Jakes R, Goedert M, Hecht L, et al. A Raman optical activity study of rheomorphism in caseins, synucleins and tau: New insight into the structure and behaviour of natively unfolded proteins. *Eur J Biochem* 2002;**269**:148–56.
31. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;**41**:415–27.
32. Bhattacharyya J, Das KP. Molecular chaperone-like properties of an unfolded protein, alpha(s)-casein. *J Biol Chem* 1999;**274**:15505–9.
33. Loomis RE, Bergey EJ, Levine MJ, Tabak LA. Circular dichroism and fluorescence spectroscopic analyses of a proline-rich glycoprotein from human parotid saliva. *Int J Pept Prot Res* 1985;**26**:621–9.
34. Fujisawa R, Mizuno M, Kuboki Y. Effect of dentin phosphophoryn of odontoblast-like cells in vitro. *International Association for Dental Research, 79th General Session & Exhibition, Chiba, Japan*. 2001;782.
35. Fedarko NS, Fohr B, Robey PG, Young MF, Fisher LW. Factor H binding to bone sialoprotein and osteopontin enables tumor cell evasion of complement-mediated attack. *J Biol Chem* 2000;**275**:16666–72.
36. Terasawa M, Shimokawa R, Terashima T, Ohya K, Takagi Y, Shimokawa H. Expression of dentin matrix protein 1 (DMP1) in nonmineralized tissues. *J Bone Miner Metab* 2004;**22**:430–8.
37. Qin C, Baba O, Butler WT. Post-translational modifications of sibling proteins and their roles in osteogenesis and dentinogenesis. *Crit Rev Oral Biol Med* 2004;**15**:126–36.
38. Meisel H, FitzGerald RJ. Biofunctional peptides from milk proteins: mineral binding and cytomodulatory effects. *Curr Pharm Des* 2003;**9**:1289–95.
39. Baxter NJ, Lilley TH, Haslam E, Williamson MP. Multiple interactions between polyphenols and a salivary proline-rich protein repeat result in complexation and precipitation. *Biochemistry* 1997;**36**:5566–77.
40. Schlesinger DH, Hay DI. Complete covalent structure of statherin, a tyrosine-rich acidic peptide which inhibits calcium phosphate precipitation from human parotid saliva. *J Biol Chem* 1977;**252**:1689–95.
41. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett* 1999;**462**:363–7.
42. Romero P, Obradovic Z, Li XH, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* 2001;**42**:38–48.
43. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 2004;**32**:1037–49.
44. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. *J Mol Biol* 1999;**293**:151–60.
45. Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 2003;**25**:847–55.
46. Wootton JC. Sequences with unusual amino-acid compositions. *Curr Opin Struct Biol* 1994;**4**:413–21.
47. Sonnhammer EL, Wootton JC. Integrated graphical analysis of protein sequence features predicted from sequence composition. *Proteins* 2001;**45**:262–73.
48. Van Den Bussche RA, Reeder SA, Hansen EW, Hooper SR. Utility of the dentin matrix protein 1 (DMP1) gene for resolving mammalian intraordinal phylogenetic relationships. *Mol Phylogenet Evol* 2003;**26**:89–101.
49. Toyosawa S, O'Huigin C, Klein J. The dentin matrix protein 1 gene of prototherian and metatherian mammals. *J Mol Evol* 1999;**48**:160–7.
50. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 1985;**2**:13–34.
51. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* 2001;**53**:290–8.
52. Prescott B, Renugopalakrishnan V, Glimcher MJ, Bhushan A., Thomas Jr GJ. A Raman spectroscopic study of hen egg

- yolk phosphovitin: structures in solution and in the solid state. *Biochemistry* 1986;**25**:2792–8.
53. Richards RI, Sutherland GR. Simple repeat DNA is not replicated simply. *Nat Genet* 1994;**6**:114–6.
  54. Pearson CE, Sinden RR. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Curr Opin Struct Biol* 1998;**8**:321–30.
  55. Borstnik B, Pumpernik D. Tandem repeats in protein coding regions of primate genes. *Genome Res* 2002;**12**:909–15.
  56. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000;**10**:967–81.
  57. Hardesty B, Kudlicki W, Odom OW, Zhang T, McCarthy D, Kramer G. Cotranslational folding of nascent proteins on *Escherichia coli* ribosomes. *Biochem Cell Biol* 1995;**73**:1199–207.
  58. Kolb VA, Makeyev EV, Kommer A, Spirin AS. Cotranslational folding of proteins. *Biochem Cell Biol* 1995;**73**:1217–20.
  59. Adzhubei AA, Adzhubei IA, Krashennnikov IA, Neidle S. Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett* 1996;**399**:78–82.
  60. Xie T, Ding D, Tao X, Dafu D. The relationship between synonymous codon usage and protein structure. *FEBS Lett* 1998;**434**:93–6.
  61. Jia M, Luo L, Liu C. Statistical correlation between protein secondary structure and messenger RNA stem–loop structure. *Biopolymers* 2004;**73**:16–26.
  62. Butler WT, Brunn JC, Qin C. Dentin extracellular matrix (ECM) proteins: comparison to bone ECM and contribution to dynamics of dentinogenesis. *Connect Tissue Res* 2003;**1**:171–8.
  63. Narayanan K, Ramachandran A, Hao J, He G, Park KW, Cho M, et al. Dual functional roles of dentin matrix protein 1: Implications in biomineralization and gene transcription by activation of intracellular Ca<sup>2+</sup> store. *J Biol Chem* 2003;**278**:17500–8.
  64. Snead ML. Amelogenin protein exhibits a modular design: implications for form and function. *Connect Tissue Res* 2003;**1**:47–51.
  65. Moradian-Oldak J, Iijima M, Bouropoulos N, Wen HB. Assembly of amelogenin proteolytic products and control of octacalcium phosphate crystal morphology. *Connect Tissue Res* 2003;**1**:58–64.