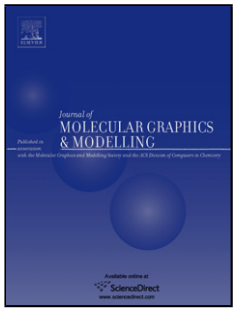
Accepted Manuscript



Title: A Structural and Functional Model for Human Bone Sialoprotein

Authors: Kevin Vincent, Marcus C. Durrant

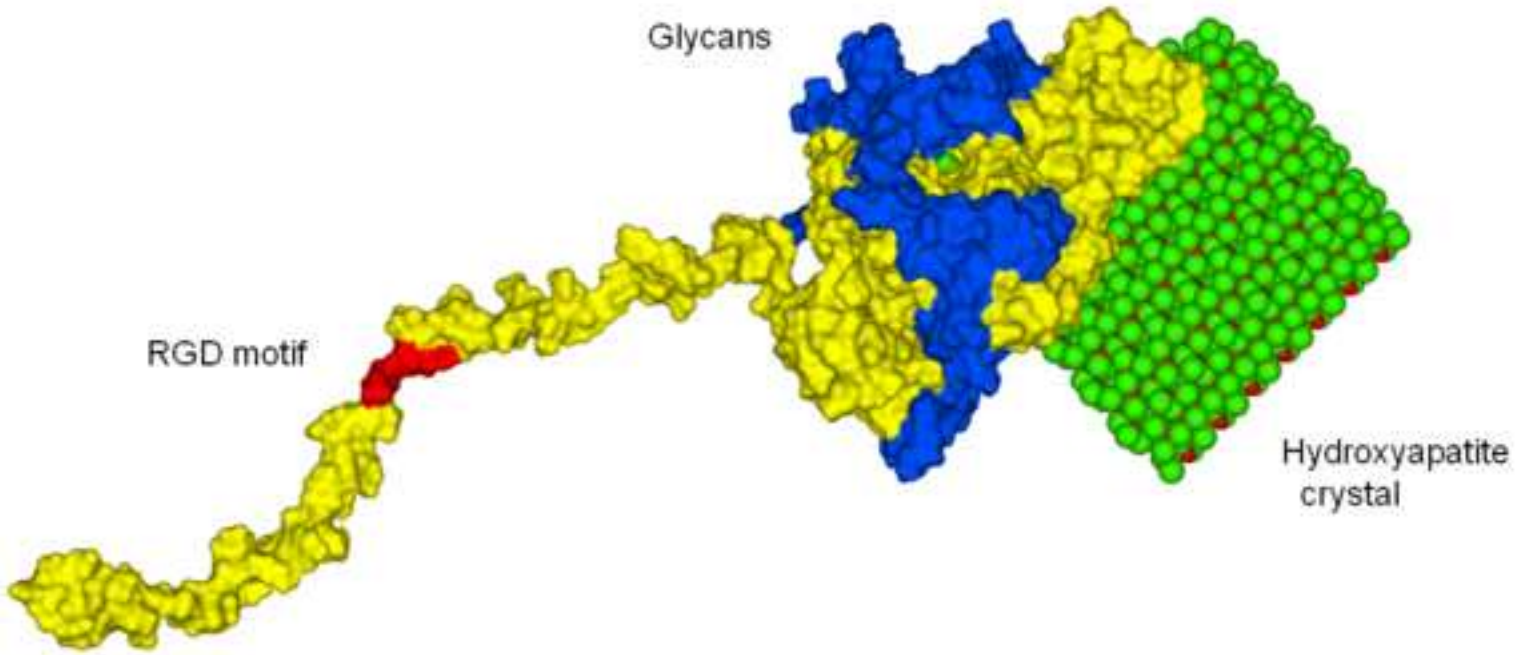
PII: S1093-3263(12)00115-5
DOI: doi:10.1016/j.jmgm.2012.10.007
Reference: JMG 6217

To appear in: *Journal of Molecular Graphics and Modelling*

Received date: 3-8-2012
Revised date: 24-10-2012
Accepted date: 25-10-2012

Please cite this article as: K. Vincent, M.C. Durrant, A Structural and Functional Model for Human Bone Sialoprotein, *Journal of Molecular Graphics and Modelling* (2010), doi:10.1016/j.jmgm.2012.10.007

Graphical Abstract (for review)



- A 3D model for bone sialoprotein has been constructed.
- The model has an acidic surface patch that can interact with calcium ions.
- The role of the protein in hydroxyapatite nucleation has been modelled.
- Simulations indicate that the protein's flexibility is important for nucleation.
- Quantum calculations have been used to probe structure-function relationships.

A Structural and Functional Model for Human Bone Sialoprotein

Kevin Vincent[a], Marcus C. Durrant[b,*]

[a]*Department of Chemistry, Durham University, South Road, Durham, DH1 3LE, United Kingdom*

[b]*Faculty of Health and Life Sciences, Northumbria University, Ellison Building, Newcastle-upon-Tyne, NE1 8ST, United Kingdom*

*Corresponding author. Tel.: +44 191 2437239.

E-mail address: marcus.durrant@northumbria.ac.uk (M.C. Durrant)

ABSTRACT

Human bone sialoprotein (BSP) is an essential component of the extracellular matrix of bone. It is thought to be the primary nucleator of hydroxyapatite crystallization, and is known to bind to hydroxyapatite, collagen, and cells. Mature BSP shows extensive post-translational modifications, including attachment of glycans, sulfation, and phosphorylation, and is highly flexible with no specific 2D or 3D structure in solution or the solid state. These features have severely limited the experimental characterization of the structure of this protein. We have therefore developed a 3D structural model for BSP, based on the available literature data, using molecular modelling techniques. The complete model consists of 301 amino acids, including six phosphorylated serines and two sulfated tyrosines, plus 92 *N*- and *O*-linked glycan residues. A notable feature of the model is a large acidic patch that provides a surface for binding Ca^{2+} ions. Density functional theory quantum calculations with an implicit solvent model indicate that Ca^{2+} ions are bound most strongly by the phosphorylated serines within BSP, along with reasonably strong binding to Asp and Glu, but weak binding to His and sulfated tyrosine. The process of early hydroxyapatite nucleation has been studied by molecular dynamics on an acidic surface loop of the protein; the results suggest that the cationic nature of the loop promotes nucleation by attracting Ca^{2+} ions, while its flexibility allows for their rapid self-assembly with PO_4^{3-} ions, rather than providing a regular template for crystallization. The binding of a hydroxyapatite crystal at the protein's acidic patch has also been modelled. The relationships between hydroxyapatite, collagen and BSP are discussed.

Keywords

Hydroxyapatite; biomineralization; molecular modelling; osteogenesis; SIBLING proteins

1. Introduction

Bone is a composite structure composed of mineral and organic components. The mineral consists of small, tile-shaped crystals of carbonated hydroxyapatite (HA), where pure HA has the stoichiometric formula [$Ca_5(PO_4)_3(OH)$]. The crystals are embedded in an organic matrix composed mainly of fibres of the structural protein type I collagen, as well as other proteins, of which the most abundant is bone sialoprotein (BSP) which accounts for up to ~15% of the non-collagenous protein [1]. Mature human BSP has 301 amino acids and includes extensive and varied post-translational modifications that account for roughly one third of the molecular mass. Zaia *et al.* analysed human BSP by MALDI-TOF and observed a peak for the as-extracted BSP centred at 52.5 kDa with a width of 11.9 kDa [2], reflecting the heterogeneous nature of the post-translational modifications. Wuttke *et al*. similarly used MALDI-TOF to characterize recombinant and bone-derived BSP samples [3]. The recombinant BSP gave a broad peak between 40 – 75 kDa, average mass 57 kDa, whilst bone-derived BSP gave a peak between 40 – 60 kDa, average mass 49 kDa. The mass of the core protein, calculated from its primary structure, is 33.5 kDa. The modifications have been identified as *N*- and *O*-glycosylation, phosphorylation of serine and/or threonine, and sulfation of tyrosine residues. Since the glycans include sialic acid residues, each of these modifications serves to increase the acidity of the mature protein, which already has a preponderance of acidic residues (60 Glu and 16 Asp, compared to 14 Lys and 9 Arg). BSP sequences from various animal species are fairly similar and all contain regions of contiguous glutamic acid residues; human BSP has two $(Glu)_8$ sequences.

In the absence of other proteins or ligands, BSP has little secondary structure and a highly variable tertiary structure. Thus, one dimensional ^{1}H NMR of recombinant BSP revealed that the protein is highly flexible along its entire length and exists as an ensemble of

completely unstructured conformations in solution [4]. Similarly, two dimensional HSQC NMR of an ^{15}N labelled recombinant BSP fragment taken from the last 59 amino acids of the protein C-terminus indicated a rapidly flexing random coil structure [5]. Secondary structure prediction programs such as PsiPred [6] also suggest that the unmodified BSP would exist as an essentially unstructured random coil. Circular dichroism (CD) spectroscopy of BSP indicated 5% α-helix, 32% β-sheet, 17% β-turn and 46% random coil [3], although it has been suggested that these values may over-estimate the helix and sheet content, due to a bias in the database of standard proteins used in the analysis [7]. Small angle X-ray scattering experiments on recombinant rat BSP gave results consistent with a random coil [7]. It seems likely that the structural elements suggested by CD are the averages for an ensemble of diverse and transient conformations. Electron microscopy of recombinant BSP showed a monomeric structure, consisting of a globule of diameter 10 $\pm$ 1 nm linked to a thread-like structure of 25 $\pm$ 6 nm length [3]. In contrast, BSP isolated from bovine bone appeared as a simpler, threadlike structure with an average length of ~40 nm [8].

The biological function of BSP has yet to be fully defined. Experiments using a steady state agarose gel system showed that BSP induces the crystallization of hydroxyapatite *in vitro* [9], supporting the commonly accepted view that BSP acts as a nucleator for growth of HA crystals within the extracellular bone matrix. Further experiments showed that removal of the phosphate groups from mature BSP did not much diminish the protein's ability to nucleate HA, whereas derivatisation of the glutamic acid residues abolished the ability. Interestingly, this ability seems to be specific for glutamic acid; thus, poly(Glu) can nucleate HA, whereas poly(Asp) cannot [10]. BSP interacts strongly with HA crystals, with a dissociation constant of ~2.6×10^{-9} M for intact BSP from a rat osteosarcoma cell line [5]. BSP also binds to type I collagen. Tye *et al.* have probed the nature of this interaction in rat proteins, using bone-extracted BSP and recombinant peptides [11]. The interaction between

BSP and collagen was found to be partially electrostatic and partially due to more specific non-bonded interactions. The poly(Glu) regions of BSP were found to be relatively unimportant in this context, whilst residues 19-46 provide specific short range interactions. These 28 residues are highly conserved across different species, and in humans include nine Tyr and Phe residues, plus five Arg and Lys residues. The ability of recombinant BSP to nucleate HA is enhanced in the presence of collagen [11]. Since type I collagen itself binds to HA [1], it seems likely that BSP serves both to nucleate HA and also to enhance the interaction between the growing HA crystals and collagen. Although it is also possible that BSP influences the shape of the HA crystals, which are remarkably thin in bone (*ca.* 3 nm), it has recently been shown that the surfaces of the crystals are decorated with strongly bound citrate, which accounts for 5.5% of the organic content of bone by weight; and this appears to be the main determinant of the crystal morphology [12].

At the cellular level, BSP is expressed at high levels by osteoblasts under bone formation conditions, and at lower levels by odontoblasts, cementoblasts and hypertrophic cartilage cells. Its patterns of expression and distribution are consistent with the hypothesis that BSP plays an essential role in HA nucleation, as well as formation and remodelling of bone [1,7]. Along with the related SIBLING (small integrin binding ligand, *N*-linked glycoprotein) protein osteopontin, BSP is expressed in malignant tissues, and indeed elevated serum levels of these proteins can be used to detect several types of cancer [13]. Expression of BSP is strongly associated with bone metastases in lung [14] and bone [15] tissues.

The primary amino acid sequences of BSP's from a range of species include a well-conserved RGD triplet near the C-terminus, which facilitates attachment of the protein to the integrins of several cell types [5]. Cell adhesion can also be accomplished independently of the RGD motif by two tyrosine-rich regions of BSP, one near the N-terminus and the other just before the RGD triplet. The N-terminal region associated with cell adhesion overlaps to

a great extent with the collagen-binding region [11]. Since BSP binds strongly to HA, it is important that individual molecules of the protein are prevented from free diffusion into solution by anchoring on cells and/or collagen fibrils, since binding of multiple BSP molecules to a single crystal of HA could coat the surface and so poison further growth, as is observed for solutions of free BSP *in vitro* [5].

Experimental determination of the structure of BSP is a highly desirable goal, but fraught with severe difficulties. The lack of a fixed structure, together with the extensive but variable post-translational modifications, makes the crystallization of BSP for X-ray crystallography exceptionally difficult. The NMR studies discussed above have shown only a lack of recognisable structural elements. In these circumstances, the most useful structural information so far has come from mass spectrometric studies. Two papers, both published in 2001, have provided most of what is known about the post-translational modifications of human BSP [2,3]. Zaia *et al.* [2] characterized BSP extracted from human bone by MALDI-TOF, as a broad peak centred at *m/z* 52,533, with a width of 11,850. The broadness of the peak is consistent with a variable array of modifications. Treatment with *N*-glycanase, *O*-glycanase and neuraminidase allowed for partial characterization of the glycans and their points of attachment, whilst use of aminopeptidase M gave information about the most likely positions of sulfated tyrosine residues. Based on their results, this group proposed two *N*-linked oligosaccharides of composition $(GlcNAc)_5(Man)_3(Gal)_3(NeuAc)_3$ and $(GlcNAc)_6(Man)_3(Gal)_4(NeuAc)_4$, with Asn-161 and -166 as the most likely sites of *N*-glycosylation (note that throughout this paper, the residue numbering does not include the 16-residue *N*-terminal signal peptide). The *O*-linked oligosaccharides were proposed to have a total composition of $(GalNAc)_{16}(GlcNAc)_4(Gal)_{15}(Glc)_6(NeuAc)_{16}$, divided among 16 individual oligosaccharides. Possible sites of *O*-glycosylation were deduced from the experimental data and bioinformatic analysis. It was also possible to identify Ser-15 as being

phosphorylated, and a number of other phosphorylated serines were more tentatively proposed. Two sulfated tyrosines were identified and traced to alternative pairs; either Tyr-259 or -262, and either Tyr-297 or -298. These pairs flank the RGD domain and so could serve to modulate the interaction between BSP and integrin.

Wuttke and co-workers investigated human BSP, both as the bone-extracted protein and as a recombinant protein expressed in a human embryonic kidney cell line [3]. There were differences between the *N*-glycans of the extracted and recombinant proteins. Enzymatic cleavage of the glycans from the protein, followed by removal of the NeuAc residues with neuraminidase, HPLC, and MALDI-TOF analysis, allowed for the structural characterization of a variety of asialo-*N*-glycans, of which the most abundant type represented 57% of the total in bone-derived BSP. Asparagines Asn-88, -161, 166 and -174 were identified as the most likely sites of *N*-glycosylation. Similarly, the *O*-glycans were cleaved from the protein and analysed by HPLC followed by MALDI-TOF. Here again, there were differences in composition between the recombinant and extracted proteins; several glycans were structurally characterized and the most abundant glycan represented 66% of the total in the bone-derived sample. Seven threonines were suggested as *O*-glycosylation sites.

Summing up, experimental studies have provided a body of data on the primary structure of BSP, which can be augmented by bioinformatic analysis. Given the central role that BSP plays in bone growth, and that an experimental structure is unlikely to appear in the near future, we decided to construct a theoretical model for human BSP. Since this protein clearly has little or no secondary structure and a highly variable tertiary structure, our main aim was to provide an initial 3D model for BSP that is consistent with all the available experimental data. In order to provide further insights into the function of this protein, we have also probed the fundamental interactions between calcium cations and the available

anionic groups of BSP by quantum calculations, and investigated the nucleation of HA by a sub-section of our model, using molecular dynamics. The results of our studies are presented in this paper.

2. Computational methods

2.1 Molecular mechanics modelling

Initial phases of the molecular modelling were carried out using Hyperchem™ release 8.0.8 [16] to build small subsections of the molecule. The complete model was assembled and refined using Accelrys Discovery Studio version 1.7 [17]. For the modified amino acids and glycan residues, formal fractional charges were set manually by dividing the total charge on the anionic group among its terminal oxygen atoms, and partial atomic charges were used as calculated by the software. Interactive analysis of the effects of individual Ramachandran angles on the protein fold was carried out using the DeepView program [18]. Geometry optimizations were carried out with the CHARMm force field [19] and adopted basis Newton-Raphson algorithm to an RMS gradient of 0.1 kcal mol^{-1} $Å^{-1}$. In the Ramachandran plot of the final model, all non-glycine residues were within the 'generously allowed' regions, and all proline residues were within the 'strictly allowed' regions. Geometry optimizations and molecular dynamics (MD) calculations used the generalized Born with simple switching (GBSW) implicit solvation model, with an implicit solvent dielectric constant of 80 and spherical cut-off electrostatics, non-bond list radius 14 Å, non-bond higher and lower cut-off distances of 12 and 10 Å respectively. The SHAKE parameter was used for MD simulations but not for geometry optimizations. The time step for all MD simulations was 0.002 ps.

2.2. Quantum calculations

Density functional theory (DFT) calculations were run with Gaussian09W [20], using the B3LYP functional and water as implicit solvent for all calculations. Geometry optimizations and frequency calculations were run using the 6-31G(d,p) basis set and Polarized Continuum Model implicit solvent corrections [21], whilst final energies were obtained for single point calculations at the optimized geometries, using the 6-311++G(3df,3pd) basis set and Solvation Model Based on Solute Electron Density implicit solvent corrections [22]. All reported structures were verified as minima by the absence of calculated imaginary frequencies.

3. Results and discussion

3.1 Primary and secondary structure of the model

The primary amino acid sequence of human BSP was retrieved from the UniProt online database, accession code P21815. The 16-residue *N*-terminal signal peptide was deleted to give the mature sequence. This is shown in Figure 1, as a sequence alignment with BSP's from seven other species. All of the sequences have at least one region of eight or more contiguous glutamate residues, preceded by aspartates and/or potentially phosphorylated serines, extending these acidic regions. In addition, there is a conserved DSSEE sequence just upstream of the first poly(Glu) region; if the serines in this sequence are phosphorylated, as suggested by the available experimental data (see below), this provides an additional patch of five contiguous acidic residues. The chicken sequence is unique in having a third region of 12 contiguous glutamates. Other important features of the sequence include a collagen binding region, a highly conserved RGD motif, and a cell binding region (see below). During the construction of our model for BSP, four types of post-translational modifications were included; namely *N*-glycosylation, *O*-glycosylation, phosphorylation, and sulfation.

Although there is a shortage of definitive information on these modifications, the literature does provide sufficient data to allow reasonable assumptions concerning their positions and nature, as follows.

We start with the potential sites of *N*-glycosylation. Wuttke *et al.* predicted four such sites, at asparagines Asn-88, -161, -166 and -174 [3]. Zaia *et al.* also noted four potential sites, but the mannose content of their protein suggested that only two of these are occupied [2]. Furthermore they were able to rule out Asn-88 from their mass spectrometric analysis, whilst Asn-174 is not universally conserved among mammalian BSP sequences, being His in the bovine sequence, for example (Figure 1). Therefore, we have chosen Asn-161 and -166 as the sites of *N*-glycosylation in our model. To construct the *N*-glycan, we started with the most common species identified by Wuttke *et al.* which constituted 57% of the total asialo *N*-glycan of extracted BSP. Browsing the Functional Glycomics Gateway [23] suggested that the most plausible way to reconstruct the sialic acid residues would be via the commonly observed NeuAc(α-2-3)Gal linkage, which was also identified by Wuttke *et al.* in BSP *O*-glycans. Therefore, we added NeuAc residues to each of the three terminal Gal residues, giving the complete *N*-glycan shown in Figure 2. This structure was subjected to preliminary MD in order to find a reasonable conformation, before splicing into the BSP model at the two sites noted above.

Progressing to the *O*-glycan, Zaia *et al.* deduced that the *O*-linked oligosaccharides have a total composition of $(GalNAc)_{16}(GlcNAc)_4(Gal)_{15}(Glc)_6(NeuAc)_{16}$, divided among 16 individual oligosaccharides. Meanwhile, Wuttke *et al.* were able to characterize a number of different *O*-linked glycans in their sample of bone-extracted BSP, finding that the most abundant of these accounted for 66% of the total. We chose this species for all the *O*-linked glycans in our model (Figure 2). At present, there is no firm evidence or consensus as to exactly which sites are *O*-glycosylated in human BSP. Experimental studies using a *Galnt1*-

null mouse model have revealed the sites of *O*-glycosylation in mouse BSP, but not all of these are conserved between the human and mouse sequences [24]. We decided to score each of the possible *O*-glycosylation sites based on the overall consensus between the mouse studies, the proposed sites of Zaia *et al.* and Wuttke *et al*., the conservation of each residue across the sequence alignment shown in Figure 1, and the predictions of the NetOGlyc server [25]. We then selected the eight most likely sites on this basis. The highest scoring site was Thr-223, which received the highest score from NetOGlyc (0.70), is conserved in six out of eight of the amino acid sequences, and was identified in all three experimental studies. The lowest scoring site selected was Thr-211, which scored 0.68 from NetOGlyc, is also conserved in six out of eight sequences, and was identified by Wuttke but not by Zaia or the mouse studies. Eight *O*-glycosylation sites may be slightly lower than the number found in a typical molecule of BSP, but this would probably not have a major effect on the overall structure of the protein.

As for the sulfation sites, Zaia *et al.* identified Tyr-259 or -262 and Tyr-297 or -298 as being sulfated from their experiments. These four residues are generally well conserved, with the exception of Tyr-259 which is Glu in the chicken sequence (Figure 1). Therefore, in order to choose between the alternatives, we searched the Protein Data Bank [26] for proteins containing sulfated tyrosines with local sequences comparable to those found in BSP (considering four residues on each side of the tyrosine and scoring alignments with the 250 PAM mutation data matrix). In this way, we identified BSP Tyr-262 as the more likely of the first pair, by local alignment with one of the three sulfated tyrosines in PDB structure 2K05 [27] [Figure 3(a)]. This also fits with the sequence conservation considerations noted above. We were unable to find a comparable sulfated sequence for either of the second pair of tyrosines in the PDB. However, internal comparison of the local sequences for the four BSP tyrosines gave a clearly better PAM score for Tyr-262 plus Tyr-298 compared to the other

three permutations [Figure3(b) and (c)]. Therefore, Tyr-298 was taken as the second of the two sulfated tyrosines.

Finally, we considered the most likely sites of phosphorylation. On average, BSP has six phosphorylated serine or threonine residues [1,28], and Ser-15 has been positively identified as one of these [2]. Once again, however, there is no hard evidence as to the other sites of phosphorylation in human BSP. Studies on rat BSP identified Ser-15, Ser-50, Ser-51 and Ser-136 as phosphorylated, all of which are conserved in the human protein; moreover Ser-136 was identified as critical for HA nucleation [28]. Detailed experimental studies of the bovine protein have identified a total of eleven phosphorylated serines [29,30], of which nine are conserved in human BSP. Using this information, together with predictions from the NetPhos server [31] and our sequence alignment, we chose the six phosphorylation sites included in our model (Figure 1). These all scored 0.988 or better in NetPhos and were conserved in at least seven of the eight sequences in our alignment; in addition, the corresponding bovine residues were identified as phosphorylated [29,30]. Having chosen all these post-translational modifications, the primary structure of our model is as shown in Figure 1. It is worth pointing out that although there is a degree of uncertainty over several of these assignments, small changes in the primary structure of our model would not materially affect its overall 3D structure or our functional modelling studies. Moreover, since natural BSP is itself inhomogeneous, with different molecules supporting a variety of post-translational modifications, our model should be considered as a representative example of the structure, rather than an attempt to provide a single definitive model.

As discussed in the Introduction, the general consensus from experimental studies is that free BSP is highly dynamic, with no regions of fixed secondary structure. Therefore, to construct our 3D model, we initially chose random backbone φ and ψ angles for each residue,

within the constraints of the Ramachandran plot. This gave an initial extended conformation which was subsequently modified, as discussed below.

3.2 Tertiary structure of the model

Wuttke *et al.* used electron microscopy to visualize recombinant BSP as a globule attached to a threadlike structure [3]. We decided to use this as a basis for building the 3D conformation of our model. In the absence of any means of identifying which part of the protein formed the globule, it was suggested on the basis of the distribution of glycans that the globule would correspond to the C-terminal part of the protein. However, our analysis of the available data, discussed above, indicates that the glycans are clustered around the central part of the amino acid sequence. We have therefore reconsidered this question from the point of view of the distribution of hydrophobic, basic and acidic residues. Since the protein is dominated by acidic residues, the hydrophobic and basic residues would seem to be the most likely candidates to aggregate into a globule; the former due to the normal hydrophobic effects, and the latter due to electrostatic attraction to the neighbouring acidic residues. Considering the primary sequence, the region of the protein most likely to form the interior of a molten globule is then the *N*-terminus, specifically residues 1 – 12 and 18 – 47. We therefore adjusted the φ and ψ angles of individual residues to form a molten globule structure, starting from the *N*-terminus. This still allows all the modified amino acid residues and glycans to occupy surface positions, as would be expected for these hydrophilic groups. According to Wuttke *et al.* [3], the diameter of the globule is 100 Å and the length of the thread is 250 Å, meaning that the globule constitutes ~30% of the total length of the molecule. We constructed our model on this basis, meaning that the residues from the *N*-terminus up to Glu-244 were built into the globule. Residue Glu-244 comes at the start of the *C*-terminal cell binding region (Figure 1). We included 30 Ca^{2+} ions, distributed among the most acidic

portions of the amino acid sequence. Our initial model was refined by local annealing to remove close contacts and/or strained local geometry, followed by a final global energy minimisation, using the GBSW implicit solvation model throughout. This gave an irregular globule of ~55 Å diameter, plus a thread of ~145 Å. Although these dimensions are both smaller than those given by Wuttke, the ratio is similar. To put these values into context, a compactly folded spherical protein of equivalent molecular weight to our entire model would have a diameter of ~52 Å (based on the mean diameter of the near-spherical protein 1UG6 [32], which has a molecular weight of 48.4 kDa). Hence, unless BSP has a very open structure, it seems possible that the experimental dimensions may be somewhat over-estimated.

Figure 4 shows the overall structure of our model. It includes 30 calcium ions and has a molecular weight of 48,870, which fits within the range observed by MALDI-TOF [2,3], bearing in mind that our model contains no water molecules. The overall charge is -30. All of the glycan residues lie in a ring on the surface of the globule, covering ~34% of its surface area. A key feature of this model is that the most acidic stretches of the amino acid sequence are contiguous such that there is a large patch of acidic residues on one face of the globule. This includes 22 Glu, 5 Asp and all 6 phosphoserines in our model, giving an overall charge for the patch of 39-. This arrangement is stabilised by the calcium ions, such that the region carries a net positive charge. When the calcium ions are deleted, the surface of the model clearly shows the acidic patch (Figure 5). A few of the calcium ions lie within the globule, but most are on the surface in a way that suggests they could be transformed fairly easily into a more crystalline arrangement. In order to test this idea, we used the published crystal structure of human dental HA [33] to construct a nanocrystal of approximate dimensions 48 × 41 × 13 Å for docking to our model. We replaced 20 of the individual calcium ions in our model with the nanocrystal; in order to achieve satisfactory results from subsequent geometry

optimization, we also deleted four of the calcium ions from the surface layer of the nanocrystal, allowing the phosphoserines to occupy the resulting voids. Overall, this allows for an intimate association between the protein and the nanocrystal, as shown in Figure 6. The contact area between the crystal and the protein is ~ 640 $Å^2$. Although this is rather less than half the size of a typical protein-protein interface [34], the nature of the BSP-nanocrystal interface is very different, being dominated by ionic rather than non-polar interactions. Taken with the results of our quantum calculations (see below), our model is entirely consistent with the idea that BSP forms a strong association with the HA crystals in bone.

Studies on rat BSP2 showed that residues 18 - 45 provide the collagen binding site [11]. This part of the sequence is well conserved and dominated by hydrophobic and basic residues (Figure 1). Since we assumed that these types of residue would help form the molten globule, in our model, this part of the sequence is located at its centre, near to but distinct from the HA binding site. It seems likely that facile rearrangement of the globule would allow BSP to effectively become threaded onto the collagen triple helix, whilst preserving the HA binding surface. Interestingly, the collagen binding region has also been shown to support attachment of BSP to cells, in addition to the cell binding region just in front of the RGD motif [5]. Hence, the *N*-terminal region can apparently bind either to the cell or to collagen, but not to both at the same time. This may provide a basis for BSP molecules to be transferred from the cell surface to collagen, in which case sulfation of the *C*-terminal tyrosines could help to complete the transfer.

3.3 Dynamic modelling of hydroxyapatite nucleation

We next considered how BSP might serve as a nucleator for HA. This question has recently been addressed by two different groups, using MD simulations. In their MD calculations, Sahai and co-workers [35,36] used a model peptide of sequence $S^PS^PEEEEEEEE$ (S^P =

phosphoserine) in a periodic box with explicit solvent plus Ca^{2+} and P_i ions (P_i = HPO_4^{2-} or $H_2PO_4^-$). The peptide was in either α-helical or random coil conformations. The first 1 ns of the simulation was used for equilibration, and the data collections lasted for up to 35 ns. On this timescale, Ca^{2+} ions remained localized on the peptide sidechains, with no exchange during the simulations. A notable feature for the α-helical peptide model was the transient formation of equilateral triangles of Ca^{2+} ions and it was suggested that the protein could serve as a template for the formation of such triangles, which are related to the geometry of Ca^{2+} ions on the (001) face of HA. Nevertheless, based on the relative scarcity of these structures in their simulations, the authors concluded that BSP is more likely to nucleate amorphous calcium phosphate rather than crystalline HA. Upon extending their studies to consider binding of the S^PS^PEEEEEEEE peptide to different faces of the HA crystal, there was no obvious structural templating effect from the peptide, which could be accommodated on the (001), (100) and (110) crystal faces. Baht *et al.* used MD to simulate the association of two different glutamate-rich 16-mers from the rat protein, both with and without phosphorylated serines, with the (100) face of HA [28]. They used a simple point charge solvent model and a periodic box for these simulations. They found that the peptides tend to show an alternating pattern of residues pointing towards and away from the crystal face.

For our MD studies, we extracted a 24 residue segment from our model for BSP, consisting of the sequence SSDS^{P}S^{P}EENGDDS^{P}S^{P}EEEEEEEETSN. The terminal residues were fixed according to their spatial arrangement in the complete model, and simulations were run using the GBSW implicit solvent model. We also used fully deprotonated PO_4^{3-} ions, with no periodic boundary conditions; although a small number of the simulations failed due to random diffusion of an ion away from the peptide, both the Ca^{2+} and PO_4^{3-} ions tended to remain associated with the protein provided that the overall charge on the system was kept close to zero. In order to validate this approach, we first modelled the calcium binding sites

of the sodium/calcium exchange protein 2QVM [37]. This protein structure has two surface-accessible Ca^{2+} ions, coordinated by Asp and Glu residues. A 20 ns simulation of the calcium binding region using our standard conditions showed that both Ca^{2+} ions remained within their experimentally determined binding sites, with only minor fluctuations in their coordination environments and the overall energy and fold of the model throughout the simulation.

Returning to BSP, since we were particularly interested in the behaviour of this system before an equilibrium state is reached, we used two different methods to probe the dynamic behaviour, which gave very similar results. First, we used a very short heating phase of 14 ps, followed directly by a production phase of 20 ns. Alternatively, we introduced a set of seven distance constraints to keep the five PO_4^{3-} ions separated by their initial distances (these ranged from 7.2 - 16.2 Å) during a heating phase of 0.2 ns, followed by an equilibration phase of 1.0 ns; the constraints were then removed for a simulation of 20 ns. Figure 7 shows a typical energy-time plot from these calculations. After an initial rapid stabilization phase, the plot shows characteristic stepwise decreases in the total energy of the system that take place over ~0.1 ns, separated by constant energy states that persist for ~1 - 10 ns. Most of the energy decreases are associated with the spontaneous self-assembly of PO_4^{3-} and/or phosphoserine phosphate groups with Ca^{2+} ions, whilst the others can be related to changes in the peptide backbone conformation. A typical Ca^{2+} - PO_4^{3-} assembly event is shown in detail in Figure 8; the clustering of the four Ca^{2+} ions with three PO_4^{3-} ions in this example is associated with a permanent decrease in the energy of the system. Throughout all our simulations, once phosphate groups became associated via bridging calcium ions, they remained together for the rest of the calculation. As expected, the overall conformation of the loop varied quite markedly throughout the simulations and also from run to run. These observations suggest that although the peptide clearly facilitates HA nucleation, it should not

be described as a template; rather, its highly flexible structure allows it to respond to the movements of the anions and cations as they undergo spontaneous self-assembly. The primary role of the acidic loops is then to attract calcium ions, which in turn attract phosphates. This is consistent with experiments that show the poly(Glu) regions are essential for HA nucleation [10], and is also supported by the results of our quantum calculations (see below). The aggregated ions then begin to pack around the peptide strand, and in particular the phosphoserines, which become tightly integrated into the growing mass of amorphous calcium phosphate. We suspect that the mature HA-BSP interface may be characterized by a transition from an intimate and amorphous protein-calcium phosphate mixture to the regular structure of the bulk nanocrystal, rather than a simple association of the protein with a crystal surface.

3.4 Ion binding energies from quantum calculations

An important consideration is the strength of interactions of calcium ions with both free phosphate, and the available acidic groups on the protein. We have investigated these interactions by DFT quantum calculations. We begin by noting that we are interested in the interactions between various ionic species at the protein-water interface. Therefore, it is essential to take solvent effects into account during geometry optimizations, since the gas phase structures of the ion pairs will be markedly different. We therefore used implicit solvent models for both geometry optimizations and final energy calculations with a large basis set (see Computational Details). The results are shown in Table 1. We have included zero point energy (ZPE) corrections but not basis set superposition corrections, which cannot easily be calculated for highly polarised systems such as $[Ca(PO_4)]^-$. Trial gas phase calculations on $[Ca(OH_2)]^{2+}$ and $[Ca(MeCO_2)]^+$ gave basis set superposition errors of 0.6 and 1.1 kJ/mole respectively, which can be considered as negligible.

We start by considering the solvation condition of the Ca^{2+} ion. This has been the subject of numerous experimental and theoretical studies; the consensus is that the coordination number of Ca^{2+} in water is 6 - 8, with Ca-O distances typically 2.4 - 2.5 Å and an exchange rate of *ca*. 10^9 - 10^{11} s^{-1} [38-40]. Our DFT calculations on the aqua ions $[Ca(H_2O)_n]^{2+}$ (n = 5 - 9) indicate that individual water ligands bind very weakly to Ca^{2+}, with binding energies less than typical hydrogen bond energies (Table 1). This is consistent with the rapid exchange rates noted above. It is worth remembering that the Ca^{2+} ion is nevertheless strongly solvated in water; our method gave calculated solvation energies ΔE_{solv} of -1606 to -1647 kJ/mole for the processes $Ca^{2+}_{(gas)} \rightarrow [Ca(OH_2)_n]^{2+}_{(aq)}$ (n = 5 - 9). Experimental estimates of the free energy of solvation ΔG_{solv} for Ca^{2+} are in the range of -1505 to -1657 kJ/mole [40,41].

We next considered the binding energies of calcium with free phosphate and the various ligands available to BSP. We modelled the individual amino acids as the fragments $CH(O)NHCH(R)C(O)NH_2$, where R is the required sidechain, and the backbone atoms were in the β-sheet conformation. All of the anionic ligands in Table 1 were bidentate in the optimized structures. We also considered a larger model, consisting of a $CH(O)[NHCH(CH_2CH_2CO_2)C(O)]_2NH_2$ fragment, designated as GluGlu in Table 1, as a model for successive Glu residues. For these larger models, several different initial backbone geometries were considered; the lowest energy conformation in each case proved to be the β-turn geometry, incorporating an internal hydrogen bond between the terminal CH(O) and NH_2 groups, although the energies of some of the alternative conformations lacking this hydrogen bond were not much higher (< 10 kJ/mole). The first point of note is that as might be expected for ionic bonding, the strength of Ca^{2+} binding by the various oxyanions correlates with the formal fractional charge on the O-donor atoms. This varies from -3/4 for PO_4^{3-} to -1/3 for sulfotyrosine; a plot of (ΔE + ZPE) *versus* fractional charge for the seven

oxyanions in Table 1 gave a straight line with $R^2 = 0.96$. The strongest interaction is thus between Ca^{2+} and PO_4^{3-}, followed by HPO_4^{2-}. The calculations on the phosphoserine model indicate that binding of Ca^{2+} to this residue is also strong, and in particular clearly stronger than the single carboxylic acids. Therefore, the phosphoserines provide the strongest association between the protein and HA. Individual carboxylate donors show reasonably favourable binding to Ca^{2+}; Asp is slightly inferior to Glu, which can be explained by a weak internal hydrogen bond between the carboxylate and the NH group of the Asp residue in the free ligand, which is lost upon Ca^{2+} binding. The extended GluGlu model shows that a single Ca^{2+} can be favourably coordinated by two adjacent carboxylate groups, however the energy gain from coordination of the second carboxylate to the single Glu-Ca^{2+} moiety (-18.7 kJ/mole including the ZPE) is less than that for capture of a second Ca^{2+} ion (-30.1 kJ/mole including the ZPE). This suggests that provided the local Ca^{2+} ion concentration is sufficiently high, the poly(Glu) regions of BSP preferentially interact with one Ca^{2+} ion per glutamate. This seems to us to be an important point, since it suggests that whereas the phosphoserines will form neutral ion pairs with single Ca^{2+} ions, the glutamates in the poly(Glu) regions of the protein are able to interact with sufficient numbers of Ca^{2+} ions to require free phosphate anions for charge neutralization, so promoting HA nucleation. This is also consistent with experiments that show that the poly(Glu) regions but not the phosphoserines are critical for HA nucleation [10]. Finally, both histidine and sulfotyrosine gave weak binding energies, not much better than the values for water ligands, therefore the interactions of these residues with Ca^{2+} are unlikely to be important for HA nucleation or binding.

Summing up, the DFT and MD calculations suggest that the surface patch of glutamate residues serves to increase the local concentration of Ca^{2+} and phosphate ions by electrostatic attraction. These loose assemblies of ions rapidly become compacted as the free

ions self-assemble into a more regular arrangement that optimizes the strong electrostatic attractions between Ca^{2+} and phosphate. The highly flexible protein chain facilitates rapid assembly of the inorganic ions, and also becomes intimately associated with the growing mineral, especially by means of the strong interactions between Ca^{2+} and phosphoserine, such that there is a strong association between BSP and the mature HA crystal.

3.5 Analysis of the composition of bone

Some interesting further insights into the biological function of BSP can be gleaned from analysis of the composition of the main components of bone. The overall composition is 65 wt. % mineral, 25 wt. % organic, and 10 wt. % water [42]. Collagen accounts for 90% of the organic component [8], while BSP comprises 8 – 15% [1,43] of the non-collagenous protein. The HA crystals are quite variable in size, but their typical dimensions from TEM have been given as (30-50) × (15-30) × (2-10) nm [42]. Taking the averages of these values, together with the density of 3.021 g/cm^3 for human dental HA [33], we can calculate that 1 g of bone contains 4.0×10^{16} individual crystals. Taking molecular weights for the collagen triple helix and BSP of 288 and 50 kDa respectively, 1 g of bone will also contain 4.7×10^{17} collagen trimers and $(2.4 - 4.5) \times 10^{16}$ molecules of BSP. In other words, the ratio of individual HA crystals to BSP molecules to collagen trimers is in the region of 1:1:12. Allowing due caution for the approximate nature of these calculations, we suggest that a ratio of one HA crystal per BSP molecule is entirely consistent with BSP being the key nucleator of HA crystal growth.

Comparison of the relative sizes of these three components is also instructive. Figure 9 shows our BSP model, together with a crystal of dimensions 40 × 22.5 × 6 nm and the experimentally determined structure of a collagen trimer [44]. Note that the latter consists only of the α-carbon trace, so is a little thinner than the complete molecule. A single BSP

molecule is clearly too small to be able to control the morphology of the crystals; a role that has in any case recently been assigned to citrate [12]. On the other hand, BSP could well anchor each crystal to a single collagen trimer within a fibril.

4. Conclusion

The model of BSP described here is consistent with a scenario in which the protein acts both to nucleate hydroxyapatite crystallization and to anchor the growing crystals to the surrounding collagen fibrils. It has been pointed out that the physical condition of mineral-nucleating proteins is important for their correct function [9,10]; proteins that can initiate crystal nucleation when isolated by immobilisation may poison crystal growth when free in solution. Therefore, we suggest that individual BSP molecules are kept separate by initial deployment to the cell surface, being attached at both the *N*- and *C*-termini. Transfer to collagen by means of the *N*-terminal collagen binding region would be followed by dissociation of the *C*-terminal from the cell (perhaps involving tyrosine sulfation). Once threaded onto the collagen fibril, BSP would then serve to nucleate HA crystallization by electrostatic attractions between the surface acidic patch and calcium ions. Finally, BSP can promote attachment of the mature HA crystal to the fibril, especially by means the strong attraction between its phosphorylated groups and calcium ions within the mineral matrix. We now plan to use our 3D model for BSP as a starting point for more extensive dynamics studies.

Acknowledgements

The Consortium for Functional Glycomics is thanked for online access to the Functional Glycomics Gateway at http://www.functionalglycomics.org/fg/. Northumbria University is thanked for funding.

Appendix A. Supplementary data

A PDB format file containing the complete BSP protein model.

References

[1] A. George, A. Veis, Phosphorylated proteins and control over apatite nucleation, crystal growth, and inhibition. Chem. Rev. 108 (2008) 4670-4693.

[2] J. Zaia, R. Boynton, D. Heinegård, F. Barry, Posttranslational modifications to human bone sialoprotein determined by mass spectrometry. Biochemistry 40 (2001) 12983-12991.

[3] M. Wuttke, S. Müller, D.P. Nitsche, . Paulsson, F.-G. Hanisch, P. Maurer, Structural characterization of human recombinant and bone-derived bone sialoprotein. Functional implications for cell attachment and hydroxyapatite binding. J. Biol. Chem. 276 (2001) 36839–36848.

[4] L.W. Fisher, D.A. Torchia, B. Fohr, M.F. Young, N.S. Fedarko, Flexible structures of SIBLING proteins, bone Sialoprotein, and Osteopontin. Biochem. Biophys. Res. Comm. 280 (2001) 460–465.

[5] J.T. Stubbs, K.P. Mintz, E.D. Eanes, D.A. Torchia, L.W. Fisher, Characterization of native and recombinant bone sialoprotein: delineation of the mineral-binding and cell adhesion domains and structural analysis of the RGD domain. J. Bone Mineral Res. 12 (1997) 1210-1222.

[6] K. Bryson, L.J. McGuffin, R.L. Marsden, J.J. Ward, J.S. Sodhi, D.T. Jones, Protein structure prediction servers at University College London. Nucl. Acids Res. 33 (Web Server issue) (2005) W36-38.

[7] C.E. Tye, K.R. Rattray, K.J. Warner, J.A.R. Gordon, J. Sodek, G.K. Hunter, H.A. Goldberg, Delineation of the hydroxyapatite-nucleating domains of bone sialoprotein. J. Biol. Chem. 278 (2003) 7949–7955.

[8] A. Franzén, D. Heinegård, Isolation and characterization of two sialoproteins present only in bone calcified matrix. Biochem. J. 232 (1985) 715-724.

[9] G.K. Hunter, H.A. Goldberg, Nucleation of hydroxyapatite by bone sialoprotein. Proc. Natl. Acad. Sci. USA 90 (1993) 8562-8565.

[10] G.K. Hunter, H.A. Goldberg, Modulation of crystal formation by bone phosphoproteins: role of glutamic acid-rich sequences in the nucleation of hydroxyapatite by bone sialoprotein. Biochem. J. 302 (1994) 175-179.

[11] C.E. Tye, G.K. Hunter, H.A. Goldberg, Identification of the type I collagen-binding domain of bone sialoprotein and characterization of the mechanism of interaction. J. Biol. Chem. 280 (2005) 13487-13492.

[12] Y.Y. Hu, A. Rawal, K. Schmidt-Rohr, Strongly bound citrate stabilizes the apatite nanocrystals in bone. Proc. Natl. Acad. Sci. USA 107 (2010) 22425-22429.

[13] N.S. Fedarko, A. Jain, A. Karadag, M.R. Van Eman, L.W. Fisher, Elevated serum bone sialoprotein and osteopontin in colon, breast, prostate, and lung cancer. Clinical Cancer Res. 7 (2001) 4060–4066.

[14] M. Papotti, T. Kalebic, M. Volante, L. Chiusa, E. Bacillo, S. Cappia, P. Lausi, S. Novello, P. Borasio, G.V. Scagliotti, Bone sialoprotein is predictive of bone

metastases in resectable non–small-cell lung cancer: a retrospective case-control study. J. Clin. Oncol. 24 (2006) 4818-4824.

[15] A. Bellahcène, M. Kroll, F. Liebens, V. Castronovo, Bone sialoprotein expression in primary human breast cancer is associated with bone metastases development. J. Bone Miner. Res. 11 (1996) 665-670.

[16] Hyperchem, release 8.0.8 for Windows, Hypercube Inc., 1995-2009.

[17] Accelrys Discovery Studio, version 1.7, Accelrys Software Inc., 2005-2006.

[18] N. Guex, M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18 (1997) 2714-2723.

[19] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. PuZ, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: the biomolecular simulation program. J. Comp. Chem. 30 (2009) 1545-1615.

[20] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, Jr., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J.M. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C.

Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09W (Version 7.0), Gaussian, Inc., Wallingford CT, 2009.

[21] K. Cossi, G. Scalmani, N. Rega, V. Barone, New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution, J. Chem. Phys. 117 (2002) 43-54.

[22] A.V. Marenich, C.J. Cramer, D.G. Truhlar, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J. Phys. Chem. B 113 (2009) 6378-6396.

[23] R. Raman, S. Raguram, G. Venkataraman, J.C. Paulson, R. Sasisekharan, Glycomics: an integrated systems approach to structure-function relationships of glycans. Nature Methods 2 (2005) 817-824; http://www.functionalglycomics.org/fg/

[24] H.E. Miwa, T.A. Gerken, O. Jamison, L.A. Tabak, Isoform-specific *O*-glycosylation of osteopontin and bone sialoprotein by polypeptide N-acetylgalactosaminyltransferase-1. J. Biol. Chem. 285 (2010) 1208–1219.

[25] K. Julenius, A. Mølgaard, R. Gupta, S. Brunak, Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15 (2005) 153-164; http://www.cbs.dtu.dk/services/NetOGlyc/

[26] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank. Nucleic Acids Res. 28 (2000) 235-242; www.pdb.org

[27] 2K05; C.T. Veldkamp, C. Seibert, F.C. Peterson, N.B. De la Cruz, J.C. Haugner, H. Basnet, T.P. Sakmar, B.F. Volkman, Structural basis of CXCR4 sulfotyrosine recognition by the chemokine SDF-1/CXCL12. Sci. Signal. 1 (2008) ra4-ra4.

[28] G.S. Baht, J. O'Young, A. Borovina, H. Chen, C.E. Tye, M. Karttunen, G.A. Lajoie, G.K. Hunter, H.A. Goldberg, Phosphorylation of Ser(136) is critical for potent bone sialoprotein-mediated nucleation of hydroxyapatite crystals. Biochem. J. 428 (2010) 385-395.

[29] E. Salih, R. Flűckiger, Complete topographical distribution of both the *in vivo* and *in vitro* phosphorylation sites of bone sialoprotein and their biological implications. J. Biol. Chem. 279 (2004) 19808–19815.

[30] F.A. Saad, E. Salih, L. Wunderlich, R. Flűckiger, M.J. Glimcher, Prokaryotic expression of bone sialoprotein and identification of casein kinase II phosphorylation sites. Biochem. Biophys. Res. Commun. 333 (2005) 443–447.

[31] N. Blom, S. Gammeltoft, S. Brunak, Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294 (1999) 1351-1362; http://www.cbs.dtu.dk/services/NetPhos/

[32] 1UG6 ; N.K. Lokanath, I. Shiromizu, M. Miyano, S. Yokoyama, S. Kuramitsu, N. Kunishima, to be published.

[33] R.M. Wilson, J.C. Elliot, S.E.P. Dowker, Rietveld refinement of the crystallographic structure of human dental enamel apatites. Am. Mineral. 84 (1999) 1406–1414.

[34] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites. J. Mol. Biol. 285 (1999) 2177-2198.

[35] Y. Yang, Q. Cui, N. Sahai, How does bone sialoprotein promote the nucleation of hydroxyapatite? A molecular dynamics study using model peptides of different conformations. Langmuir 26 (2010) 9848-9859.

[36] Y. Yang, D. Mkhonto, Q. Cui, N. Sahai, Theoretical study of bone sialoprotein in bone biomineralization. Cells Tissues Organs 194 (2011) 182-187.

[37] 2QVM ; G.M. Besserer, M. Ottolia, D.A. Nicoll, V. Chaptal, D. Cascio, K.D. Philipson, J. Abramson, The second Ca^{2+}-binding domain of the Na^{+} Ca^{2+} exchanger is essential for regulation: crystal structures and mutational analysis. Proc. Natl. Acad. Sci. USA 104 (2007) 18467-18472.

[38] F. Bruni, S. Imberti, R. Mancinelli, M.A. Ricci, Aqueous solutions of divalent chlorides: ions hydration shell and water structure. J. Chem. Phys. 136 (2012) 064520.

[39] H.J. Kulik, N. Marzari, A.A. Correa, D. Prendergast, E. Schwegler, G. Galli, Local effects in the X-ray absorption spectrum of salt water. J. Phys. Chem. B 114 (2010) 9594-9601.

[40] A. Sigel, H. Sigel, R.K.O. Sigel (Eds.) Interplay between metal ions and nucleic acids. Springer, Dordrecht, 2012.

[41] Y. Marcus, Thermodynamics of solvation of ions part 5. Gibbs free energy of hydration at 298.15 K. J. Chem. Soc., Faraday Trans. 87 (1991) 2995-2999.

[42] M.J. Olszta, X. Cheng, S.S. Jee, R. Kumar, Y.-Y. Kim, M.J. Kaufman, E.P. Douglas, L.B. Gower, Bone structure and formation: A new perspective. Mater. Sci. Eng. R 58 (2007) 77-116.

[43] B. Ganss, R.H. Kim, J. Sodek, Bone sialoprotein. Crit. Rev. Oral Biol. Med. 10 (1999) 79-98.

[44] 3HQV ; J.P. Orgel, T.C. Irving, A. Miller, T.J. Wess, Microfibrillar structure of type I collagen *in situ*. Proc. Natl. Acad. Sci. USA 103 (2006) 9001-9005.

Table 1

Calculated bond lengths and binding energies for Ca^{2+} complexes of ligands available in the BSP environment.

complex	Ca-O/N bond lengths, Å	ΔE, kJ/mole[a]	ΔE + ZPE, kJ/mole[a]
$[Ca(PO_4)]^-$	2.279, 2.280	-103.9	-99.6
$[Ca(HPO_4)]$	2.352, 2.352	-77.0	-73.6
[Ca(phosphoserine)]	2.358, 2.366	-72.6	-68.9
$[Ca(MeCO_2)]^+$	2.432, 2.433	-54.9	-50.5
$[Ca(glutamate)]^+$	2.425, 2.430	-53.3	-47.3
$[Ca(aspartate)]^+$	2.435, 2.453	-47.0	-42.0
[Ca(GluGlu)][b]	2.426, 2.435	-54.2	-48.8
[Ca(GluGlu)][c]	2.305, 2.458, 2.463	-76.6	-67.5
$[Ca_2(GluGlu)]^{2+}$	2.385 - 2.489	-56.2[d] (-35.8)	-48.8[d] (-30.1)
$[Ca(histidine)]^{2+}$	2.532	-19.3	-15.9
$[Ca(sulfotyrosine)]^+$	2.519, 2.549	-13.1	-9.9
$[Ca(H_2O)_5]^{2+}$	2.435 - 2.519	-13.7[e]	-6.0[e]
$[Ca(H_2O)_6]^{2+}$	2.461 - 2.538	-11.7[e] (-1.9)	-3.0[e] (+11.6)
$[Ca(H_2O)_7]^{2+}$	2.456 - 2.574	-10.3[e] (-1.9)	-1.2[e] (+9.9)
$[Ca(H_2O)_8]^{2+}$	2.552 - 2.566	-9.1[e] (-0.9)	+0.4[e] (+11.8)
$[Ca(H_2O)_9]^{2+}$	2.533 - 2.645	-7.9[e] (+1.8)	+1.3[e] (+8.5)

[a] Values in parentheses are for addition of a successive Ca^{2+} or water to the species in the row immediately above. [b] Model with Ca^{2+} bound to a single η^2 carboxylate. [c] Model with Ca^{2+} bound to both carboxylates (η^1,η^2 coordination). [d] average value for a single Ca^{2+}. [e] Average value for a single water ligand.

Figure legends

Fig. 1. Sequence alignment for BSP's from man (P21815), mouse (mse, Q61711), rat (Q3HLN4), golden hamster (ham, P70113), cow (Q28862), pig (P31936), dog (XP_851449) and chicken (hen, P79780). The labels on the top line indicate the post-translational modifications in our model; ***P*** = phosphorylated serine, ***G*** = glycosylated asparagine or threonine, ***S*** = sulfated tyrosine. In addition, the two patches of contiguous glutamate residues in the human sequence are indicated by numbers in square brackets, and the RGD sequence common to all the proteins is indicated. Sequences were retrieved from the UniProt database, except for the dog sequence which was retrieved from the NCBI database.

Fig. 2. (a) *N*-glycan and **(b)** *O*-glycan used in the model. NeuAc = *N*-acetyl neuraminic acid; Gal = galactose; GlcNAc = *N*-acetyl-*D*-glucosamine; Man = mannose; Fuc = fucose; GalNAc = *N*-acetyl-*D*-galactosamine.

Fig. 3. Local sequence comparisons to identify the most likely sites of tyrosine sulfation. The numbers in parentheses are the PAM 250 scores (neglecting the highlighted central tyrosine). **(a)** Comparisons of tyrosines 259 and 262 with a known sulfated tyrosine from protein 2K05; **(b)** internal comparisons of tyrosine 297 with 259 and 262; **(c)** internal comparisons of tyrosine 298 with 259 and 262. Comparison of the 2K05 sequence with tyrosines 297 and 298 scored -7 and +3 respectively.

Fig. 4. Picture of the BSP model. The protein chain is rendered as a yellow ribbon; regions of five or more contiguous acidic residues (including phosphoserines) are coloured red. Glycans are rendered in stick mode; *N*-glycans and *O*-glycans are coloured dark blue and light blue respectively. In this projection, the *N*-glycans lie behind the protein globule and

most of the *O*-glycans are in front. Ca^{2+} ions are rendered as green spheres. The sidechains of the RGD motif are shown in magenta, and the sidechains of the two sulfated tyrosines are shown in orange. H atoms are omitted for clarity.

Fig. 5. Solvent surface (2.5 Å probe) of the BSP model after deletion of all the Ca^{2+} ions, coloured according to electrostatic potential (red = negative, blue = positive). **(a)** View of the complete molecule, showing the negatively charged surface patch; **(b)** view of the opposite side of the globule.

Fig. 6. Complex between the BSP model and a crystal of calcium hydroxyapatite of approximate dimensions 48 × 41 × 13 Å. Calcium atoms are coloured green; sugars are coloured blue, and the protein is coloured yellow. H atoms and the *C*-terminal region of the protein (residues 263-301) are omitted for clarity. **(a)** Overall view; the collagen binding region is highlighted in magenta. **(b)** Slab view, showing a section through the protein-crystal interface; Asp and Glu residues are coloured red, and phosphoserine residues are coloured cyan.

Fig. 7. Typical plot of total energy *versus* time for MD simulation of the interactions of the acidic loop with Ca^{2+} and PO_4^{3-} ions. The numbered regions are associated with self-assembly of Ca^{2+}, PO_4^{3-} and phosphoserine groups; event **1** involves phosphoserines 50 and 59 plus two PO_4^{3-} ions; event **2** involves phosphoserines 51 and 58; event **3** involves phosphoserine 59 plus one PO_4^{3-}; and event **4** involves three PO_4^{3-} ions. The average P···P distance for the associating groups is 8.8 ± 0.6 Å immediately before and 5.3 ± 0.8 Å immediately after these events.

Fig. 8. A typical phosphate association event from the molecular dynamics calculations. **(a)** structure immediately before the association; **(b)** structure immediately after the association; **(c)** plots of P···P distances and total energy *versus* time during the association event. The associating Ca^{2+} and PO_4^{3-} ions are highlighted; other Ca^{2+} ions are rendered as small green spheres. The P···P distances between the three PO_4^{3-} groups are 4.1, 10.4 and 13.0 Å before association, reducing to 4.1, 4.8 and 6.5 Å after association. The four highlighted Ca^{2+} ions are located between the PO_4^{3-} ions (the Ca···P distances are 2.8 - 3.7 Å). Note also phosphoserine-58, whose P atom is coloured magenta; the phosphate of this group is also part of the self-assembled calcium phosphate cluster.

Fig. 9. Relative sizes of BSP, a typical hydroxyapatite crystal, and a collagen trimer. The collagen is rendered as CPK atoms scaled × 2, since only the α-carbons are included.

```
                   P  ·····collagen binding·····     PP       PP··[1]···
man FSMKNLHRRVKIEDSEENGVFKYRPRYYLYKHAYPYPHLKRFPVQGSSDSSEENGD-DSSEEEEEEETSNEGENN---EESNEDEDSE--AENTTLS---ATTLGYG  99
mse FSMKNFHRRIKAEDSEENGVFKYRPRYFLYKHAYPYPPLKRFPVQGGSDSSEENGDGDSSEEEGEEETSNEEENN---EDSEGNEDQEAEAENATLSTLSGVTASYG 105
rat FSMKNFHRRIKAEDSEENGVFKYRPRYFLYKHAYPYPPLKRFPVQGGSDSSEENGDGDSSEEEGEEETSNEEENN---EDSEGNEDQEAEAENATLS---GVTASYG 102
ham FSMKNFHRRVKAEDSEDNGVFKYRPRYFLYKHAYPYPSLKRFSVQGGSDSSEENGDGDSSEEEGEEE-TSNEEENN---EESEGNEDEEAEAENTTLS---SVTTSYG 101
cow LSMKNLNRRAKLEDSEENGVFKYRPQYYVYKHGYPYPALKRFAVQSSSDSSEENGNGDSSEEEEEEETSNEEGNNGGNEDSDENEDEESEAENTTLS---TTTLGYG 105
pig FSMKNFHRRAKLEDPEENGVFKYRPRYYLYKHAYPYPPLKRFPVQSSSDSSEENGNGDSSEEEEEEENSNEEENNEENEDSDGNEDEDSEAENITLS---TTTLGYG 105
dog FSMKNLHRRAKLEDSEENGVFKYRPRYYLYKHSYPYPPLKRFPVQSSSDSSEEDGDGDSSEEEEEEETSNEEENNEENANSDENEDE-SDAENSTLS----ATPGYG 103
hen FSVRSWLRRARAGDSEENAVLKSRHRYYLYRYA--YPPLHRY---KGSDSSEEEGDG--SEEEEEGGAPSHAGT--------------QAAGEGLTLG----------  77

      G                          P ··[2]···                G    G
man EDATPGTGYTGLAAIQLPKK--AGDITNKATKEKESDEEEEEEEGNENEESEAEVDENEQGINGTSTNST-EAENGNGSSGGDNGEE--GEEESVTGANAEDTTETG 202
mse AETTPQAQTFELAALQLPKK--AGDAESRAPKVKESDEEEEEEEEENENEEAEVDENELAVNGTSTNST-EVDGGNGSSGGDNGEEAEAEEASVTEAGAEGTTG-G 209
rat VETTADAGKLELAALQLPKK--AGDAEGKAPKMKESDEEEEEEEEE-NENEEAEVDENEQVVNGTSTNST-EVDGGNGPSGGDNGEE--AEEASVTEAGAEGTTAGV 204
ham AETTTGTGNIGLAALQLPKK--AGNAESKAAKKKESDEEEEEEENE---NEEAEVEENEQVTNGTSTNST-EVYGGNGSSGGYNGEE--GEEQSVTEAGVEGTTV-G 200
cow -EITPGTGDIGLAAIWLPRK--AGATGKKATKEDESDEEEEEEEE----NEAEVDDNEQGINGTSSNST-EVDNGHGSSGGDNGEED-GEEESVTEANTEGITVAG 204
pig GDVTPGTASIGLAALQLPKK--AGDIGKKSAKEEESDEDEEEEEENEE---NEAEVDDNEQGTNGTSTNST-EVDSGNGHSGGDNGEE--GDQESVTEA--QGTTVAG 203
dog EEITPGTGYIGLAAIQLPKK--AGDIRHKATKEEESDEEEEDEENEE---NEAEVDENGQGINSTSSNST-EAENGNGSSAGDNGEG---EEESVTEAHSEGTTEAG 202
hen -----DVGPGGDAASAHQDCKGGQKGTRGDSGDEDSDEEEEEEEEEE----EEEVEEQDVSVNGTSTNTTAETPHGNNTVAAEEEEDDDEEEEEEEEEEEAEATTA 176

           GGG        GG       GG            ···cell binding····S···    RGD                        S
man RQGKGTSKTTTSPN-GGFEPTTPPQ-VYRTTSPPFGKTTTVEYEGEYEYT-GA-NEYDNGYEIYESENGEPRGDNYRAYEDEYSYPKGQGYDGYDGQNYY-HHQ 301
mse RELTS-VGTQTAVLLNGFQQTTPPPEAYGTTSPPIRKSSTVEYGGEYEQT-G--NEYNNEYEVYDNENGEPRGDTYRAYEDEYSYYKGHGYEGYEGQNYYY-HQ 308
rat RELTS-YGTTTAVLLNGFQYTTPPPEAYGTTSPPARKSSTVEYGEEYEQT-G--NEYNTAYETYDENNGEPRGDTYRAYEDEYSYYKGHGYEGYEGQDYYY-HQ 303
ham REQIS-DGPTTAVLMNGFQYTTPPPEAYGTTSPPFRKPTTVEYWGEYEQT-GN-NEYNGEYQIYDNENGEPRGDNYRAYEDEYSYYKGRGYEGYDGQDYYY-HQ 300
cow -------ETTTSPN-GGFKPTTPHQEVYGTTPPPFGKITTP---GEYEQT-GT-NEYDNGYEIYESENGDPRGDNYRAYEDEYSYYKGRGYDSYDGQDYYS-HQ 294
pig EQDNGGAKTTTSPN-GGLEPTPPPQDISGTTLPPSGKTTTPEYEGEYEQT-GA-HEYDNGYEIYESENGEPRGDSYRAYEDEYSYYKGRSYNSYGGHDYY---- 300
dog KQNNGGSKTTLSPD-GGFEPTTPPPELYGTTTRPSGEATPNGYEEEYEQT-GT-NEYDNGYEVYESENGEPRGDNYRAYEDEYSYYKGHSYDSYDGQDYYYHHQ 303
hen AATTAQDEVTTLGD---------EQRSEVTTAGE---------QWEYEVTVGARGDEGPTESSYGDQEEPARGDSYRAYEDEYGYYKGHGYDMY-GQDYYY-NQ 260
```

Figure 1

```
NeuAc(α2-3)Gal(β1-4)GlcNac(β1-2)Man(α1-6)                                  NeuAc(α2-6)
                                         \                                            \
NeuAc(α2-3)Gal(β1-4)GlcNAc(β1-4)         Man(β1-4)GlcNAc(β1-4)                        GalNAc
                                 \       /                   \                        /
                                 Man(α1-3)                   GlcNAc     NeuAc(α2-3)Gal(β1-3)
                                 /                           /
NeuAc(α2-3)Gal(β1-4)GlcNAc(β1-2)                   Fuc(α1-6)
```

(a) **(b)**

Figure 2

```
BSP Y259: YDNGYEIYE (-13)
2K05:     GISIYTSDN
BSP Y262: GYEIYESEN (+16)
```

(a)

```
BSP Y259: YDNGYEIYE (-6)
BSP Y297: DGQNYYHHQ
BSP Y262: GYEIYESEN (-7)
```

(b)

```
BSP Y259: YDNGYEIY (-11)
BSP Y298: GQNYYHHQ
BSP Y262: GYEIYESE (+3)
```

(c)

Figure 3

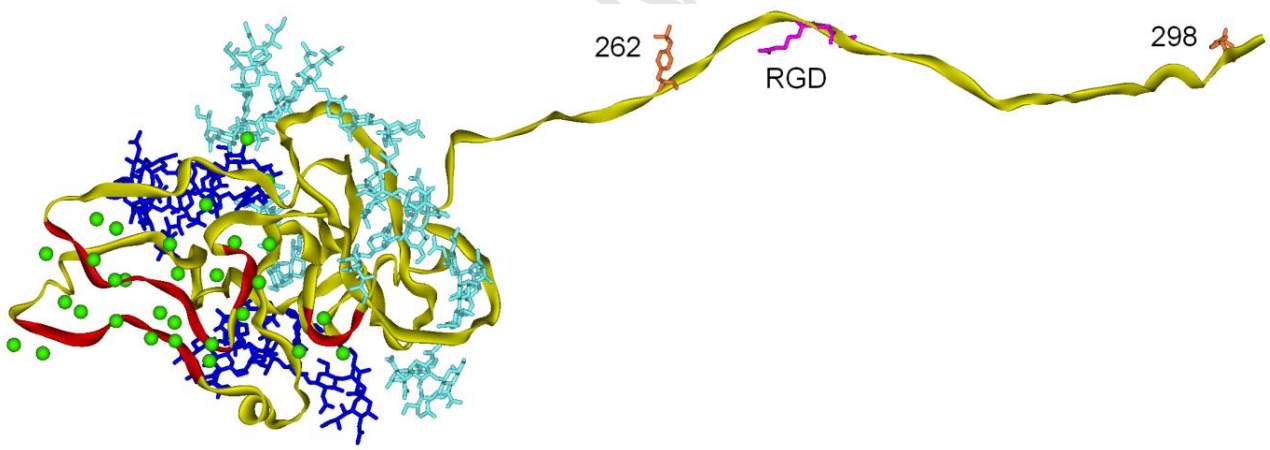


Figure 4

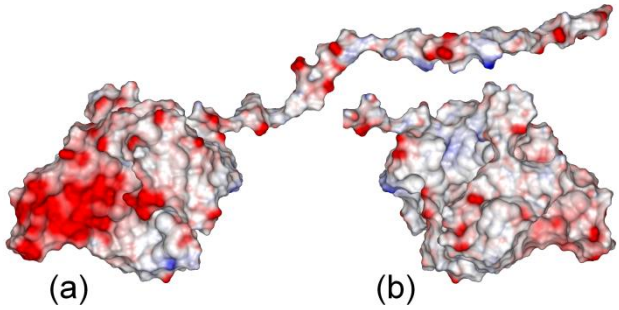


Figure 5

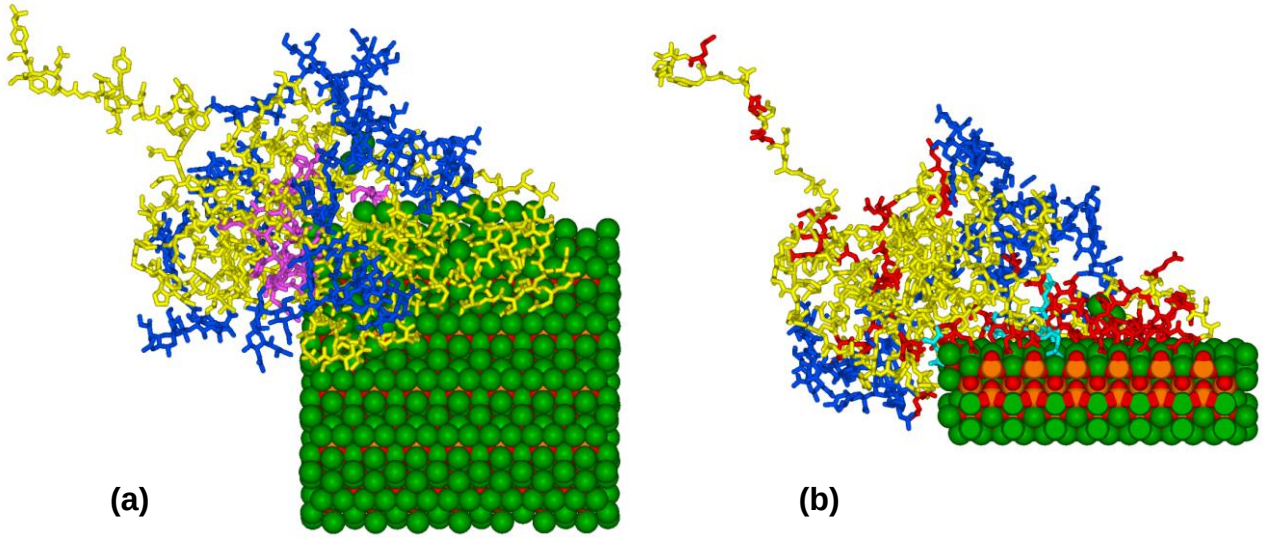


Figure 6

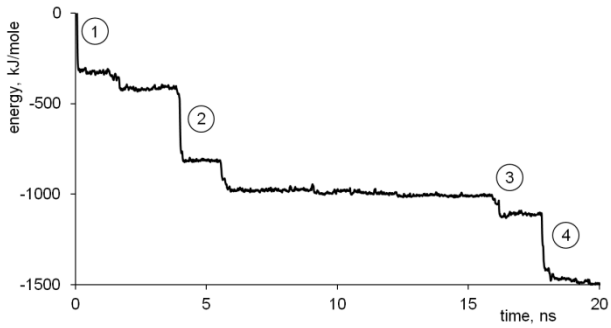


Figure 7

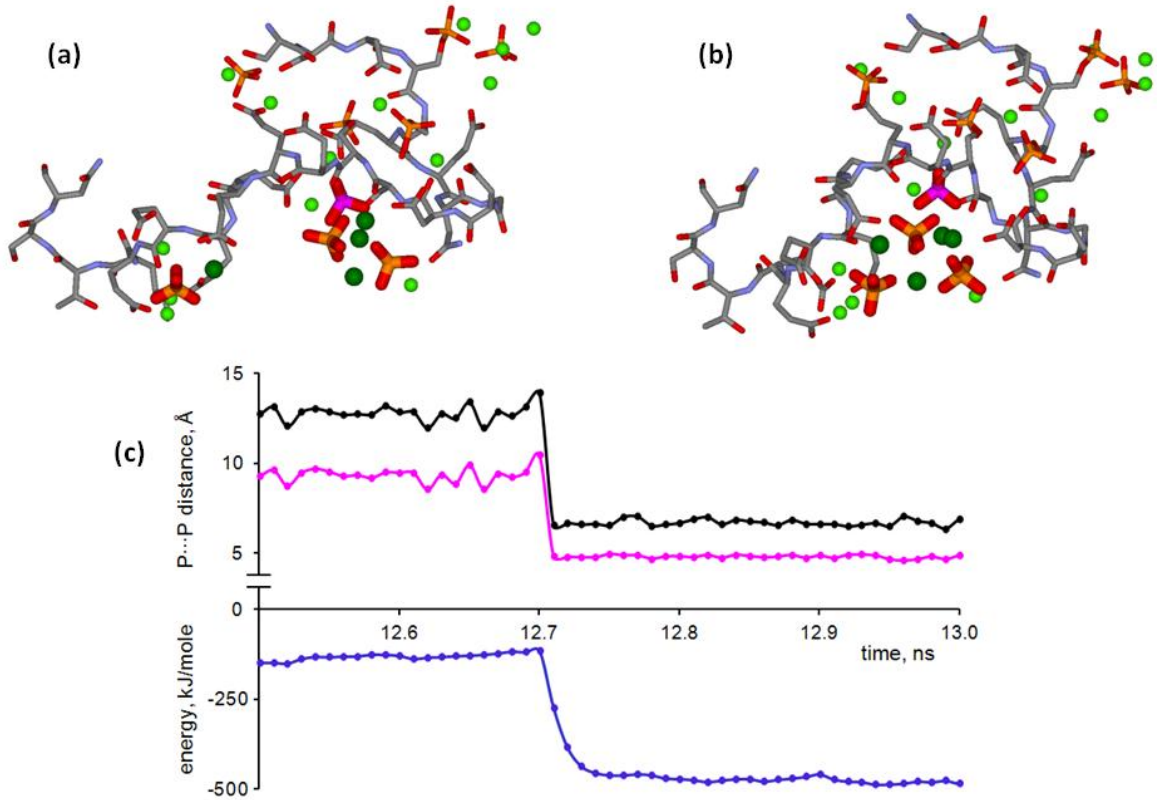


Figure 8

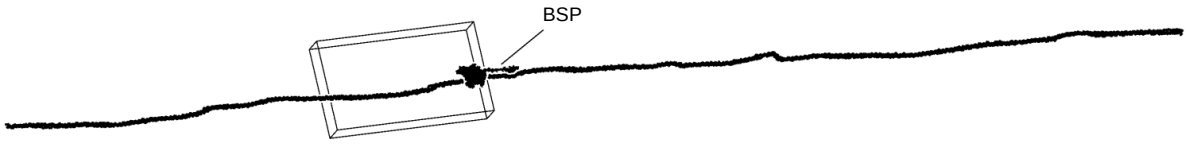


Figure 9