Theoretical model of the three-dimensional structure of a sugar-binding protein from *Pyrococcus horikoshii*: structural analysis and sugar-binding simulations

Anna MARABOTTI*†, Sabato D'AURIA‡§, Mosé ROSSI‡ and Angelo M. FACCHIANO*†¹

*Laboratory of Bioinformatics, Institute of Food Science, Italian National Research Council, Via Roma 52A/C, 83100 Avellino, Italy, †Research Center of Computational and Biotechnological Sciences (CRISCEB), Second University of Naples, Via S. Maria di Costantinopoli, 16, 80138 Naples, Italy, ‡Institute of Protein Biochemistry, Italian National Research Council, Via Castellino, 111, 80131 Naples, Italy, and §Center for Fluorescence Spectroscopy, UMAB, Baltimore, MD, U.S.A.

The three-dimensional structure of a sugar-binding protein from the thermophilic archaea *Pyrococcus horikoshii* has been predicted by a homology modelling procedure and investigated for its stability and its ability to bind different sugars. The model was created by using as templates the three-dimensional structures of a maltodextrin-binding protein from *Pyrococcus furiosus*, a trehalose–maltose-binding protein from *Thermococcus litoralis* and a maltodextrin-binding protein from *Escherichia coli*. According to the suggestions from the CASP (Critical Assessment of Structure Prediction) meetings, the homology modelling strategy was applied by assessing an accurate multiple sequence alignment, based on the high structural conservation in the family of ATP-binding cassette transporters to which all these proteins

belong. The model has been deposited in the Protein Data Bank with the code 1R25. According to the origin of the protein, several characteristics in the organization of the secondary-structure elements and in the distribution of polar and non-polar amino acids are very similar to those of thermophilic proteins, compared with proteins from mesophilic organisms, and are analysed in detail. Finally, a simulation of the binding of several sugars in the binding site of this protein is presented, and interactions with amino acids are highlighted in detail.

Key words: molecular simulation, sugar-binding protein, theoretical model, thermophilic protein, three-dimensional structure.

INTRODUCTION

ATP-binding cassette (ABC) transporters represent an exceptionally large superfamily of transporters present in Nature, and consist of at least 45 recognized families, most of them found only in bacteria [1]. It is known that transporters belonging to this superfamily are multi-domain membrane proteins, responsible for the controlled efflux and influx of substances across cellular membranes, and they function as pumps that couple substrate translocation to ATP hydrolysis [2,3]. Sugar transporters characterized in thermophilic archaea generally belong to the ABC transporters family and are composed of an extracellular binding component, two transmembrane proteins and two nucleotide-binding domains, generally associated as homodimers, responsible for coupling the energy of ATP hydrolysis to conformational changes in the transmembrane domains that allow the intake of ligands. The best-characterized component of this class is maltose transporter from *Escherichia coli*, consisting of the ATPase subunit, two integral cytoplasmic membrane proteins and the extracellular domain responsible for maltose binding [2]. The structure of the last domain, complexed to maltose, maltotriose and maltotetraose, was solved in 1997 at high resolution [4]; subsequently, two other sugar-binding domain structures from thermophilic organisms (maltodextrin-binding protein from *Pyrococcus furiosus* complexed to maltotriose [5] and trehalose–maltose-binding protein from *Thermococcus litoralis* complexed to trehalose [6]) were solved by X-ray crystallography.

Previously, the complete sequence of the genome of a hyperthermophilic archaeobacterium, *Pyrococcus horikoshii*, has been determined, and a list of sequences showing similarities to registered genes with known and unknown functions was published

[7]. One of these sequences (PH1214) was identified as a 'hypothetical ABC transporter extracellular binding protein' and it was included in the SWISS-PROT Protein Sequence database with the ID code YC14.PYRHO. There is experimental evidence that this protein is able to bind to sugar substrates (S. D'Auria, M. Staiano, M. R. Sapio, V. Scognamiglio, A. Marabotti, A. Facchiano, P. Bazzicalupo and M. Rossi, unpublished work). However, the lack of data on its 3D (three-dimensional) structure impairs a complete knowledge of conformational properties and interactions between the protein and its substrates.

In recent years, a massive development of methods to predict the 3D structure of a protein starting from its sequence has increased both their accuracy and their statistical robustness [8] and at present a comparative modelling of the 3D structure of a protein for which homologous templates are known can give very reliable results, especially when the proteins share a high sequence identity or when they belong to the same structural family, with a high conservation of functions and structural features [9]. Indeed, from the high conservation of sequence similarity, a high conservation of structures can also be inferred, but it is well known that in a single family the function and the fold can be retained even if proteins have a low sequence similarity. Thus, in the presence of well-defined structural information, the strategy of comparative modelling can be applied with success by assessing with care the sequence alignment, even if the sequence identity between the target and template proteins is lower than 40% [10].

In this paper, we present results on modelling the structure of Ph-SBP [SBP (sugar-binding protein) from *P. horikoshii*] with a comparative modelling strategy, starting from the 3D structures of proteins belonging to the same functional and structural family

Abbreviations used: ABC, ATP-binding cassette; 3D, three-dimensional; SBP, sugar-binding protein; PDB, Protein Data Bank; Ph-SBP, SBP from *Pyrococcus horikoshii*.

¹ To whom correspondence should be addressed (e-mail angelo.facchiano@isa.cnr.it).

and available in the PDB (Protein Data Bank) [11]. The results show that this protein maintains the fold of the extracellular domain of ABC transporters, with several characteristics typical of thermophilic proteins. The analysis of the 3D model allows us to confirm predictions made on the sequence and to increase our knowledge of substrate binding and interaction.

EXPERIMENTAL

The structure of Ph-SBP was predicted using as templates the 3D crystallographic structures of sugar-binding domains of proteins belonging to the ABC transporters family, chosen with a BLAST [12] search (<http://www.ch.embnet.org/software/aBLAST.html>) within the non-redundant SWISS-PROT Protein Sequence database [13] and PDB database. The selected templates were maltodextrin-binding protein from *P. furiosus* (PDB code: 1ELJ) and trehalose–maltose-binding protein from *T. litoralis* (PDB code: 1EU8). In addition, the maltodextrin-binding protein from *E. coli* was also considered as a template in an early phase of the modelling (see the Results and discussion section), as suggested by a TOPITS prediction-based threading [14,15] performed to find a reliable template by means of fold recognition strategy. Analysis of architectural motifs and topology of the proteins with known crystallographic structure was made according to SCOP [16] and CATH [17] classifications. The PredictProtein server (<http://cubic.bioc.columbia.edu/predictprotein/>) was used for the TOPITS analysis, as well as for searching functional motifs in Ph-SBP with PROSITE [18] and the prediction of structural features in a sequence with PHD [19]. The position of secondary structures on the templates was calculated with the program DSSP [20]. The alignment of the proteins was made with the program CLUSTALW [21] using the following parameters: scoring matrix Blosom, opening and end gap penalty 10 and extending gap penalty 0.05. A few manual refinements were added to account for the position of secondary structures. For the modelling procedure, only the sequence region for which the 3D structure of the templates is available was considered, i.e. the putative 'mature' domains of the proteins [5,6]. As a consequence, we excluded from the model of Ph-SBP the first 41 amino acids as well as the last nine amino acids that did not correspond to any of the template sequences. Moreover, in both crystallographic structures of maltodextrin-binding protein from *P. furiosus* and trehalose–maltose-binding protein from *T. litoralis*, a short N-terminal segment of amino acid residues coming from the mature domain of maltodextrin-binding protein from *E. coli* is present as a result of an artifact of the heterologous expression in *E. coli* [5,6]. Therefore these segments were not considered in the multiple alignment and both templates were managed to delete this artifact before starting the modelling procedure.

The program MODELLER 4 [22] implemented in the Quanta molecular simulation package (version 00.1110, 2000; Accelrys, San Diego, CA, U.S.A.) was used to create all the full-atom protein models by setting 4.0 Å (1 Å = 0.1 nm) as the root-mean-square deviation among the crystal structures of the templates and fully optimized models, with multiple cycles of refinement with conjugate gradient minimization and molecular dynamics with simulated annealing. All other molecular simulations and analyses were performed with the InsightII package (version 2000.1, 2000; Accelrys). The best model among those obtained was chosen by evaluating the stereochemical quality of the models with the program PROCHECK [23]. The ADIT validation tool available online (<http://deposit.pdb.org/validate/>) was also used to check the quality of the structure before deposition in the PDB database. Secondary structures on the final 3D model were calculated again with the program DSSP. Solvent accessibility of the amino acids

was calculated with the program NACCESS [24] by rolling a probe atom of 1.40 Å on the van der Waals surface of the protein models. The helix stability was calculated by the method of Facchiano et al. [25].

Docking of maltotriose, trehalose and maltose into the binding site of Ph-SBP was performed using as starting reference the positions of the oligosaccharides in the crystallographic templates. The sugars were modelled in the binding site of Ph-SBP, and the CVFF force field developed for InsightII was adopted to assign potentials and charges both to the protein and the ligands. The structures of the resulting complexes were optimized to allow a better accommodation of the ligand in the binding groove and to decrease structural conflicts due to steric hindrance. Optimization was performed using 1000 minimization steps in the steepest-descent method, eventually followed by the conjugate-gradient method if the energy gradient reached the threshold value requested by the default options. H bonds were calculated with the tool Measure/H bonds provided by InsightII. The intermolecular energy of the optimized complexes was calculated with the Energy/Intermolecular tool in the docking module of InsightII. As a comparison, intermolecular energy was also calculated for crystallographic complexes. In this case, before energy evaluation, hydrogen atoms were added to the molecules with the Modify/Add Hydrogens tool and force field potentials and charges were calculated.

RESULTS AND DISCUSSION

Search and selection of templates

To create a model of the structure of Ph-SBP, we first performed a BLAST search in protein databases for proteins with similar sequence and known 3D structure. Significant similarities were found with sugar-binding domains of several proteins belonging to the ABC sugar transporter family. Instead, no significant alignment was found with binding proteins belonging to other families, e.g. the chemotactic receptor family. These results suggest that Ph-SBP may belong to the family of ABC sugar transporters. Table 1 lists the top 15 proteins found with the BLAST search in the non-redundant SWISS-PROT Protein Sequence database and their related scores and *E* values, together with their percentage of identity calculated excluding low-complexity regions of the sequence. In addition, proteins belonging to the family of ABC transporters with a crystallographically determined structure of the sugar-binding domain are shown at the bottom of the Table. The first two are maltodextrin-binding protein from *P. furiosus* (PDB code: 1ELJ) [5] and trehalose–maltose-binding protein from *T. litoralis* (PDB code: 1EU8) [6], found with a BLAST search against the PDB database. These two models were taken into account as possible templates. Moreover, to investigate which protein folds are compatible with Ph-SBP, we performed a TOPITS prediction-based threading, and the only protein found with a significant (>3) zscore is the maltodextrin-binding protein from *E. coli*, for which the 3D structure complexed with different sugars was solved [4] (PDB codes: 1ANF, 3MBP and 4MBP). This protein is also present in the last row of Table 1 with its similarity parameters, calculated with BLAST on its sequence. The analysis by CATH and SCOP on the crystallographic structures of trehalose–maltose-binding protein from *T. litoralis* and of maltodextrin-binding proteins from both *P. furiosus* and from *E. coli*, confirmed that they all belong to the periplasmic binding protein-like II superfamily. Their fold is well conserved in the entire superfamily, even if the sequence identity between the proteins is low. On the basis of these results, we can conclude with confidence that our protein may be folded similarly to these proteins. Therefore we used

Table 1 Proteins used for multiple alignment

Proteins found with BLAST search producing significant alignments (E value $< 1 \times 10^{-25}$) with the sequence of putative Ph-SBP from *P. horikoshii* are represented together with proteins with known 3D structure. These sequences were used in a multiple alignment for the comparative modelling procedure. The meaning of Score and E value for BLAST search are reviewed in [12].

Database code	Description	Organism	Score (bits)	E	Sequence identity (%)
TrEMBL: Q9K687	ABC transporter (substrate-binding protein)	<i>Bacillus halodurans</i>	299	5×10^{-80}	40
TrEMBL: Q92KZ7	Putative periplasmic binding protein	<i>Rhizobium meliloti</i>	184	2×10^{-45}	29
TrEMBL: Q98H33	Maltose-binding protein component of ABC sugar transporter	<i>R. loti</i>	184	2×10^{-45}	29
TrEMBL: Q8DG15	ABC-type sugar transport system, periplasmic component	<i>Vibrio vulnificus</i>	181	2×10^{-44}	32
TrEMBL: Q9HZ48	Probable binding protein component of ABC sugar transporter	<i>Pseudomonas aeruginosa</i>	181	2×10^{-44}	31
TrEMBL: Q8FVX5	ABC sugar transporter, periplasmic SBP	<i>Brucella suis</i>	176	8×10^{-43}	28
TrEMBL: Q988P6	ABC transporter, substrate-binding protein	<i>R. loti</i>	174	2×10^{-42}	31
TrEMBL: Q8YCE2	SBP	<i>Brucella melitensis</i>	174	3×10^{-42}	28
TrEMBL: Q88P38	ABC sugar transporter, periplasmic SBP	<i>Ps. putida</i>	163	5×10^{-39}	31
TrEMBL: Q9F7V0	Maltose-binding protein	<i>Burkholderia pseudomallei</i>	158	2×10^{-37}	30
TrEMBL: Q8ZYN8	Multiple SBP	<i>Pyrobaculum aerophilum</i>	157	4×10^{-37}	28
TrEMBL: Q887K0	Glucose ABC transporter, periplasmic glucose-binding protein, putative	<i>Ps. syringae</i>	154	3×10^{-36}	29
TrEMBL: Q87FL6	Putative binding protein component of ABC sugar transporter	<i>Vibrio parahaemolyticus</i>	141	2×10^{-32}	27
TrEMBL: Q97UZ1	Periplasmic SBP	<i>Sulfolobus solfataricus</i>	125	2×10^{-27}	25
TrEMBL: Q8RPQ7	Putative binding protein component of ABC sugar transporter	<i>Ps. putida</i>	115	9×10^{-25}	32
Proteins with known 3D structure					
SWISS-PROT: P58300*	Maltotriose-binding protein precursor (maltodextrin-binding protein, MMBP)	<i>Pyrococcus furiosus</i>	72	2×10^{-11}	23
TrEMBL: Q51923*	Trehalose-maltose-binding protein	<i>T. litoralis</i>	49	1×10^{-04}	20
SWISS-PROT: P02928†	Maltodextrin-binding protein	<i>E. coli</i>	39	0.18	26

* Found with BLAST search in PDB.

† Found with TOPITS search.

these experimental 3D models as templates to predict the 3D structure of Ph-SBP by using the comparative modelling strategy (see below).

A PROSITE analysis on Ph-SBP identified a sequence (ALVIS-VVASGC) in the N-terminal region, typical for cysteine lipid modification. Instead, the PHD predictions identified with high reliability a 'membrane helix' in the segment LLSLLLGVGALVIS-VVASG that is partly superimposed with the PROSITE prediction. Thus this portion is presumably useful for anchoring the protein to the membranes, although the unambiguous identification of this structural feature remains controversial.

Comparative modelling of Ph-SBP

The availability of experimental 3D templates allowed us to create a 3D model of Ph-SBP by using the comparative modelling strategy, taking into account the difficulties existing when the sequence identity comprises between 20 and 40%, a borderline case that has to be treated carefully [9,26–28]. The main problem is related to the consideration that sequences with low identity can be aligned in different solutions with similar scores, and gaps may fall in the middle of secondary-structure elements or represent loops to be modelled without any reference, with a high risk of producing incorrect models. Nevertheless, when proteins used for alignment and modelling belong to the same family in which the structure is well conserved, functional information and overall structure similarity can supply the low-sequence identity and a good multiple sequence alignment can be obtained, suitable for applying the comparative modelling procedure [9]. Moreover, information such as the position of secondary-structure elements can be used to verify the quality of the sequence alignment and to optimize the position of gaps. A similar approach was successfully applied in our laboratory to model the human eIF-5A starting from IF-5A proteins with less than 30% identity [10]. The results of the CASP5 edition, recently discussed in the work of Venclovas

[29], highlighted the crucial role of the sequence alignment and confirmed the validity of our procedure.

Therefore we performed a multiple sequence alignment, which includes Ph-SBP and the proteins listed in Table 1. The alignment of the sequences of Ph-SBP and of templates was then extracted from the multiple alignment, and the positions of the experimentally observed secondary-structure elements of templates and of the predicted secondary structure of Ph-SBP were superimposed to the aligned sequences. We found good agreement between the predicted and the experimental secondary-structure positions, as well as the position of gaps, which fall mainly into loop regions or at the end of experimentally observed structural elements. These results confirmed the good quality of the multiple alignment obtained, and after a few manual refinements to correct gaps falling in the middle of predicted secondary structures, we used this final alignment (Figure 1) as the starting point to predict the 3D structure of Ph-SBP.

By considering the selected templates, two from thermophilic and one from mesophilic organisms, and the low significance of sequence similarity to the target protein for the mesophilic protein, a peculiar procedure was adopted to create the 3D model of Ph-SBP. Two sets of ten structural models were created by using MODELLER, in two distinct sessions, either excluding (SET-MODEL-A) or including (SET-MODEL-B) the structure of maltodextrin-binding protein from *E. coli*. The aim of this procedure was also to find if we could detect some differences between the models created by using only proteins from thermophilic organisms as templates, and those created by using both thermophilic and mesophilic proteins. We then performed a PROCHECK analysis on the stereochemical quality of the 20 models resulting from both modelling procedures and found that the average of the percentages of residues in the most favoured regions, used as a measure of the global quality of the models, was slightly higher for the SET-MODEL-A (85.26%) obtained without using maltodextrin-binding protein from *E. coli*

Figure 1 Final alignment of Ph-SBP sequence with sequences of templates maltodextrin-binding protein from *P. furiosus* (MALE.PYRFU), trehalose-maltose-binding protein from *T. litoralis* (TMBP) and maltose-binding protein from *E. coli* (MALE.ECOLI)

Amino acids in italics were not modelled because of the absence of corresponding regions in the templates. Amino acids in boldface represent artifacts of the heterologous expression of thermophilic proteins in *E. coli* and were thus discarded both in alignment and modelling procedures. Thin lines, secondary structures predicted with PHD for Ph-SBP; thick lines, secondary structures calculated with DSSP on PDB files of templates. Boxes with (—) represent α -helices and boxes with (· · ·) represent β -sheets. Residues in white with black background are in contact with ligands; residues with grey background are involved in thermal stabilization of Ph-SBP.

as a template, compared with the SET-MODEL-B (83.05%). Moreover, the average of percentages of residues in disallowed regions was 1.15% in SET-MODEL-A and 1.43% in SET-MODEL-B. We selected the best model created in each session, i.e. MODEL-A obtained by using only thermophilic templates and MODEL-B obtained by using thermophilic and mesophilic templates (PROCHECK results for these two models are reported in Table 2) and used them as templates to create a third set of ten models. After that, we again performed a PROCHECK analysis on the ten models obtained. The most reliable of them in terms of stereochemical quality, named FINAL-MODEL, shows 90.6% of residues in most favoured regions with no residues in disallowed regions (Table 2), a result expected for crystallographic models with at least 2.0 Å resolution and the *R* factor lower than 20% [23]. By comparing the root-mean-square deviation between the FINAL-MODEL and the two template models, we found a mean value of 0.53 Å from MODEL-A and 1.16 Å from MODEL-B. We hypothesize that some structural features typical of thermophilic proteins and

Table 2 Ramachandran plot statistics

Statistical results on the structure of models obtained with the three modelling sessions. Ramachandran plot statistics were calculated with PROCHECK.

Distribution of residues in the plot	MODEL-A (%)	MODEL-B (%)	FINAL-MODEL (%)
Residues in most favoured regions (A, B, L)	88.8	85.2	90.6
Residues in additional allowed regions (a, b, l, ρ)	9.4	12.0	7.3
Residues in generously allowed regions (-a, -b, -l, ρ)	1.8	2.1	2.1
Residues in disallowed regions	0	0.8	0.0
Total	100.0	100.0	100.0

Figure 2 Schematic representation of the 3D structure of the final model for Ph-SBP

Secondary-structure elements (cylinders, helices; arrows, sheets) were identified by DSSP analysis. N- and C-terminal amino acids are labelled.

hidden in Ph-SBP sequence were recognized by the modelling strategy. A complete validation analysis on this model was also performed with ADIT Validation server at RCSB (Research Collaboratory for Structural Bioinformatics) (results not shown) and confirmed its good quality. The coordinates of FINAL-MODEL were deposited at PDB and accepted with the PDB code 1R25.

Structural features of Ph-SBP

The 3D structure of the model of Ph-SBP is presented in Figure 2. The modelled protein shows a global fold consisting of two separate domains, mainly composed of α/β structures with a 'sandwich' disposition in which a central β -sheet is surrounded by α -helices. The position of secondary-structure elements was identified by a DSSP analysis on the model and found to be essentially in agreement with that predicted by PHD. The sugar-binding site is easily identified in the groove between the two domain structures that are connected with a hinge structure made of two antiparallel β -strands and an α -helix. This structural

Table 3 Stabilizing factors in helices

Analysis of the stabilizing factors in helices of the Ph-SBP model and the three templates. See [25] for a full description of the factors analysed.

	Ph-SBP (<i>P. horikoshii</i>)	1ELJ (<i>P. furiosus</i>)	1EU8 (<i>T. litoralis</i>)	1ANF (<i>E. coli</i>)
Helix stability contributions (kJ/mol)*				
$\Delta G_{\text{H bond}}$	− 496.14	− 460.45	− 531.79	− 450.70
$\Delta G_{\text{intrinsic}}$	722.07	691.45	797.18	679.90
ΔG_{nonH}	− 12.30	− 10.33	− 17.15	− 4.52
ΔG_{dipole}	− 2.05	4.02	− 6.74	− 4.60
ΔG_{SD}	− 62.80	− 65.86	− 84.94	− 36.32
ΔG_{total}	148.78	158.83	156.56	183.76
Putative stability factors				
β -Branched residues in helices†	13	16	23	17
β -Branched residues in helices (%)	13.13	17.02	20.91	18.68
Charged residues difference at the N-terminus‡	− 6	− 10	− 9	− 6
Charged residues difference at the C-terminus‡	9	− 2	5	1
Other details§				
Number of helices	18	16	18	16
Helix content (%)	17.81	25.4	37.7	26.98
Glu–Lys interactions (kJ/mol)	− 11.72	− 13.77	− 18.79	− 14.10
Potential N-capping boxes	0	3	2	1
N-cap free energy (kJ/mol)	− 3.97	− 3.47	− 4.02	− 3.10

* Five terms contribute to the total free energy of helices, which contributes to the overall stability of the protein. The most stabilizing effect corresponds to the lower value of this free energy.

† The presence of β -branched residues (isoleucine, valine and threonine) in α -helices is a destabilizing factor. A lower number of these residues corresponds to a higher stability in α -helices. The absolute number of residues and the percentage are reported.

‡ The value reported is the difference: (positively charged residues)–(negatively charged residues). Owing to the interaction with the helix dipole, stabilization of helices occurs when negatively charged residues are at the N-terminus of the helix, and positively charged residues are at the C-terminus. Therefore stabilizing effect of charged residues is obtained with a negative difference at the N-terminus results and a positive difference at the C-terminus.

§ Other helix stabilizing factors show few differences among the models analysed. Schellman motif, staple motif, α L motif, Phe-Cys/Met interactions are absent. Reported here are the remaining factors and general information about helices of the models.

organization is in perfect agreement with SCOP and CATH analyses performed on crystallographic templates.

The superposition of the modelled protein with the two crystallographic structures 1ELJ and 1EU8 was made, and the overall RMSD for this assembly is 0.42 Å. This result confirms that no major differences are present among the three proteins in terms of global architecture.

Stability of Ph-SBP

The thermophilic origin of our protein prompted us to verify the quality of the model by analysing its stability features in comparison with the templates and in consideration of the known structural features of thermophilic proteins. Some theories on stabilizing factors in thermophilic proteins assign a major importance to the tight packing of hydrophobic residues, and in particular to the formation of isoleucine clusters [30]. By performing a comparative solvent accessibility analysis with NACCESS, we found that only 43.3% of isoleucine residues in our model are totally buried (compared with 80.6% for 1ELJ and 80% for 1EU8). Nevertheless, we noted that it is possible to find in our model at least three clusters of isoleucine residues buried in the interior of Ph-SBP, formed by Ile 5-32-34, 192-196-381-382 and 123-135-140-265. Moreover, three leucine clusters, formed by Leu 3-51-54-56-281, 146-153-163-232-242 and 179-203-358-361-368 are in close contact with isoleucine clusters and determine a tight disposition of hydrophobic residues in the core of the model. Similar clusters are not present in the structure of the mesophilic template [5].

Another interesting consideration was made about buried charged residues. The NACCESS analysis of solvent accessibility indicated that their percentage with respect to totally buried amino acids is 9.4%, a value higher than that for the templates. Glu¹⁴ (one of the most conserved residues) is near the sugar-binding

cavity and can be involved in polar interactions with substrates. A few other charged residues are dispersed into the structure, but a very interesting cluster of buried charged residues formed by Glu¹⁷⁶-Asp²⁷³-Arg¹¹⁹-Asp¹⁶⁵-Arg¹⁶⁶-Lys²²⁴-Asp²⁶¹ is present in the C-terminal domain of the model. This cluster of salt bridges sequestered by the aqueous environment is probably co-responsible, together with other components, for the high thermostability of this protein. This is also in good agreement with results of other studies on factors that cause the thermostability of proteins and which consider the presence of ion pairs as the determinant for this property [31,32].

The number of proline residues in our sequence is 26, which means 6.6% of total residues. This value is similar to that of template thermophilic proteins and higher than that of mesophilic proteins [5]. Moreover, six proline residues are found at the end and five at the beginning of α -helices. This particular location of proline residues seems to be typical of thermophilic proteins, because it can increase their thermostability [33,34]. The region 309–324 is particularly rich in proline residues, where six proline residues are found either in the loop zone or in the middle of two helices which, however, do not seem to be disrupted by the presence of this amino acid.

Finally, we analysed the stability of helical regions by the method of Facchiano et al. [25]. In [25], ten different helix stabilizing factors were reviewed and, although no single factor was found to discriminate specifically between helices of mesophilic and thermophilic proteins, the overall greater stability of helices in thermophilic proteins with respect to mesophilic counterparts was found to be related to a combination of stabilizing factors, the most common being the low content of β -branched residue. In the present study, we applied the same analysis to our model and the three template structures, two from thermophilic and one from mesophilic source. As shown in Table 3, calculation of the energetic contribution of α -helices to the global

energy of the structure indicates that, in the Ph-SBP model, the helices are more stabilized compared with the other three templates. A detailed analysis of the putative stabilizing factors indicates that in the Ph-SBP helices there is a lower content of destabilizing β -branched residues (both as absolute number and in percentage) and a higher number of stabilizing positive charges at the C-terminus of helices, whereas other factors (staple motif, Schellman motif, α L motif, Phe-Cys/Met interactions) do not change significantly among the four models, or are totally absent.

Taken together, all these considerations allow us to consider this modelled structure to be highly compatible with the structural features expected for a thermophilic protein, and confirm the good quality of the prediction. Therefore the model can be used as a reasonable starting point to study the functional properties of this protein.

Simulation of sugar binding

When a model is created starting from a low-sequence homology with its template, it is very difficult to perform reliable simulations on ligand binding and thermodynamic calculations [9]. Nevertheless, since in our case structural and functional information coming from other members of the ABC SBPs were available, in addition to sequence, we could identify with high confidence the substrate-binding cavity on our model. The proteins used as templates lack strict specificity for sugars [4–6] and this is also typical of other proteins belonging to the ABC sugar transporters family [1]. Since the proteins we chose as templates are all complexed with an oligosaccharide (maltose, maltotriose and trehalose), we decided to analyse the interactions of Ph-SBP with those sugars, with the aim of identifying residues interacting with the ligand and possibly to increase our knowledge of its ligand specificity. Therefore we performed a preliminary binding simulation of different sugars. Structures of the oligosaccharides were modelled in the binding site, the resulting complexes were optimized (see the Experimental section for details) and here we present the analysis of the results.

Owing to the low specificity of these binding proteins, Ph-SBP seems to be capable of accommodating all these different sugars without steric conflicts in its cavity. We defined for all complexes a sphere of 5 Å radius starting from the centre of gravity in the sugar and we identified the residues with at least one atom included in this distance. Looking at the amino acids that are present in that distance in all complexes of our protein with the different oligosaccharides, we could define a 'minimal binding site' for Ph-SBP, composed of the following residues: Trp⁸, Trp⁹, Thr¹⁰, Glu¹⁴, Ser⁴⁰, Gly⁴¹, Gly⁴², Gly⁴³, Gly⁴⁴, His⁶⁷, Ala⁶⁸, Gly⁶⁹, His¹¹⁸, Pro¹⁷⁰, Thr¹⁷², Gln¹⁷³, Val¹⁷⁴, Phe¹⁷⁵, Trp²²⁶, Met²⁴³, Trp²⁴⁶, Leu²⁷⁴, Val²⁷⁵ and Ile²⁷⁶. It is interesting to note that these residues are well aligned to amino acids known to interact with ligands in templates (see Figure 1), further confirming the high quality of the starting alignment and of the model obtained.

Looking at this list, it is possible to note that the binding site of Ph-SBP is characterized by a relevant presence of apolar amino acids, and especially of aromatic amino acids, that form favourable interactions with the apolar rings of the sugars. This is in agreement with recurrent features reported in the literature for ABC SBPs [2,3,35,36] as well as with other examples of aromatic residue–sugar interactions observed in the crystallographic structures of the templates used for our modelling [4–6]. Polar amino acids are also present in the binding groove and they are capable of creating a network of hydrogen bonds, together with the backbone of other residues in the binding site. This enables the sugar to remain in place and probably increases the binding affinity. Apparently, the presence of protic amino acids

Figure 3 Simulation of the binding of maltotriose

Residues localized within 5 Å from the centre of the maltotriose molecule are shown. The sugar is black and shown by ball-and-stick representation; amino acids that form H bonds with the sugar are shown in dark grey and labelled and H bonds are drawn with a dotted line. (A) Maltotriose in the binding site of Ph-SBP. (B) Crystallographic complex between maltotriose and maltodextrin-binding protein from *P. furiosus*.

in the binding site of Ph-SBP is limited to Glu¹⁴, a residue that is conserved as an acid residue not only in all template sequences (see Figure 1), but also in almost all sequences of ABC transporters used for the multiple alignment (results not shown), but which apparently in our protein does not interact directly with the ligand. Probably, its negative charge can polarize the hydrogen atoms and this is necessary to create stronger hydrogen bonds with other residues. Interestingly, no positively charged residues seem to be present in the proximity of the binding core of our protein.

Figure 3 shows the comparison between the binding of maltotriose in the ligand cavity of Ph-SBP and the crystallographic complex of maltodextrin-binding protein from *P. furiosus* used as a template. It may be noted that interactions with sugars are similar, and also the number of H bonds is almost the same for both complexes (6 for Ph-SBP and 7 for the crystallographic template complex), but the binding groove of Ph-SBP seems to be composed of more hydrophobic residues than that of the template. In particular, four tryptophan residues (Trp⁸, Trp⁹, Trp²²⁶ and Trp²⁴⁶) are grouped together and create for all complexes a pocket that forms a side of the binding groove. Moreover, the apolar rings of phenylalanine and tyrosine are also present and form another side of the pocket. Finally, other hydrophobic residues enclose the sugar in a highly hydrophobic pocket.

Figure 4 Simulation of the binding of trehalose

Residues localized within 5 Å from the centre of the trehalose molecule are shown. The sugar is black and shown by ball-and-stick representation; amino acids that form H bonds with the sugar are shown in dark grey and labelled, and H bonds are drawn with a dotted line. (A) Trehalose in the binding site of Ph-SBP. (B) Crystallographic complex between trehalose and maltotriose-binding protein from *T. litoralis*.

Figures 4 and 5 show the comparisons between the binding of trehalose and maltose respectively in the cavity of Ph-SBP and the crystallographic complexes of trehalose–maltotriose-binding protein from *T. litoralis* and maltodextrin-binding protein from *E. coli*, used as templates. Trehalose complexed to the ABC protein from *T. litoralis* shows seven residues that interact with multiple H bonds, compared with three residues in Ph-SBP, and the same appears for maltose complexed in *E. coli* maltodextrin-binding protein that is kept in place by more than ten H bonds from six residues, whereas only four H bonds from four different residues are used to complex maltose in Ph-SBP. This is interesting since it clearly shows that the structural features of the binding site of Ph-SBP (from *P. horikoshii*) are more similar to those of the protein from the most similar organism (*P. furiosus*), and this means that the comparative modelling strategy created a model of this protein consistent with its taxonomy.

To understand how favourably these sugars interact with the Ph-SBP, the intermolecular energy of the complexes was computed, and the resulting values were compared with those of crystallographic complexes. Total values together with their

Figure 5 Simulation of the binding of maltose

Residues localized within 5 Å from the centre of the maltose molecule are shown. The sugar is black and shown in ball-and-stick representation; amino acids that form H bonds with the sugar are shown in dark grey and labelled and H bonds are drawn with a dotted line. (A) Maltose in the binding site of Ph-SBP. (B) Crystallographic complex between maltose and maltodextrin-binding protein from *E. coli*.

Table 4 Intermolecular energies for the complexes between the protein and the sugar for templates and our model

These energies were calculated for Ph-SBP on the complexes between protein and sugars after a mild minimization, and for templates on the crystallographic structures of the complexes after addition of hydrogen atoms (see also the Experimental section).

Complex	van der Waals energy (kJ/mol)	Electrostatic energy (kJ/mol)	Total energy (kJ/mol)
Ph-SBP + maltotriose	−190.2	−105.8	−296.0
Ph-SBP + trehalose	−72.4	−51.0	−123.4
Ph-SBP + maltose	−128.0	−80.3	−208.3
1ELJ + maltotriose	−236.8	−43.1	−279.9
1EU8 + trehalose	−182.0	−30.5	−212.5
1ANF + maltose	−154.0	−58.6	−212.6

electrostatic and van der Waals components are shown in Table 4. Their negative values give an indication of the possibility for this protein to create stable complexes with all these different sugars. The energy values for the Ph-SBP complexes are in the same range with respect to template complexes and, hence, we can argue that our docking simulation gives a realistic model

of the complex structure. The intermolecular energy values are not directly related to the binding affinity in terms of ΔG or K_{diss} , since they do not take into account all components that generally are involved in binding (e.g. solvent interactions or entropic factors); however, when they are compared with each other in relative terms, they can give an indication of the most stable complex. Looking at the results obtained in the present study, it is possible to argue that the complex between Ph-SBP and maltotriose is more stable than those with other sugars, and trehalose, whose absolute value is smaller than those of any other complex, is the less preferred sugar for binding. Obviously, these inferences must be experimentally confirmed, but they can give a preliminary indication of putative preferred ligands for this protein. Moreover, it is interesting to note that a great deal of the variance seen in intermolecular energy arises from the difference in van der Waals contribution, whose absolute value varies more significantly among the different complexes. This is also in line with previous analysis, further confirming the importance of non-polar interactions to stabilize and promote the binding of ligands in this category of proteins.

In conclusion, the model of the Ph-SBP described in the present study was found to be highly compatible with expected features of a thermophilic SBP belonging to the ABC transporter family. Stability considerations and characterization of its binding site could open an interesting perspective to modify or increase the affinity of this protein for specific sugars by means of site-directed mutagenesis, in view of its possible biotechnological applications, e.g. in the development of biosensors resistant to the technological processes and capable of titrating the sugar concentration in blood or interstitial fluids.

We thank Mrs Patricia Reynolds for carefully reading the paper before its submission. This project was realized in the framework of CRdC-ATIBB POR UE-Campania Mis 3.16 activities (S.D'A. and M.R.). This work was also partially supported by FIRB grants RBNE0157EH_003 (to A. M. and A. M. F.) and RBAU015B47_006 (to S.D'A. and M.R.) and AA Dgrant (S.D'A.).

REFERENCES

- Saier, Jr, M. H. (2000) Families of transmembrane sugar transport proteins. *Mol. Microbiol.* **35**, 699–710
- Koning, S. M., Albers, S.-V., Konings, W. N. and Driessen, A. J. M. (2002) Sugar transport in (hyper)thermophilic archaea. *Res. Microbiol.* **153**, 61–67
- Horlacher, R., Xavier, K. B., Santos, H., DiRuggiero, J., Kossmann, M. and Boos, W. (1998) Archaeal binding protein-dependent ABC transporter: molecular and biochemical analysis of the trehalose/maltose transport system of the hyperthermophilic archaeon *Thermococcus litoralis*. *J. Bacteriol.* **180**, 680–689
- Quioco, F. A., Spurlino, J. C. and Rodseth, L. E. (1997) Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **5**, 997–1015
- Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. and Waugh, D. S. (2001) Structural basis for oligosaccharide recognition by *Pyrococcus furiosus* maltodextrin-binding protein. *J. Mol. Biol.* **305**, 891–904
- Diez, J., Diederichs, K., Grell, G., Horlacher, R., Boos, W. and Welte, W. (2001) The crystal structure of a liganded trehalose/maltose-binding protein from the hyperthermophilic Archaeon *Thermococcus litoralis* at 1.85 Å. *J. Mol. Biol.* **305**, 905–915
- Kawarabayashi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S. I., Kosugi, H., Hosoyama, A. et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**, 55–76
- Thornton, J. M. (2001) From genome to function. *Science* **292**, 2095–2097
- Tramontano, A. (1998) Homology modeling with low sequence identity. *Methods* **14**, 293–300
- Facchiano, A. M., Stiuso, P., Chiusano, M. L., Caraglia, M., Giuberti, G., Marra, M., Abbruzzese, A. and Colonna, G. (2001) Homology modelling of the human eukaryotic initiation factor 5A (eIF-5A). *Protein Eng.* **14**, 881–890
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. L. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programmes. *Nucleic Acids Res.* **25**, 3389–3402
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I. et al. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in (2003). *Nucleic Acids Res.* **31**, 365–370
- Rost, B. (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 314–321
- Rost, B., Schneider, R. and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480
- Murzin, A. G., Brenner, S. E., Hubbard, T. and Clothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108
- Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219
- Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539
- Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680
- Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) PROCHECK: a programme to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291
- Hubbard, S. J., Campbell, S. F. and Thornton, J. M. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.* **220**, 507–530
- Facchiano, A. M., Colonna, G. and Ragone, R. (1998) Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study. *Protein Eng.* **11**, 753–760
- Rodriguez, R., Chinea, G., Lopez, N., Pons, T. and Vriend, G. (1998) Homology modeling, model and software evaluation: three related resources. *Bioinformatics* **14**, 523–528
- Westhead, D. R. and Thornton, J. M. (1998) Protein structure prediction. *Curr. Opin. Biotechnol.* **9**, 383–389
- Venclovas, C., Zemla, A., Fidelis, K. and Mout, J. (2001) Comparison of prediction performance in successive CASP4 experiments. *Proteins* **5** (Suppl.), 163–170
- Venclovas, C. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins* **53** (Suppl. 6), 380–388
- Danson, M. J. and Hough, D. W. (1998) Structure, function and stability of enzymes from the Archaea. *Trends Microbiol.* **6**, 307–314
- Vogt, G., Woell, S. and Argos, P. (1997) Protein thermal stability, hydrogen bonds and ion pairs. *J. Mol. Biol.* **269**, 631–643
- Kumar, S., Tsai, C. J. and Nussinov, R. (2000) Factors enhancing protein thermostability. *Protein Eng.* **13**, 179–191
- Matthews, B. W., Nicholson, H. and Becktel, W. J. (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6663–6667
- Korkhin, Y., Kalb(Gilboa), A. J., Peretz, M., Bogin, O., Burstein, Y. and Frolow, F. (1998) NADP-dependent bacterial alcohol dehydrogenases: crystal structure, cofactor-binding and cofactor specificity of the ADHs of *Clostridium beijerinckii* and *Thermoanaerobacter brockii*. *J. Mol. Biol.* **278**, 967–981
- Quioco, F. A. (1986) Carbohydrate-binding proteins: tertiary structures and protein–sugar interactions. *Annu. Rev. Biochem.* **55**, 287–315
- Vyas, N. K. (1991) Atomic features of protein–carbohydrate interactions. *Curr. Opin. Struct. Biol.* **1**, 732–740