J Mol Model (2010) 16:1743–1751
DOI 10.1007/s00894-010-0787-4

ORIGINAL PAPER

Protein subunit interfaces: a statistical analysis of hot spots in Sm proteins

Srđan Đ. Stojanović · Božidarka L. Zarić · Snežana D. Zarić

Received: 18 December 2009 / Accepted: 16 June 2010 / Published online: 23 July 2010

Abstract The distinguishing property of Sm protein associations is very high stability. In order to understand this property, we analyzed the interfaces and compared the properties of Sm protein interfaces with those of a test set, the Binding Interface Database (BID). The comparison revealed that the main differences between the interfaces of Sm proteins and those of the BID set are the content of charged residues, the coordination numbers of the residues, knowledge-based pair potentials, and the conservation scores of hot spots. In Sm proteins, the interfaces have more hydrophobic and fewer charged residues than the surfaces, which is also the case for the BID test set and other proteins. However, in the interfaces, the content of charged residues in Sm proteins (26%) is substantially larger than that in the BID set (22%). Hot spots are residues that make up a small fraction of the interfaces, but they contribute most of the binding energy. These residues are critical to protein–protein interactions. Analyses of knowledge-based pair potentials of hot spot and non-hot spot residues in Sm proteins show that they are significantly different; their mean values are 31.5 and 11.3, respectively. In the BID set, this difference is smaller; in this case, the mean values for hot spot and non-hot spot residues are 20.7 and 12.4, respectively. Hence, the pair potentials of hot spots differ significantly for the Sm and BID data sets. In the interfaces of Sm proteins, the amino acids are tightly packed, and the coordination numbers are larger in Sm proteins than in the BID set for both hot spots and non-hot spots. At the same time, the coordination numbers are higher for hot spots; the average coordination number of the hot spot residues in Sm proteins is 7.7, while it is 6.1 for the non-hot spot residues. The difference in the calculated average conservation score for hot spots and non-hot spots in Sm proteins is significantly larger than it is in the BID set. In Sm proteins, the average conservation score for the hot spots is 7.4. Hot spots are surrounded by residues that are moderately conserved (5.9). The average conservation score for the other interface residues is 5.6. The conservation scores in the BID set do not show a significant distinction between hot and non-hot spots: the mean values for hot and non-hot spot residues are 5.5 and 5.2, respectively. These data show that structurally conserved residues and hot spots are significantly correlated in Sm proteins.

Keywords Protein interface · Sm proteins · Hot spots · Statistical analysis

Introduction

One of the fundamental goals of molecular biology is to study protein–protein interactions in an organism, as well as their biochemical and biological functions [1]. Protein–protein interactions are central to most biological processes. A very important problem is to determine the contributions of specific amino acid residues to the specificity and strength of protein interactions. The anatomy, characteristics, and statistics of protein–protein interfaces have been broadly and extensively studied [2, 3]. A protein-binding interface consists of two relatively large, spatially close

S. Đ. Stojanović · B. L. Zarić
ICTM—Department of Chemistry,
University of Belgrade,
Belgrade, Serbia

S. D. Zarić (✉)
Department of Chemistry,
University of Belgrade,
Belgrade, Serbia
e-mail: szaric@chem.bg.ac.rs

protein surfaces with good shape and chemical complementarity. The formation of protein chain interfaces is driven by various natural forces, such as van der Waals contacts and electrostatic interactions, resulting in the removal of water molecules from the binding sites [4, 5]. A detailed, comprehensive knowledge of the principles that govern complex formation can be obtained by studying the crystallographic structures of proteins co-crystallized with various ligands [6], from structural and thermodynamic studies [7, 8] that identify structural epitopes, and through the alanine-scanning mutagenesis of protein–protein interfacial residues [9]. An understanding of protein–protein associations is useful for linking the structures and functions of biomolecular systems, and it allows the energetics of molecular complexes to be characterized [10]. A number of studies have focused on the physical and chemical properties of protein–protein interfaces of complexes in order to determine their unique features [3, 11, 12].

Studies of protein interfaces have revealed that binding energies are not uniformly distributed. Instead, there are certain critical residues called *hot spots* that comprise only a small fraction of the interface but account for the majority of the binding energy [12, 13]. Experimentally, a hot spot can be found by evaluating the change in free energy upon mutating it to an alanine. Hot spot information from experimental studies is only available for a very limited number of complexes, so there is a need for computational methods that identify hot spots in protein interaction sites [3, 14]. The identification of these critical binding residues on proteins enables the rational design of complexes of high affinity and specificity, which are typical of protein–protein complexes.

Sm and Sm-like (LSm) proteins are a widespread protein family which has members that are found in all kingdoms of life. Phylogenetic distributions suggest that Sm proteins were present in the last ancestor common to all present-day life forms, and that this protein family underwent rapid diversification with the advent of eukaryotes [15]. Sm proteins primarily occur as small (~9–29 kDa) standalone proteins that lack other domains [16, 17], and which assemble to form characteristic homomorphic or heteromorphic rings containing six or seven proteins. Members of the family are characterized by the conserved bipartite Sm domain or “Sm fold,” which functions, at least in part, to bind to neighboring Sm proteins within the rings [18, 19]. One highly conserved characteristic of Sm rings is the direct interaction of the central pore of the ring with short uracil-rich stretches of RNA in both prokaryotes [20, 21] and eukaryotes [19, 22]. The Sm family has undergone considerable diversification in eukaryotes, with a variety of heteromorphic Sm rings participating in many RNA-processing pathways and snRNP complexes [16, 19, 23].

Sm/Lsm proteins are able to build defined or undefined highly ordered structures with a ring-like morphology. Such assemblies are highly stable [24]. To disrupt these higher order assemblies, it is necessary in some instances to use a chaotropic agent such as urea at semidenaturing concentrations, or even higher concentrations [24].

We analyzed the interface hot spot residues in subunits of Sm proteins in an attempt to understand the high stability of Sm protein associations. Sm protein complexes are noncovalent assemblies of proteins that fold separately and subsequently oligomerize in order to carry out a particular function. We performed an analysis of the X-ray structures of 15 Sm proteins and analyzed their amino acid compositions and several interface properties. We also compared the properties of Sm protein interfaces with those of a test set, the Binding Interface Database (BID) [25].

Methods

For this study, we used the Protein Data Bank’s (PDB’s) 10 March 2009 list of 56366 structures. The following criteria were employed to assemble the set: (1) no theoretical model structures and no NMR structures were accepted; (2) only crystal structures with a resolution of 3.0 Å or better and a crystallographic R-factor of 25.0% or lower were accepted; (3) crystal structures of proteins containing an Sm-like fold (SCOP classification, version 1.75) [26] without RNA binding were accepted. If not already present, all hydrogen atoms were added and optimized using the program REDUCE [27] with default settings.

Using these criteria, we created a dataset of 15 Sm proteins (presented in Table 1). After the interface dataset had been assembled, several interfaces that contained ligands were rejected. In this way, 213 interfaces were used as the dataset in our analysis.

In order to understand the specificity of Sm protein interfaces, we compared the results with the properties of interfaces in a test set derived from the Binding Interface Database (BID) [25]. The BID contains experimental data on binding free energies. The redundancy in this dataset was removed using the PISCES sequence culling server [40], utilizing a sequence identity of not more than 35%, as in the procedure of Darnell et al. [41]. We only considered residues with known conservation scores and accessibilities, and the set contained 112 residues (54 hot spots and 58 non-hot spots) on 25 monomers.

Interface areas and interface residues were calculated using the Protein Interfaces, Surfaces and Assemblies Service (PISA) at the European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) [42].

Table 1 Dataset of the Sm proteins used for the interface analysis

Protein	Genetic source	Number of subunits	Number of amino-acid residues in single subunit	Resolution (Å)	PDB code	Reference
Hetero-oligomers						
SmD1	Human	1	119 (D1)	2.50	1b34	[28]
SmD2		1	118 (D2)			
SmD3	Human	6	75 (D3)	2.00	1d3b	[28]
SmB		6	91 (B)			
Homo-oligomers						
SmD1	*Pyrococcus abyssi*	28	71	1.90	1h64	[21]
HFQ	*Escherichia coli*	6	74	2.15	1hk9	[29]
AF-Sm1	*Archaeoglobus fulgidus*	28	77	2.50	1i4k	[30]
Sm	*Pyrobaculum aerophilum*	7	81	1.75	1i8f	[31]
Mth649	*Methanobacterium thermautotrophicum*	7	86	1.85	1jbm	[32]
HFQ	*Staphylococcus aureus*	12	77	1.55	1kq1	[20]
SmAP3	*Pyrobaculum aerophilum*	28	130	2.00	1m5q	[33]
Sm	*Methanobacterium thermoautotrophicum*	7	83	1.70	1mgq	[34]
SmF	*Saccharomyces cerevisiae*	7	93	2.80	1n9r	[35]
Sm	*Sulfolobus solfataricus*	14	81	1.68	1th7	[36]
HFQ	*Pseudomonas aeruginosa*	6	82	1.60	1u1s	[37]
LSm5	Cryptosporidium parvum	2	121	2.14	2fwk	[38]
Lsm3	*Saccharomyces cerevisiae*	2	96	2.50	3bw1	[39]

Interface hot spot residues were calculated using the HotPOINT web tool (http://prism.ccbb.ku.edu.tr/hotpoint). Interface residues with observed binding free energies of$\geq$ 2.0 kcal mol^{-1} were considered hot spots, while interface residues with binding free energies of <0.4 kcal mol^{-1} were labeled non-hot spots [3]. The accessible surface areas (ASA) of each residue in the monomer state and in the complexed state in our dataset were calculated using Naccess [43]. These ASAs were then converted into relative accessibilities:

$$\mathrm{relCompASA}_i = \frac{\mathrm{ASA\ in\ Complex}_i}{\mathrm{maxASA}_i} \times 100 \tag{1}$$

$$\mathrm{rel}\Delta\mathrm{ASA}_i = \frac{[\mathrm{ASA\ in\ monomer}_i] - [\mathrm{ASA\ in\ Complex}_i]}{\mathrm{maxASA}_i} \times 100, \tag{2}$$

where *relCompASA$_i$* is the relative ASA in the complexed state of the *i*-th residue, and *relΔASA$_i$* is the relative difference in ASA between the complexed and monomer states of the *i*-th residue; in other words, the change in the ASA of the residue upon complexation. *maxASA$_i$* is the maximum ASA of the residue in the tripeptide state [44].

We used knowledge-based solvent-mediated inter-residue potentials [45] extracted from protein interfaces. The contact potential between two residues *i* and *j* is found using

$$\mathrm{Pair}(i,j) = \left\{ \begin{array}{ll} \text{contact potential of type } (i,j) \text{ if } d(i,j) \leq 7.0 & \\ \text{and } |i-j| \geq 4 & \\ 0 \qquad\qquad \text{otherwise} & \end{array} \right\}, \tag{3}$$

where Pair(*i*, *j*) is the contact potential for residues *i* and *j*, and d(*i*, *j*) is the distance between the centers of the residues [46]. For each residue, we found the neighbors with side-chain center of masses that were closer than the cutoff (7.0 Å). The overall contact potential of residue *i* was defined as the absolute value of the sum of its pair potentials:

$$\mathrm{PP}_i = abs\left(\sum_{j=1}^{n} \mathrm{Pair}(i,j)\right) \quad \mathrm{for}|i-j| \geq 4. \tag{4}$$

A Bayesian method was used to calculate amino acid conservation scores. Homologs were collected from SWISS-PROT; the maximum number of homologs was 50, the number of PSI-BLAST iterations was 1 (PSI-

BLAST E-value=0.001) [47], and conservation scores ranged from 9 (conserved) to 1 (variable).

Results and discussion

The interface properties of the 15 Sm proteins shown in Table 1 were analyzed. We used a selected set of properties to investigate the interface hot spot residues of subunits. The properties used in this study were: (1) size of the subunit interface and number of interface residues; (2) hydrophobic character of the interface; (3) interface residue composition; (4) amino acid preferences in hot spots; (5) distributions of hot spot and non-hot spot features; (6) structurally conserved residues in the interfaces.

Size of the subunit interface

Lo Conte et al. [48] noted that in protein–protein complexes, most interface areas are in the range 1200–2000 Å^2. Interfaces with areas of <1200 Å^2 were considered "small" interfaces, while interfaces with areas of >2000 Å^2 were "large" interfaces.

We estimated the sizes of the interfaces in oligomers of Sm proteins by measuring the area of the protein surface buried in subunit contacts.

Figure 1 is a plot of interface area versus protein size for the 15 oligomers. The data show that the interface area increases with protein chain length. The interface areas range from 400 Å^2 to above 2,000 Å^2. Although the smaller proteins obviously cannot form very large interfaces, the correlation with size is mediocre. Interfaces bury 27% of the subunit surface on average, but this fraction varied from 5 to 31% in our Sm protein dataset. In this dataset, there were also 34±9 residues per interface on average.

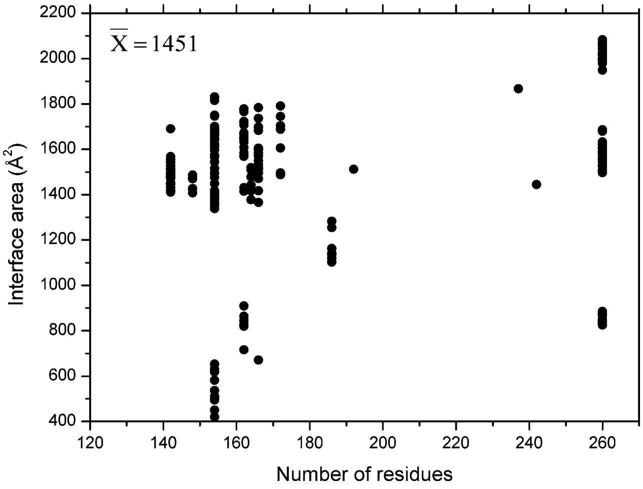

Fig. 1 Interface area versus number of residues (representing the protein chain length) for Sm proteins

Hydrophobic character of the interface

On average, 18% of the protein surface that is buried at the interface belongs to main-chain atoms, and 82% to side-chain atoms. These interfaces are 66% nonpolar and 34% polar, if we count all carbon-containing groups as being nonpolar, and nitrogen-, oxygen-, and sulfur-containing groups as being polar. A histogram of the contribution of polar groups to the subunit interface area for the 213 interfaces is shown in Fig. 2. There is no systematic tendency for the nonpolar/polar area ratio to change with the size of the interface.

We compared our results for the interfaces of Sm proteins with the data for the interfaces in the Binding Interface Database (BID) [25], which was used as a test set. We observed that the interfaces of the Sm proteins are very slightly less polar (34%) than those in the BID test set (35%).

Interface residue composition

Interfaces have been shown to be more hydrophobic than the surface of the protein, but are less hydrophobic than the interior of the protein. In a study conducted on 340 dimer structures containing both homo- and heterodimers, 47% of the interface residues were hydrophobic, 31% were polar and 22% were charged [49].

Figure 3 shows the fractions of hydrophobic, hydrophilic and charged residues that contribute to the solvent-accessible protein surface and to the interfaces in our Sm dataset. The surfaces contained 29% hydrophobic amino acids, 28% hydrophilic, and 43% charged residues. The interfaces had a high fraction of hydrophobic amino acids

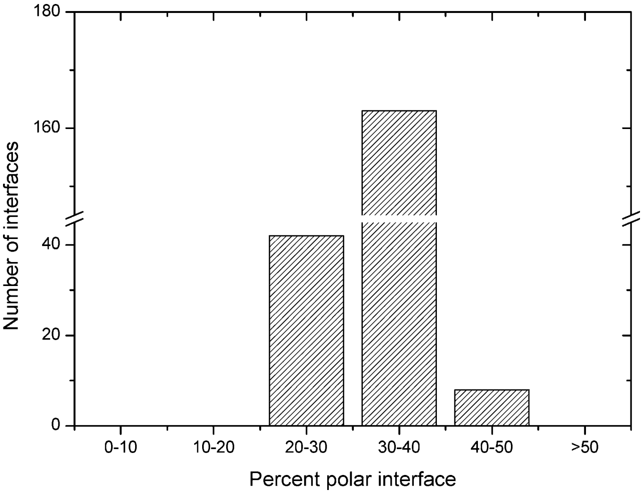

Fig. 2 Histogram of the contributions of polar (N-, O-, and S-containing) groups to the interfaces, expressed as a fraction of the interface area

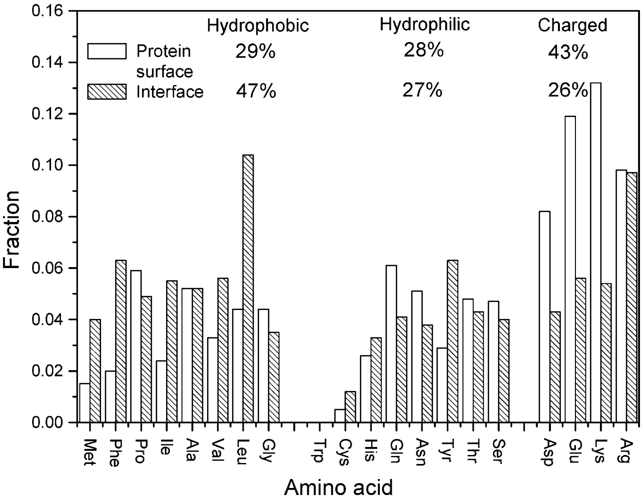

Fig. 3 Amino acid compositions of protein surfaces and interfaces. Hydrophobic, hydrophilic and charged residue fractions are shown

(47%), and smaller fractions of hydrophilic (27%) and charged (26%) residues.

The most abundant residue at interfaces is leucine, which contributes about 11% of the buried surface area. Other aliphatic residues (Leu, Ile, Val, and Met) together contribute 26% of the interface area, whereas they form only 12% of the protein surface. The interfaces are also enriched in aromatic residues: Phe and Tyr are more abundant in interfaces by a factor of 2–3 than on the protein surface. Hydrophilic residues are distributed equally between the surface and the interfaces. In contrast, the latter are depleted by a factor of about two in the charged residues Asp, Glu, and Lys. Together, these three residues contribute 32% of the accessible surface area of the proteins, but only 15% of their interface area.

Remarkably, arginine is not excluded from interfaces, despite its charge. Arg contributes about 10% of the accessible surface area and the surface buried at interfaces. Arg is a significant contributor to interfaces—the second largest after Leu—but it ranks after Lys and Glu on the protein surface. The high abundance of arginine at interfaces has also been seen in other protein–protein complexes [48].

Data on the fractions of hydrophobic, hydrophilic and charged residues in Sm proteins were compared with the corresponding data for the BID test set [25]. On protein surfaces, the amino acid compositions were similar for both sets of proteins. Namely, the surfaces from the BID test set consisted of 31% hydrophobic, 27% hydrophilic, and 42% charged amino acids, while the surfaces of the Sm proteins consisted of 29% hydrophobic, 28% hydrophilic, and 43% charged amino acids. In the interfaces, the hydrophobic amino acid contributions were similar for the Sm and BID sets, while they differed in hydrophilic and charged amino acid contributions. Namely, the interfaces in the BID test set consisted of 48% hydrophobic amino acids, 30% hydrophilic, and 22% charged amino acids. The data show that hydrophilic residues are less common in Sm proteins (27%, Fig. 3), while charged residues are significantly more common in Sm protein interfaces (26%, Fig. 3) than in interfaces from the BID set (22%). This difference is particularly large for positively charged amino acids; Sm proteins contain 0.1% Arg and 0.06% Lys, while the BID set contained 0.06% Arg and 0.03% Lys.

Amino acid preferences in hot spots

The distribution of amino acids in the hot spots in the Sm proteins is strikingly nonrandom (Table 2).

Only two amino acids appear as hot spots with a frequency of more than 10%: arginine (18.7%) and isoleucine (13.6%). However, many amino acids are found in hot spots only very rarely. Less than 3% of the alanine, aspartate, cysteine, glycine, histidine, lysine, serine and tryptophan residues in our database of Sm proteins are in hot spots. A relative abundance of arginine and methionine as well as a relative scarcity of alanine, valine, lysine and serine were noted. The nonrandom composition of the hot spots demonstrates that certain amino acids are preferred in the high-energy interactions between protein chains in interfaces (Table 2). We do not see a preference for a single type of amino acid, such as hydrophobic or charged residues. In fact, the most abundant amino acids in the hot spots analyzed here (Arg, Asn, Ile, Met, Tyr, Phe, Pro) include hydrophobic, polar residues, and one positively charged amino acid. A possible explanation for this is that amino acids that are capable of undergoing multiple types of favorable interactions are preferred as hot spots. Tyrosine, for example, offers a hydrophobic surface and both aromatic π-interactions and the hydrogen bonding ability of its 4-hydroxyl group. Presumably, the ability of tyrosine to hydrogen bond explains why it is more likely to be found in hot spots than phenylalanine. Arginine is also capable of undergoing multiple types of favorable interaction. It has the ability to form a hydrogen bond network with up to five hydrogen bonds, and a salt bridge with its positively charged guanidinium motif. The electron delocalization of the guanidinium π-system has pseudo-aromatic character. Also, arginine has three methylene carbon atoms, which are all hydrophobic in character. It is also interesting to note that asparagine and glutamine are over twice as abundant as aspartate and glutamate in hot spots (Table 2). Curiously, we see that isoleucine, which appears in hot spots with a frequency of 13.6%, is almost twice as common in hot spots as leucine (8.7%), despite the fact that they are isomers with essentially identical chemistry.

Table 2 Amino acid preferences in hot spots

Residue	All amino acids in the Sm dataset		Hot spots		Enrichment in hot spots[a]
	Number	%	Number	%	
Ala	269	2.9	2	0.2	0.1
Arg	770	8.3	163	18.7	2.2
Asn	613	6.6	76	8.7	1.3
Asp	506	5.5	24	2.8	0.5
Cys	26	0.3	5	0.6	2.0
Gln	205	2.2	29	3.3	1.5
Glu	675	7.3	45	5.2	0.7
Gly	541	5.8	0	0	0
His	332	3.4	21	2.4	0.7
Ile	800	8.6	118	13.6	1.6
Leu	1066	11.5	76	8.7	0.8
Lys	643	6.9	11	1.3	0.2
Met	165	1.8	58	6.7	3.7
Phe	444	4.8	53	6.1	1.3
Pro	280	3.0	54	6.2	2.0
Ser	352	3.8	18	2.1	0.5
Thr	233	2.5	34	3.9	1.6
Trp	0	0	0	0	0
Tyr	383	4.1	47	5.4	1.3
Val	967	10.4	37	4.2	0.4

[a] Value of 2 indicates that the residue is twice as frequent in hot spots than in the dataset as a whole

Distribution of features of hot spots and non-hot spots

relCompASA and relΔASA In order to analyze the features of hot spots, we prepared histograms of relative accessible surface area in the complex (relCompASA), relative difference in accessible surface area between the complexed and monomer states (relΔASA), and pair potentials for the hot spot and non-hot spot residues. Further, *t*-tests were performed to determine if the difference between the distributions of hot and non-hot spots is statistically significant for each feature.

In Fig. 4, histograms of the distributions of relCompASA and relΔASA for Sm proteins are presented. The data show that relCompASA values for non-hot spot and hot spot residues are different (Fig. 4a); the mean value for hot spots is 5.1%, while that for non-hot spots is 29.1%. The *P* value for the relCompASA values of hot and non-hot spots is less than <0.05, which implies a significant difference between the hot and non-hot spot distributions. This difference indicates that hot spots are located near the center of the interface and are largely protected from the bulk solvent (corresponding to low relCompASA). The resulting hydrophobic hot spot environment should therefore favor residues capable of both hydrogen bonding and hydrophobic interactions. This is also consistent with previous studies indicating that hot spots are buried [12, 50, 51]. Figure 4b shows the distribution of the change in accessible surface area (ASA) (relΔASA) upon oligomerization. RelΔASA indicates the change in the accessibility of a residue to solvent, and correlates significantly with relCompASA. For relΔASA, the mean values are 50.5% for hot spots and 37.2% for non-hot spots. The *P* value for the relΔASA values of hot and non-hot spots is less than 0.05, which indicates a significant difference.

We compared these features of the Sm protein hot spots with the corresponding data on the interfaces in the BID test set [25]. There is no clear distinction between the set of Sm proteins and the BID test set in this respect. The difference between the two sets is insignificant for both relCompASA and relΔASA (*P* values are 0.16 and 0.21, respectively).

Knowledge-based pair potentials The histograms for knowledge-based pair potentials of residues in the Sm and BID data sets in Fig. 5 show a difference between the interface residues of the Sm and BID data sets. The histogram for Sm proteins shown in Fig. 5a indicates that the knowledge-based pair potentials of hot spot and non-hot spot residues are significantly different. The mean values for hot spots and non-hot spots in the Sm dataset are 31.5 and 11.3, respectively. The knowledge-based pair potentials of the residues are different enough to statistically discriminate hot spots from non-hot spots (*P* value=5.7×10^{-6}) in Sm proteins.

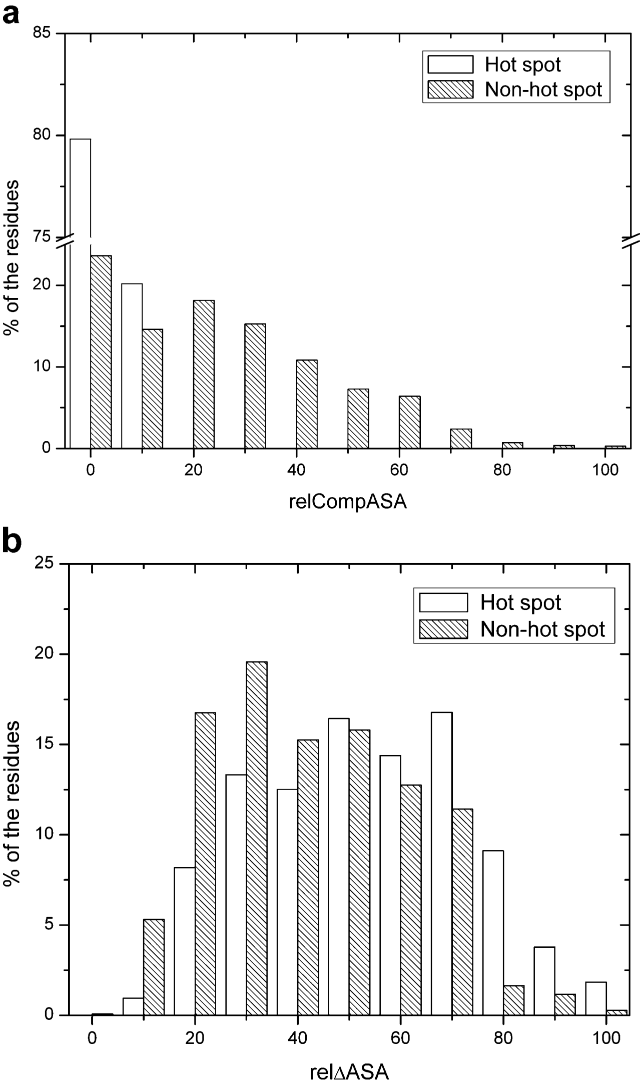

Fig. 4 **a**–**b** Distributions of **a** relCompASA and **b** relΔASA values for hot spot and non-hot spot residues in the Sm protein dataset

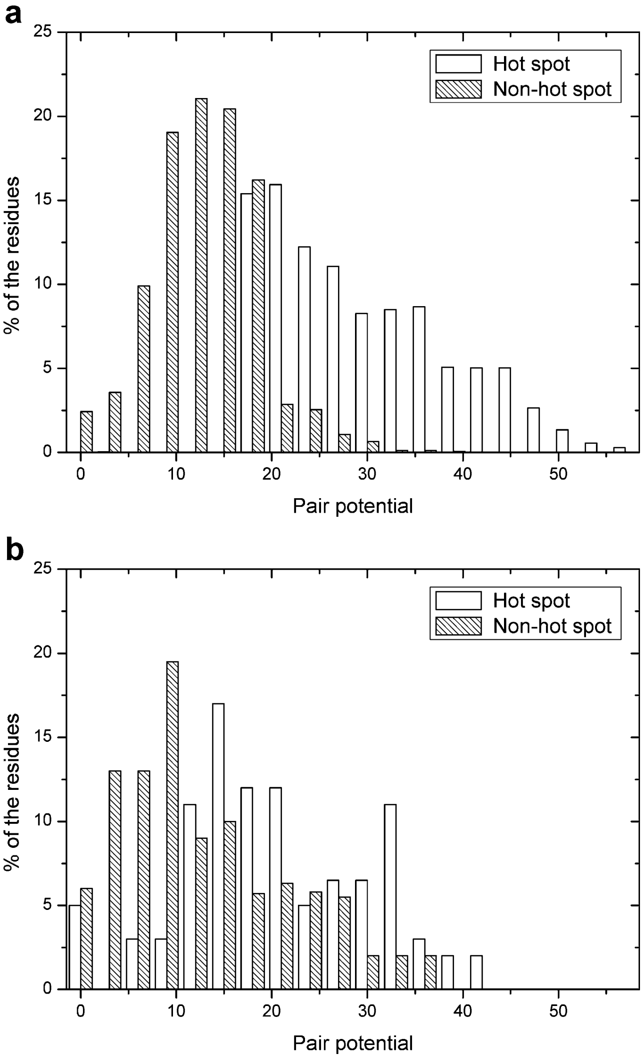

Fig. 5 **a**–**b** Pair potential distributions of hot spot and non-hot spot residues in **a** the dataset of Sm proteins and **b** the BID test set

The histogram in Fig. 5b shows a smaller difference between the pair potentials for hot spot and non-hot spot residues in the BID set than between the pair potentials for hot spot and non-hot spot residues in the Sm proteins (Fig. 5a). The data show that the mean values were similar for non-hot spot residues in the Sm (11.3) and BID (12.4) data sets, while the hot spot residues of Sm proteins have much higher pair potentials (mean value is 31.5) than those in the BID data set (mean value is 20.7). The knowledge-based pair potentials of hot spots in our Sm dataset and the BID test set are significantly different (P value=1.4×10^{-2}). Hence, comparing the pair potentials for interfaces in the Sm and BID data sets shows that the main difference occurs in the pair potentials of hot spots; hot spots of Sm proteins have very high pair potentials that are in accord with the high stabilities of Sm protein oligomers.

Coordination number We analyzed the residue packing densities around the hot spots and around the other interface residues. To study the packing, we investigated the number of nonbonded neighbors (coordination number, *CN*) at 6.5Å around each residue when the residues were represented by their C^a atom positions. Figure 6 shows histograms of coordination numbers around the hot spots and the other interface residues in the Sm proteins and the BID test set. In the Sm proteins (Fig. 6a), the average coordination number for the hot spots was 7.7, while it was 6.1 for the non-hot spot residues. Thus, packing around the hot spots is significantly tighter than in the rest of the interface. These high-density motifs are reminiscent of densely packed protein cores [52, 53]; indeed, the *CN* of the hot spots is very similar to the *CN* of protein cores [54]. This similarity

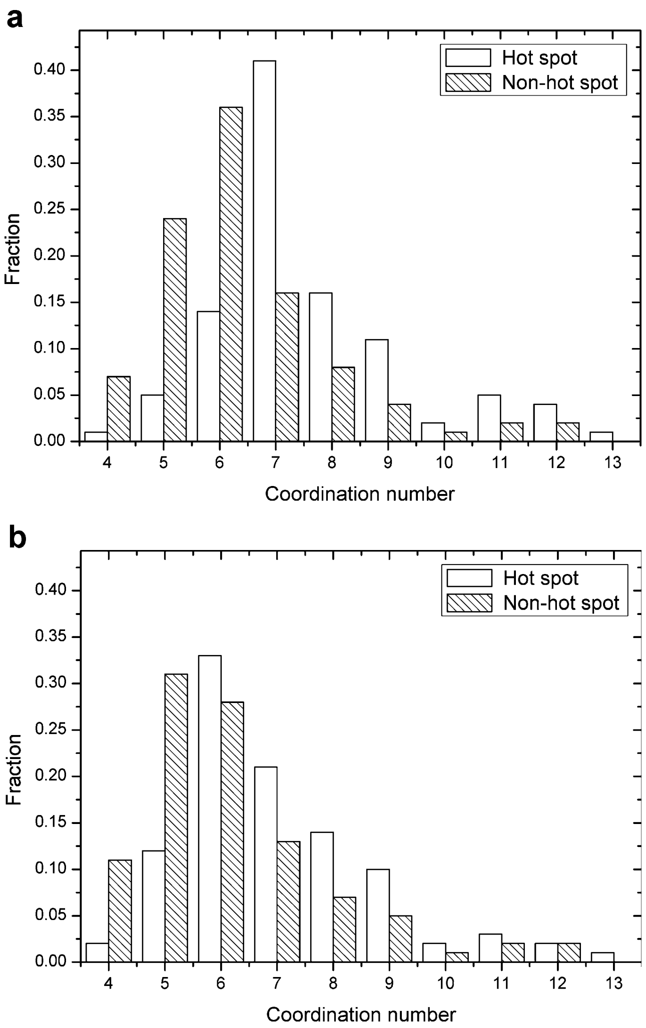

Fig. 6 **a**–**b** Histograms of coordination number for hot spot and other interface residues in **a** Sm proteins and **b** the BID test set

may suggest that binding and folding are similar processes [55].

In the interfaces of the BID data set (Fig. 6b), the average coordination number of the hot spots was 6.3, while it was 5.4 for the other interface residues. These data indicate that the coordination numbers are smaller in the BID set than in the Sm set for both hot spots and non-hot spots. Comparing the residue packing densities around the hot spots of Sm proteins with the data for the interfaces of the BID test set, we found a significant difference between the two sets (P value=2.3×10^{-2}).

Structurally conserved residues in the interfaces

Analyses of conservation patterns of protein–protein interfaces with respect to protein surfaces have shown that the interfaces have been conserved more than the protein surfaces during the course of evolution [56, 57], and it is considered that structurally conserved residues are important in protein stability [14].

Using a Bayesian method for calculating conservation scores for amino acids that are involved in protein interfaces, we found that most of the amino acids are highly conserved: most of them had a conservation score of 9, the highest number on the scale. The calculated average conservation score for the hot spots was 7.4. As expected, the hot spots are surrounded by residues that are moderately conserved (5.9). The average conservation score for the other interface residues was 5.6, which is statistically significantly lower than the value for the hot spots. Further, *t*-tests were performed to determine if the difference between conservation scores for hot and non-hot spots was significant. The difference was indeed statistically significant (P value=2.5×10^{-4}).

The conservation scores in the BID test set did not show a significant distinction between hot and non-hot spots. The mean value for the hot spot residues was 5.5 and that for the non-hot spot residues was 5.2. The difference between the conservation scores of the hot spot amino acids in the Sm proteins and in the BID test set was significant (P value=6.4×10^{-3}).

Conclusions

In order to understand the high stability of Sm protein associations, we analyzed the properties of interfaces and hot spot residues. We performed an analysis of the X-ray structures of 15 Sm-like fold proteins with 213 protein–protein interfaces. We compared the properties of the Sm protein interfaces with the properties of a test set, the Binding Interface Database (BID) [25]. This comparison revealed that the main differences between the interfaces of Sm proteins and those in the BID set were the content of charged residues, the coordination numbers of residues, the knowledge-based pair potentials, and the conservation scores of hot spots.

In Sm proteins, the interfaces have more hydrophobic and fewer charged residues than the surfaces, and this is also the case for the BID test set and other proteins. However, in the interfaces of Sm proteins, the fraction of charged residues is substantially larger than it is in the BID set. Also, in the interfaces of Sm proteins, the amino acids are more tightly packed and the coordination numbers are larger than those in the interfaces of the BID set, for both hot spots and non-hot spots. At the same time, in the Sm protein interfaces, the coordination number is higher for hot spots than for non-hot spots.

The knowledge-based pair potentials of hot spot and non-hot spot residues in Sm proteins are significantly different, while this difference is smaller in the BID set.

Hence, the main difference between the pair potential data for the Sm and that for the BID set is the high pair potentials of the hot spots in Sm proteins.

The difference between the calculated average conservation scores for the hot spots and the non-hot spots in Sm proteins is significantly larger than the corresponding difference in the BID set, and the average conservation score for the hot spots is significantly larger in the Sm proteins than in the BID set. The data show that structurally conserved residues and hot spots are significantly correlated. This demonstrates that hot spots play an important role in the stability of Sm protein associations.

We observed that the hot spots of Sm proteins are located within densely packed regions, they are highly conserved, and they make large energy contributions to interfacial interactions. These properties of hot spots, together with the significant content of charged residues in the interfaces, can explain the high stability of Sm assemblies.

Acknowledgments This work was supported by the Foundation of the Serbian Ministry of Science under grant no. 142037.

References

1. Bordner AJ, Abagyan R (2005) Proteins 60:353–366
2. Jones S, Thornton JM (1996) Proc Natl Acad Sci USA 93:13–20
3. Tuncbag N, Gursoy A, Keskin O (2009) Bioinformatics 25:1513–1520
4. Fernandez A, Scott R (2003) Biophys J 85:1914–1928
5. Privalov PL, Dragan AI, Crane-Robinson C, Breslauer KJ, Remeta DP, Minetti CA (2007) J Mol Biol 365:1–9
6. Erlandsen H, Abola EE, Stevens RC (2000) Curr Opin Struct Biol 10:719–730
7. Davies DR, Cohen GH (1996) Proc Natl Acad Sci USA 93:7–12
8. Brooijmans N, Sharp KA, Kuntz ID (2002) Proteins 48:645–653
9. Pal G, Ultsch MH, Clark KP, Currell B, Kossiakoff AA, Sidhu SS (2005) J Mol Biol 347:489–494
10. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2003) Proteins 53:201–219
11. Ofran Y, Rost B (2003) J Mol Biol 325:377–387
12. Moreira IS, Fernandes PA, Ramos MJ (2007) Proteins 68:803–812
13. Bogan AA, Thorn KS (1998) J Mol Biol 280:1–9
14. DeLano WL (2002) Curr Opin Struct Biol 12:14–20
15. Anantharaman V, Aravind L (2004) BMC Genomics 5:45
16. Anantharaman V, Koonin EV, Aravind L (2002) Nucleic Acids Res 30:1427–1464
17. Pillai RS, Grimmler M, Meister G, Will CL, Luhrmann R, Fischer U, Schumperli D (2003) Genes Dev 17:2321–2333
18. Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R (1995) EMBO J 14:2076–2088
19. Khusial P, Plaag R, Zieve GW (2005) Trends Biochem Sci 30:522–528
20. Schumacher MA, Pearson RF, Moller T, Valentin-Hansen P, Brennan RG (2002) EMBO J 21:3546–3556
21. Thore S, Mayer C, Sauter C, Weeks S, Suck D (2003) J Biol Chem 278:1239–1247
22. Urlaub H, Raker VA, Kostka S, Luhrmann R (2001) EMBO J 20:187–196
23. Wilusz CJ, Wilusz J (2005) Nat Struct Mol Biol 12:1031–1036
24. Zarić B, Chami M, Remigy H, Engel A, Ballmer-Hofer K, Winkler FK, Kambach C (2005) J Biol Chem 280:16066–16075
25. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J (2003) Bioinformatics 19:1453–1454
26. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) J Mol Biol 247:536–540
27. Word JM, Lovell SC (1999) J Mol Biol 285:1735–1747
28. Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K (1999) Cell 96:375–387
29. Sauter C, Basquin J, Suck D (2003) Nucleic Acids Res 31:4091–4098
30. Toro I, Thore S, Mayer C, Basquin J, Seraphin B, Suck D (2001) EMBO J 20:2293–2303
31. Mura C, Cascio D, Sawaya MR, Eisenberg DS (2001) Proc Natl Acad Sci USA 98:5532–5537
32. Mura C, Kozhukhovsky A, Gingery M, Phillips M, Eisenberg D (2003) Protein Sci 12:832–847
33. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D (2003) Proc Natl Acad Sci USA 100:4539–4544
34. Collins BM, Harrop SJ, Kornfeld GD, Dawes IW, Curmi PM, Mabbutt BC (2001) J Mol Biol 309:915–923
35. Collins BM, Cubeddu L, Naidoo N, Harrop SJ, Kornfeld GD, Dawes IW, Curmi PM, Mabbutt BC (2003) J Biol Chem 278:17291–17298
36. Kilic T, Thore S, Suck D (2005) Proteins 61:689–693
37. Nikulin A, Stolboushkina E, Perederina A, Vassilieva I, Blaesi U, Moll I, Kachalova G, Yokoyama S, Vassylyev D, Garber M, Nikonov S (2005) Acta Crystallogr D 61:141–146
38. Vedadi M, Lew J, Artz J, Amani M, Zhao Y, Dong A, Wasney GA, Gao M, Hills T, Brokx S, Qiu W, Sharma S, Diassiti A, Alam Z, Melone M, Mulichak A, Wernimont A, Bray J, Loppnau P, Plotnikova O, Newberry K, Sundararajan E, Houston S, Walker J, Tempel W, Bochkarev A, Kozieradzki I, Edwards A, Arrowsmith C, Roos D, Kain K, Hui R (2007) Mol Biochem Parasitol 151:100–110
39. Naidoo N, Harrop SJ, Sobti M, Haynes PA, Szymczyna BR, Williamson JR, Curmi PM, Mabbutt BC (2008) J Mol Biol 377:1357–1371
40. Wang G, Dunbrack RL Jr (2003) Bioinformatics 19:1589–1591
41. Darnell SJ, Page D, Mitchell JC (2007) Proteins 68:813–823
42. Krissinel E, Henrick K (2007) J Mol Biol 372:774–797
43. Hubbard SJ, Thornton JM (1993) Naccess. Department of Biochemistry and Molecular Biology, University College, London
44. Miller S, Lesk AM, Janin J, Chothia C (1987) Nature 328:834–836
45. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL (1998) Protein Sci 7:2578–2586
46. Bahar I, Jernigan RL (1996) Fold Des 1:357–370
47. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben Tal N (2005) Nucleic Acids Res 33:W299–W302
48. Lo Conte L, Chothia C, Janin J (1999) J Mol Biol 285:2177–2198
49. Lu H, Lu L, Skolnick J (2003) Biophys J 84:1895–1901
50. Li X, Keskin O, Ma B, Nussinov R, Liang J (2004) J Mol Biol 344:781–795
51. Keskin O, Ma B, Nussinov R (2005) J Mol Biol 345:1281–1294
52. Liang J, Dill KA (2001) Biophys J 81:751–766
53. Mirny L, Shakhnovich E (2001) J Mol Biol 308:123–129
54. Miyazawa S, Jernigan RL (1996) J Mol Biol 256:623–644
55. Tsai CJ, Xu D, Nussinov R (1998) Fold Des 3:R71–R80
56. Lichtarge O, Bourne HR, Cohen FE (1996) J Mol Biol 257:342–358
57. Valdar WS, Thornton JM (2001) Proteins 42:108–124