Theor Chem Acc (1999) 101:27–32
DOI 10.1007/s00214980m123

Theoretical
Chemistry Accounts

Regular article

Stabilization centers in various proteins*

Zsuzsanna Dosztányi, István Simon

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, P.O. Box 7, H-1518 Budapest, Hungary

Received: 24 April 1998 / Accepted: 17 September 1998 / Published online: 7 December 1998

Abstract. Interactions among residues together with their interactions with the surrounding medium determine the unique structure of globular proteins. An algorithm was recently developed to locate residues participating in cooperative long-range interactions, called stabilization center residues, that are primarily responsible for preventing the decay of the 3D structure. While our statistical analysis showed that interactions of stabilization center residues hardly influence the formation of the various secondary structure elements, the distribution of the stabilization center residues is rather uneven among the secondary structure elements. Here we analyzed the frequency and distribution of the stabilization center residues and their interacting pairs in secondary structure classes to learn about the effect of secondary structure on the formation and properties of stabilization centers and about the types of interactions responsible for stabilization of proteins of various secondary structure classes. It was found that residues from the same secondary structure tend to interact with each other in the stabilization centers of all classes. It is also suggested that the folding-unfolding equilibrium is governed by different principles for class all-α than for the rest of the classes.

Key words: Stabilization center – Long-range interactions – Secondary structure classes

1 Introduction

The native structure of a protein is in fact an ensemble of very similar structures, the number of which is almost negligible compared to the number of possible structures of an unfolded protein. The ΔTS part of the free-energy difference between the folded (native) and the unfolded protein is almost as much as the enthalpy gain during folding, i.e. the sum of the energy of all local and nonlocal interactions in the folded state minus the sum of the energy of those interactions that appear in a random structure. Therefore the free-energy difference between the two states is about 10 kcal/mol [1]. This free-energy difference means an equilibrium constant in the range 10^6–10^9, thus the unfolding rate must be very small relative to the folding rates. Since folding is a complicated process which does require a certain time, the unfolded rate should also be very small in absolute terms. It is in good agreement with H-D exchange experiments: some amide H exchanges have half-lives of several months at room temperature and at neutral pH [2, 3].

The fascinating phenomenon of spontaneous folding directed the attention of many researchers to the folding process and less attention was paid to describing the phenomenon related to the remarkably low rate of spontaneous decay of the native structure. This low rate of decay contributes exactly as much to the equilibrium constant as the high rate of folding. This decay is driven by the thermal fluctuation of the structure held together by interatomic interactions whose energies are usually only a few times higher than the thermal energy per degree of freedom.

An obvious explanation for the low rate of decay is that individually weak interactions might act cooperatively [4]. If there are structural elements in the protein which assembled from sequentially remote parts of the polypeptide chain and these structural elements can be disintegrated only by simultaneous breaking of several interactions this can keep the protein structure intact for an extended period.

Recently, we have demonstrated that these kinds of structural elements do exist: we called them stabilization centers, (SCs) [5]. An algorithm was developed to locate the amino acid residues of these SCs in proteins. The biological significance of these centers was demonstrated not only by their significantly more long-range interatomic interactions compared to the other part of the protein but also by their evolutionary conservative

*Contribution to the proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: István Simon
e-mail: simon@enzim.hu

characters. Methods were also developed to predict the residues which belong to the SCs from the amino acid sequence of a single protein, or from homologous sequences of related proteins. To identify residues of SCs the following definitions were applied.

Two residues were considered to be in long-range interaction if they were separated by at least ten residues in the sequence and at least one of their heavy atom contact distances was less than the sum of the van der Waals radii of the two atoms plus 1.0 Å. Then these residues were considered together with their flanking tetrapeptides on both sides of the sequence. Finally, two residues were considered as elements of an SC if they were in long-range contact and it was possible to select residues from both flanking tetrapeptides of both residues that make at least seven contacts out of the possible nine contacts between these two residues' triplets. A detailed justification of the definition applied can also be found in Ref. [5]. A WWW server is available at http://www.enzim.hu/scpred/scpred.html.

SCs should be considered in protein design and the predictability of their elements can help conformational energy calculation of proteins, which practically cannot be done without incorporating information from databank analyses into the energy calculations [6, 7].

Various proteins need different numbers and different kinds of interresidue interactions to stabilize their structure. This might depend on the number and distribution of disulfide bonds (covalent crosslinks) and other factors characteristic of the structure itself and its surrounding medium. In fact, it was found that both local and nonlocal interresidue interactions have different frequencies in extracellular and intracellular proteins. Since interresidue interactions can be considered as noncovalent crosslinks and as there is a strong correlation between disulfide content and extra- or intracellular location of proteins, it was surprising to find that the presence or absence of disulfides had nothing to do with the different amounts of local and nonlocal interactions. Instead, it was found that the difference was related to the different secondary structure composition of extracellular and intracellular proteins and it was related to the fact that helical structures, more abundant in intracellular proteins, are stabilized mainly by local interactions, while β sheets, more abundant in extracellular proteins, are stabilized mainly by nonlocal interactions [8]. Since residues of SCs are involved mainly in long-range interactions, some relations between SCs and secondary structure elements are expected. Concerning these relations, it was reported that about half of the SC residues belong to extended (β) structures while the other half is distributed among helix, turn and coil structures. Moreover, half of the interactions in SCs connect residues of extended structures and the other half is distributed among the other nine different types of contacts, such as helix-helix, helix-extended, etc [5]. At the same time it was found that secondary structure prediction methods, using only information on small segments of the sequence, work as accurately on residues of SCs, involved in a large number of nonlocal interactions, than any other residue of the protein, indicating that the formation of SCs has little effect (if any) on the formation of secondary structure elements [9]. This reflects the hierarchy of structure organization, since the SC is a typical tertiary structure element.

To learn more about the effect of secondary structure on the formation of SCs, its properties were analyzed by comparing data of SCs of proteins belonging to different secondary structural classes, as all-α, all-β, α/β, $\alpha+\beta$, and "other".

2 Database

We used the database PDB-select release 1997-May-6, with 25% sequence similarity cutoff [10]. An extra filtering was made to leave out structures with resolutions above 2.5 or with refinements above 0.2. Structures with only Cα atoms or model proteins were also omitted. This led to 527 unrelated polypeptide chains containing 140151 residues. The structural class definition was based on the SCOP database [11]. Our database consists of the following classes: class all-α with 77 proteins, class all-β with 98 proteins, class α/β with 141 proteins and class $\alpha+\beta$ containing 91 proteins. A fifth group was made containing small or multidomain proteins, and proteins composed of two or more domains of different classes (120 proteins).

On the study of lengths of sequence segments built up from SC residues, a randomized sample was also used. It represented the average of 1000 random sequences with the same size and the same number of marked residues as the real proteins and the number of its SC residues.

SC residues were identified as described in Section 1. The secondary structure was identified by the DSSP program [12].

3 Results and discussion

Figure 1 shows proteins representing class all-α, myoglobin, and class all-β, IgG light chain, with their SCs marked.

Table 1 shows the secondary structure composition of the whole dataset of the five classes and of the subsets of residues which are involved in long-range interactions, but not SC residues, and of the subset of SC residues.

The absolute or the relative dominance of residues from extended structures in SCs can be seen in all cases. Although the extended structures represent only 21.52% of the whole dataset, their frequency in SCs is 51.91%. Even in the class all-α, where there is only 3.52% extended structure, 11.71% of the SC residues have extended structure origin. In the whole dataset and in all but one of the classes (all-α) more than half of the residues of extended structures also appear in SCs. Helical structure residues are only dominant in the class all-α (51.67%). Helices are about twice as abundant in class β than extended structures in class α but the ratio of their occurrence in SCs is 1:9. In general, only about 11% of helical structure residues can be found in SCs. They represent 14% of all SC residues. The SCs of class α have

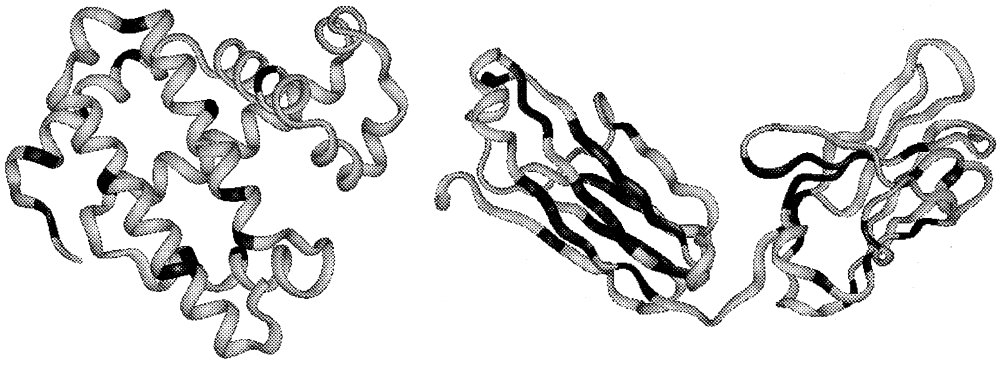

Fig. 1. Proteins representing class all-α, myoglobin (left), and class all-β, IgG light chain (right), with their stabilization center residues marked

Table 1. The secondary structure composition (in percent) for the whole dataset and for the five different classes (all-α, all-β, α/β, $\alpha+\beta$, and "other") in the case of all residues, the subset of residues involved in long-range interactions, but not in stabilization center (LRI), and the subset of SC residues (*SC*). In the "Total" rows the number of residues is given for the subsets

	Whole			All-α			All-β		
	All	LRI	SC	All	LRI	SC	All	LRI	SC
Helix	30.70	35.70	14.06	54.64	58.26	51.67	6.34	7.22	1.30
Extended	21.52	13.76	51.91	3.52	2.51	11.71	39.71	30.09	67.38
Turn	11.97	12.48	5.24	10.74	9.44	7.33	12.29	14.34	4.04
Coil	35.81	38.06	28.80	31.10	29.79	29.29	41.66	48.34	27.27
Total	140151	81166	34322	16078	9895	2127	20761	20761	6923
	α/β			$\alpha+\beta$			Other		
	All	LRI	SC	All	LRI	SC	All	LRI	SC
Helix	36.50	42.44	16.41	27.18	31.02	13.22	28.96	33.89	12.76
Extended	17.58	9.19	48.41	23.49	17.01	52.12	22.76	14.62	53.37
Turn	12.04	12.59	5.66	12.02	12.12	4.99	12.19	12.95	5.27
Coil	33.88	35.78	29.51	37.31	39.85	29.66	36.10	38.55	28.60
Total	44299	26242	10646	18739	10604	4605	40274	23615	10021

about half the relative amount of residues as the rest of the dataset.

Turns appear slightly more often in class all-β than in class all-α (12.29% vs. 10.74%), perhaps because many extended structures are involved in hairpin structures where they are connected by a turn. However, in the SCs their abundance is less in class all-β than in class all-α (4.04% vs. 7.33%). Note that turns in hairpin structures are suggested as good candidates for folding nuclei. It is likely that folding nuclei, stabilized mainly by local interactions, are independent of SCs.

Residues from coil structures account for about 29% of the SC residues. This is slightly less than their frequency in the dataset which is about 36%. This may reflect the dominance of extended structure residues in SCs: it can be seen in the example of class all-β. Concerning the relative abundance of coils in the five classes studied, the relative abundance of coil residues is the highest in class all-β. In fact there are more coil residues than extended structure residues in class all-β (41.66% vs. 39.71%) and this is the class where the relative abundance of coil residues in SCs is slightly smaller than in all other classes, perhaps because more than two-thirds of SC residues come from extended structures in this class.

The dominance of extended structure residues in SCs can also be seen in the three mixed classes: $\alpha/\beta, \alpha+\beta$, and "other".

Considering the ten kinds of interactions, such as helix-helix, helix-extended, etc., the dominance of extended structure residues is even more pronounced. Table 2 shows that almost half of the interactions in SCs (49.30%) connects extended structure elements, while the other half is distributed among the other nine pairs. Note that from the frequency of extended structure residues in the SC, only 26.95% would be expected if the SC residues were paired randomly. However, it is not a unique feature of this kind of pairing, it also fits the general pattern, i.e. residues from the same secondary structure tend to interact with each other: there are 5.77% helix-helix contacts instead of 1.98%, the value for random pairing; there are 0.51% turn-turn contacts instead of 0.27% and there are 13.91% coil-coil contacts instead of 8.29% .

The composition and distribution of SCs differ in the five classes studied. In class all-α, helix-helix contacts are 30.7% in SCs, which is almost the same as the expected value for random pairing (26.70%). For extended structure pairs in this class this value is 12.13% instead of 1.37% and the apparently rare helix-helix contacts of

Table 2. The percentage of the observed number of links between the various secondary structure elements of the whole dataset and of the five different classes (all-α, all-β, α/β, $\alpha+\beta$ and "other") in the case of those long-range interactions which are not members of SCs (LRI) and in the case of SC residues (*SC*)

	LRI (221118 links)					SC (28074 links)				
		Helix	Extended	Turn	Coil		Helix	Extended	Turn	Coil
Whole	Helix	12.20				Helix	5.77			
	Extended	10.35	17.04			Extended	1.34	49.30		
	Turn	4.58	3.45	1.09		Turn	1.83	1.26	0.51	
	Coil	14.60	16.38	7.07	13.24	Coil	6.03	15.69	4.36	13.91
	LRI (19845 links)					SC (1443 links)				
All-α	Helix	41.18				Helix	30.70			
	Extended	2.48	1.63			Extended	1.87	12.13		
	Turn	8.21	0.61	0.91		Turn	6.17	0.62	0.55	
	Coil	26.38	2.85	5.85	9.90	Coil	21.28	4.99	5.20	16.49
	LRI (35337 links)					SC (5993 links)				
All-β	Helix	0.52				Helix	0.27			
	Extended	3.44	33.62			Extended	0.43	61.19		
	Turn	0.90	5.68	1.04		Turn	0.25	1.72	0.15	
	Coil	2.73	28.51	7.57	15.98	Coil	0.77	17.80	4.12	13.30
	LRI (71992 links)					SC (8729 links)				
α/β	Helix	12.38				Helix	6.36			
	Extended	14.67	12.78			Extended	1.49	46.49		
	Turn	5.39	2.67	1.17		Turn	2.10	0.93	1.02	
	Coil	17.95	13.53	6.97	12.49	Coil	7.38	16.04	4.39	13.82
	LRI (28640 links)					SC (3717 links)				
$\alpha+\beta$	Helix	8.45				Helix	4.74			
	Extended	13.57	17.56			Extended	2.50	48.53		
	Turn	4.31	3.47	1.04		Turn	1.83	1.16	0.19	
	Coil	13.75	16.59	6.87	14.40	Coil	7.08	14.53	4.95	14.50
	LRI (65304 links)					SC (8192 links)				
Other	Helix	11.19				Helix	5.25			
	Extended	10.32	17.21			Extended	1.21	50.49		
	Turn	4.68	3.95	1.11		Turn	1.94	1.43	0.38	
	Coil	14.11	16.98	7.38	13.08	Coil	5.30	16.17	4.10	13.73

class all-β (0.52%) are in fact 30 times more frequent than is expected for random pairing (0.017%). For extended-extended contacts the increase is not too dramatic: 61.19% instead of 45.40%, since the expected value is already rather high, while helix-helix contact is almost negligible at 0.27%, which is close to the expected value of 0.16% from random pairing. In the SCs of both classes all-α and all-β, helix-extended contacts are rare: 1.87 and 0.43% instead of 6.05 and 0.86%, the values for random pairing.

For the mixed classes, $\alpha/\beta, \alpha+\beta$ and "other", there are a large number of helix-extended contacts between non-SC residues: for the class α/β among long-range interactions between non-SC residues helix-extended contacts are the most abundant. However, between SCs they fall below the expected value (random pairing): for class α/β it is 1.49% instead of 7.94% , while for class $\alpha+\beta$ it is 2.50% instead of 6.89%.

Table 3 shows the amino acid compositions of the whole dataset, of the five classes and their SCs. In the whole dataset almost all kinds of residues are either significantly preferred in SCs or are significantly not preferred there. There are only three residues, H, M and R for which the difference in their frequencies in SCs and in the whole dataset is smaller than three standard deviation units. The most hydrophobic residues prefer to be in SCs while polar and charged residues prefer to remain outside SCs. Note that R is charged but it has a large hydrophobic part of its side chain and H has a pK close to neutral so its charge depends on its environment.

The distribution of residues in SCs also shows that properties of residues characteristic from the viewpoint of protein structure formation are not always the same as their physicochemical properties. In good agreement with one of our previous findings, for example, alanine behaves quite differently to all the other hydrophobic residues, while methionine does not show significant hydrophobic character, etc. [13].

Considering the amino acid composition of SCs relative to the composition of the corresponding protein classes, they show similar patterns for each class, however, SC residues show different distributions in the primary structures of proteins. Figure 2 shows the percentage of residues of the SC subset in sequence segments of various length. As a general rule, valid for all classes, there are significantly fewer single residues and dipeptides in real SCs than in the "randomized sample" (see Sect. 2) while from pentapeptide and from longer segments there are many more in the real SCs than in the "randomized sample".

Table 3. The percentage of amino acids in the whole dataset and for the five different classes (all-α, all-β, α/β, $\alpha+\beta$ and "other") in the case of all residues and in the case of SCs, and the observed composition differences between the SC residues and the whole dataset, in deviation units

	Amino acid composition								
	Whole			All-α			All-β		
	All	SC	$\frac{(SC-All)}{\delta_{SC}}$	All	SC	$\frac{(SC-All)}{\delta_{SC}}$	All	SC	$\frac{(SC-All)}{\delta_{SC}}$
A	8.44	7.16	−9.8	8.91	7.33	−2.8	6.69	5.98	−2.8
C	1.46	1.97	10.2	1.41	2.07	2.9	1.67	1.88	1.8
D	6.09	4.24	−16.8	6.06	3.62	−5.3	6.02	4.29	−7.9
E	6.02	4.06	−17.8	6.79	4.98	−3.8	5.01	3.81	−5.7
F	4.06	5.24	13.1	4.58	6.35	4.1	4.09	5.50	7.8
G	7.98	6.91	−8.9	6.44	5.08	−2.8	8.85	7.09	−6.1
H	2.28	2.35	1.0	2.39	1.97	−1.3	2.11	2.02	−0.6
I	5.43	7.87	24.4	4.89	7.15	5.4	5.07	7.12	9.3
K	5.81	4.43	−12.5	5.99	4.89	−2.3	6.00	5.00	−4.3
L	8.40	10.28	14.5	9.67	13.92	7.5	7.04	8.72	6.7
M	2.14	2.28	2.3	2.34	2.49	0.5	1.61	1.95	2.8
N	4.68	3.77	−9.1	4.51	3.24	−3.2	5.47	4.51	−4.6
P	4.75	3.76	−11.0	4.85	4.14	−1.7	5.08	3.65	−6.8
Q	3.73	2.86	−9.7	4.09	2.91	−3.0	3.48	3.11	−2.1
R	4.69	4.65	−0.4	5.14	5.59	1.0	4.05	4.42	1.9
S	6.01	5.15	−8.6	5.94	4.42	−3.0	7.43	6.40	−4.1
T	5.76	6.16	3.6	5.31	5.55	0.5	7.31	7.37	0.2
V	6.90	10.23	30.3	5.59	7.52	4.3	7.43	10.15	10.5
W	1.59	1.96	6.2	1.66	2.59	3.7	1.80	2.28	3.7
Y	3.77	4.67	10.0	3.43	4.18	2.1	3.81	4.74	4.9
Total	140151	34322		16078	2127		20761	6923	
	α/β			$\alpha+\beta$			Other		
	All	SC	$\frac{(SC-All)}{\delta_{SC}}$	All	SC	$\frac{(SC-All)}{\delta_{SC}}$	All	SC	$\frac{(SC-All)}{\delta_{SC}}$
A	9.24	8.08	−4.8	8.18	7.27	−2.8	8.39	6.92	−6.1
C	1.15	1.49	3.8	1.57	1.93	2.4	1.67	2.52	7.7
D	6.21	4.32	−9.4	6.09	4.36	−5.8	6.02	4.19	−9.1
E	5.92	3.70	−10.6	6.18	4.21	−6.6	6.27	4.33	−9.2
F	4.00	4.98	6.1	3.86	5.36	6.3	4.01	5.05	6.1
G	7.85	7.53	−1.4	8.14	6.67	−4.3	8.20	6.64	−6.5
H	2.16	2.30	1.2	2.72	3.04	1.6	2.24	2.38	1.2
I	5.77	8.46	14.2	5.28	7.99	9.0	5.53	7.84	12.2
K	5.84	3.98	−9.8	5.42	4.36	−3.5	5.79	4.46	−6.6
L	8.76	11.12	9.8	8.33	9.88	4.4	8.21	9.87	7.2
M	2.21	2.37	1.3	2.01	1.91	−0.6	2.34	2.53	1.5
N	4.67	3.72	−5.6	4.64	3.67	−3.6	4.37	3.47	−5.3
P	4.68	3.82	−4.8	4.67	3.80	−3.3	4.65	3.65	−5.9
Q	3.65	2.52	−7.1	4.00	2.84	−4.6	3.66	3.06	−4.0
R	4.51	4.26	−1.4	4.96	4.47	−1.8	4.93	5.12	1.1
S	5.67	4.69	−4.9	5.97	4.89	−3.6	5.71	5.07	−3.2
T	5.54	5.91	1.9	5.43	5.99	2.0	5.54	5.81	1.3
V	7.01	10.65	17.3	6.63	10.01	10.9	7.15	10.51	14.6
W	1.39	1.60	2.1	1.75	2.00	1.6	1.61	1.98	3.4
Y	3.77	4.50	4.6	4.15	5.32	4.7	3.71	4.60	5.6
Total	44299	10646		18739	4605		40274	10021	

Considering the characters of the plot, the curve of class all-α is rather close to the "randomized sample" showing a hyperbola-like curve, while all the other curves have wide bell shapes.

We can summarize that there are several characters of SC residues valid for all secondary structure classes: the high relative frequency of extended-extended contacts, the rare occurrence of helix-extended and other mixed contacts, the dominance of hydrophobic residues in SCs and the longer sequence segments in SCs. At the same time there are many differences in SC character for the various classes. The most pronounced differences can be seen between class all-α and class all-β, while the other three classes, α/β, $\alpha+\beta$ and "other" have characters much closer to class all-β than to class all-α. Class all-α is the only one where residues from the α helix are the most abundant in SCs and it has fewer SC residues than any other class. Also, this is the only class where helix-helix contacts are dominant in pairing, although it is practically the

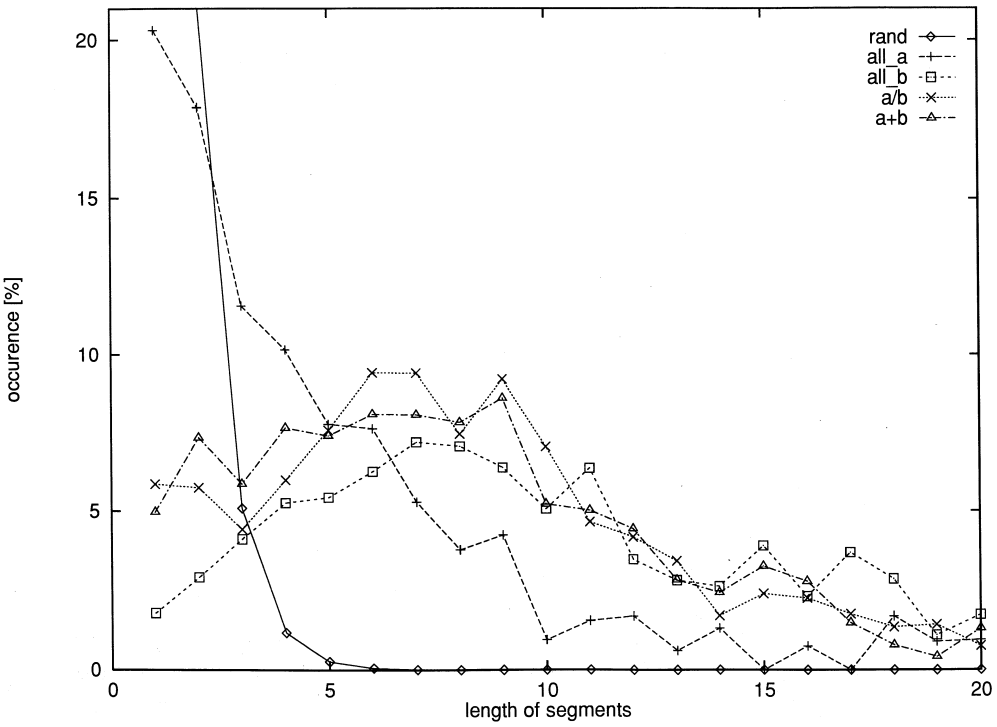

Fig. 2. The percentage of residues in segments with a given length for a randomized sample, and for the five classes (all-α, all-β, α/β, $\alpha+\beta$ and "other")

same as expected for random pairing. The sequence-segment size distribution of class all-α has similar character to the "randomized sample" while all the other classes show rather different distributions.

We can conclude that in all classes residues from the same secondary structure tend to interact with each other. In some aspects, for example, the relative number of SC residues, the SC residue distribution, connections of secondary structure elements etc., class all-α differs from the rest of the dataset. The nature of the uniqueness of class all-α reflects the fact that the helix can be formed and stay stable by itself [14], suggesting that unfolding of this class is governed by a different mechanism than the rest of the dataset.

Acknowledgements. We thank to Á. Simon, G. Szirtes and G.E. Tusnády for their critical comments on the manuscript. This work was supported by research grants OTKA T017652, T022050 and F019008, by the Hungarian–Austrian and the Hungarian–French Intergovernmental S & T Cooperation Programmes. Zs. D. acknowledges support from the Soros Foundation (Hungary).

References

1. Privalov PL, Khechinashvili NN (1974) J Mol Biol 86: 665–684
2. Hvidt A, Nielsen SO (1966) Adv Protein Chem 21: 287–386
3. Simon I, Tüchsen E, Woodward C (1984) Biochemistry 23: 2064–2068
4. Creighton T (1993) Proteins: structures and molecular properties. Freeman, New York
5. Dosztányi Z, Fiser A, Simon I (1997) J Mol Biol 272: 597–612
6. Ortiz AR, Kolinski A, Skolnick J (1998) Proc Natl Acad Sci USA 95: 1020–1025
7. Simon I, Glasser L, Scheraga HA (1991) Proc Nat Acad Sci USA 88: 3661–3665
8. Gugolya Z, Dosztányi Z, Simon I (1997) Proteins Struct Funct Genet 27: 360–366
9. Fiser A, Dosztányi Z, Simon I (1997) Comput Appl Biosci 13: 297–302
10. Hobohm U, Sander C (1994) Protein Sci 3: 522–524
11. Murzin AG, Breuner SE, Hubbard T, Chothia C (1995) J Mol Biol 247: 536–540
12. Kabsch W, Sander C (1983) Biopolymers 22: 2577–2637
13. Tusnády GE, Tusnády G, Simon I (1995) Protein Eng 8: 417–423
14. Chakrabartty A, Baldwin RL (1995) Adv Protein Chem 48: 141–176