Theor Chem Acc (1999) 101:33–40
DOI 10.1007/s002149800m92

Theoretical
Chemistry Accounts

Regular article

Analyzing patterns between regular secondary structures using short structural building blocks defined by a hidden Markov model*

A.C. Camproux[1], **P. Tuffery**[2], **L. Buffat**[3], **C. André**[1], **J.F. Boisvieux**[1], **S. Hazout**[2]

[1] Département de Biomathématiques, CHU Pitié-Salpêtrière, 91 boulevard de l'Hôpital, F-75013 Paris Cedex 13, France
[2] Équipe de Bioinformatique Moléculaire, INSERM U155, Université Paris VII, 2 Place Jussieu, F-75251 Paris Cedex 05, France
[3] Centre de Bioinformatique, INSERM U444, Université Paris VII, 2 Place Jussieu, F-75251 Paris Cedex 05, France

Received: 14 May 1998 / Accepted: 4 August 1998 / Published online: 16 November 1998

Abstract. Hidden Markov models were used to identify recurrent short 3D structural building blocks (SSBBs) describing protein backbones. Polypeptide chains were broken down into successive short segments defined by their inter-alpha-carbon distances. Fitting the model to a database of nonredundant proteins identified 12 distinct SSBBs and described the preferred pathways by which SSBBs were assembled to form the 3D structure of the proteins. Protein backbones were labelled in terms of these SSBBs. The observed SSBB preferences for fragments located between regular secondary structures suggested that they depended more on the following regular structure than on the preceding one. Extraction of repeated series of SSBBs between regular secondary structures showed some structural specificity within different connection types. These results confirm that SSBBs can be used as building blocks for analyzing protein structures, and can yield new information on the structures of the coils flanking secondary structures.

Key words: Protein structure – Pattern classification – Hidden Markov chain – Connecting fragments

1 Introduction

Protein conformation has long been the subject of experimental and computational interest. The database of known protein structures clearly indicates that proteins use recurrent structural motifs at all levels of organization [1]. This recurrence can be seen at the level of secondary structure elements, of local three-dimensional (3D) structures, and of protein domain topology. Fragments of a single class, all with the same particular secondary structure assignment, vary substantially in their 3D structures [2]. Unger and Sussman [3] have pointed out that a classification into 3D building blocks crosses the boundaries between traditional secondary structure assignments. Building blocks, unlike secondary structure elements, have a tertiary significance, because concatenating them in an overlapping manner produces a 3D chain. Hence, rigorous and objective categorization of these structural blocks may lead to a deeper understanding of the modular architecture of proteins.

This paper sets out to identify recurrent short structural 3D building blocks (SSBBs) in a database of polypeptide chains. To do this, polypeptide chains were broken down into a series of protein backbone segments four residues long. Each conformation segment is described by a distance vector between nonsuccessive alpha carbons (C_α). However, the conformations of protein structures are not uniform; they are inherently flexible. We therefore gave our model a stochastic form in order to solve the associated problems of describing SSBBs, the heterogeneity of their corresponding short segments, and their global organization by quantifying their connections. These goals are achieved by using hidden Markov models (HMMs) [4]. HMMs have been successfully used in molecular biology to distinguish coding from noncoding regions of DNA [5], to model protein families and domains [6, 7] and to generate multiple alignments for them [8–10], to predict the secondary structures of proteins from their amino acid sequences [11, 12], and to identify protein folds [13].

First, HMMs were estimated in a representative library of nonredundant known protein structures extracted from the Protein Data Bank (PDB) of Bernstein et al. [14]. We identified 12 different SSBBs. We then divided the protein backbones into series of SSBBs, and all contiguous polypeptide chain fragments which connected two consecutive regular secondary structures were extracted. Lastly, we used an automated procedure

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

Correspondence to: A.C. Camproux
e-mail: acc@biomath.jussieu.fr

to find exact repeated contiguous backbone chain fragments of defined length which formed words made up of identical successions of SSBBs.

2 Methods

2.1 Description of the database

The proteins selected from the PDB met the following criteria: a crystallographic resolution of less than 2.5 Å, and, only a limited sequence homology (25% or less sequence identity) [15]. The information used was the set of atomic coordinates and secondary structure assignments obtained from a prediction consensus [16] that were merged into three categories – helices, strands, and coils. Because the structure of the HMMs was based on local dependence of successive residues in each protein, all noncontiguous protein chains were eliminated, resulting in a training database of 100 proteins for HMMs.

Polypeptide chains were broken down into a succession of protein backbone segments four residues long. The conformation of each segment can be displayed adequately with a four-dimensional distance vector between nonsuccessive C_α, as illustrated in Fig. 1. The vector y_j associated with the j^{th} four-residue segment of the database comprises the three distances ($d_1(j)$, $d_2(j)$, $d_3(j)$) between nonsuccessive C_α and $d_4(j)$, the orientation of the last residue relative to the plane defined by the first three C_α.

2.2 Structure of the HMM

A series of successive four-dimensional distance vectors y_1, y_2,...y_N describing the sequence of N four-residue segments was used as the input data extracted from the database of 100 proteins. Each four-residue segment was assigned to one of a limited number (R) of the SSBBs identified from the polypeptide chains. The interdependence of the conformations of contiguous segments was taken into account by using a Markov chain to model a "hidden" sequence of states that is the succession of underlying SSBBs. HMMs of first order presume that the SSBBs of the polypeptide chains are related through state-transition probabilities, written as $\pi = (\pi_{ii'})_{1\le i,i'\le R}$, common to all the polypeptide chains. We, however, specified an already-started process specific to each protein, i.e., a different probability law $P(X_1 = i)$ that each polypeptide chain would start in the i^{th} SSBBs ($1 \le i \le R$). An associated output function $f_i(y,\theta_i)$, ($1 \le i \le R$) described the variability of the four-residue segments resulting from each state of the Markov chain.

Unknown parameters of the HMM were estimated by finding the best fit to the observed y_1, y_2, y_N [4]. We first set $R = 3$, a simple model, and then considered cases of increasing complexity by increasing R. The Bayesian Information Criteria [17] based on the likelihood, were used to select the number of states incorporated into the HMM. Given a set of y_1, y_2, y_N and the HMM parameter estimated from the database, the most likely underlying sequence of SSBBs was determined using the Viterbi algorithm [18, 19] and used to infer a classification of protein structural building blocks for the proteins in the database.

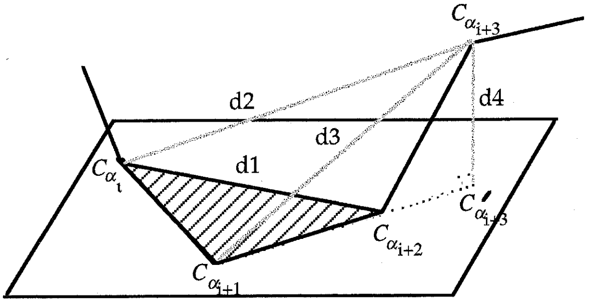

Fig. 1. Four-residue segment as defined by the four distances between nonsuccessive C_α

2.3 Analysis of regular secondary structures connections

Once the SSBBs had been identified, the protein sequences were labeled with this new code. Each SSBB was assigned to the third residue of the corresponding four-residue segment. We focused on contiguous polypeptide chain fragments, X, which connected two consecutive regular secondary structures. Each structural fragment X excluded flanking residues that are regular secondary structures and contained only some isolated residue assigned to be either an α-helix or a β-strand. When the two residues preceding X were assigned to be either an α-helix or a β-strand, fragments X were written as fragments αX or βX. When the two residues following X were likewise assigned to be an α-helix or a β-strand, fragments X were called fragments $X\alpha$ or $X\beta$. This classification resulted in four types of connecting fragments: fragments $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$ and $\beta\beta$.

The distribution of SSBBs into four connection types of different lengths was then explored. The Shannon entropy H_l was computed to find the equivalent number of SSBBs N_{eq} (denoted by eSSBBs) from

$$H_l = -\sum_{1\le i\le R} f_{i,l} \times \ln(f_{i,l}) \enspace .$$

where R is the number of possible SSBBs and $f_{i,l}$ the proportion of SSBB of category i involved for a fixed length l. Then $N_{eq} = e^{H_l}$ It was found out that N_{eq} varied between 1 (only one block involved) and R (all the blocks involved in this type of connecting fragment of length l).

Exact patterns were then looked for by an exhaustive extraction of repeated motifs in the four types of connecting fragments. This program took as its input the SSBB assignment for each four-residue segment and detected the series of SSBBs that occurred at some frequency above a user-fixed value. Each repeated pattern that was found was identified by the fragment length of each pattern (the number of SSBBs), and the number of occurrences of this pattern in the particular type of connecting fragments.

3 Results

3.1 The SSBB categories

We identified 12 distinct SSBBs by fitting the HMMs to 19,017 experimental four-residue segments obtained from the database of 100 polypeptide chains. Progressively increasing the number of states from 3 to 12 significantly improved the model, as tested by using Bayesian Information Criteria. We were more interested in identifying a representative base of SSBBs, associated with at least 3% of the database of four-residue segments, than in identifying a detailed list of SSBBs. We therefore limited the number of SSBBs to 12. The appropriateness of applying our results to a general case was checked by fitting the HMMs to two nonoverlapping datasets of 50 different polypeptide chains. This identified very similar SSBBs. The 12-state HMM were labeled (α_1, α_2, α'_-,α'_+, α', γ_1, γ_2, γ_3, β'_-,β'_+, β_2, β_1).

Table 1 shows each of the 12 SSBBs with the means of the four distances (in Å) describing their associated average four-residue-segment conformation; the proportion of each SSBB; an index of similarity for each SSBB, as estimated from the average root-mean-square deviation (RMSD) of its associated four-residue seg-

Table 1. The 12 short structured 3D building blocks (SSBBs) identified by hidden Markov models analysis of a database containing 100 non redundant proteins, with the means of the four distances [in Å] of the corresponding average conformation of four-residue segments, their fraction of occurrence in the database (F, in %), the variation of each SSBB as estimated by the RMSD of the four-residue segments, and the correspondence with the usual secondary structures. The conventional secondary structure was broken down into SSBBs, this suggested that the 12 SSBBs, form five clusters: A, A′, B, B′, C

12 Identified SSBBs		Corresponding four-residue segments								
		Mean values [Å]				F	RMSD	Distribution of secondary structure in the third residue		
Cluster	SSBB	d_1	d_2	d_3	d_4	[%]	[Å]	α	coil	β
A	α_1	5.46	5.13	5.45	2.92	23.03	0.09	63.3	4.5	
	α_2	5.48	5.42	5.52	3.00	14.86	0.20	28.3	11.6	0.8
A′	α'	5.80	5.59	5.91	1.67	3.50	0.26	3.3	5	0.2
	α'_-	5.57	7.40	5.65	−3.18	2.97	0.56	0.5	6	0.3
	α'_+	5.64	7.46	5.67	3.38	3.66	0.38	3.4	5.2	0.2
C	γ_1	6.66	6.75	5.61	−0.28	3.00	0.59	0.5	5.9	0.2
	γ_2	6.21	9.10	5.67	−0.20	4.28	0.38		7.6	3.1
	γ_3	6.81	9.18	6.72	−0.61	3.80	0.84		6.7	3.8
B′	β'_+	5.70	8.26	6.74	1.60	11.44	0.73	0.7	19.5	10.2
	β'_-	6.68	8.57	5.55	−2.54	5.62	0.33		10.5	2.4
B	β_2	6.74	9.41	6.46	−2.35	8.78	0.32		9.8	21.1
	β_1	6.65	10.11	6.74	−0.66	15.06	0.34		7.6	57.6

ments; and their correspondence to the usual secondary structures.

Different SSBBs have relatively distinct average conformations as shown by the 4 mean distances. Distances d_1 and d_3 correspond to the distances between residues separated by one residue and have the smallest variations for the various SSBBs. The value d_2 describes the extent of the average four-residue segment conformation corresponding to each SSBB, and increases from SSBB α_1 to SSBB β_1. The fourth distance, d_4, indirectly describes the volume of the four-residue segments and also indicates their topological orientation. The volume and the topological orientation distinguish between similar average conformations (in terms of the first three distances) associated with SSBB α'_- versus SSBB α'_+, and SSBB β'_+ versus SSBB β'_-. The RMSD of the 12-state HMM are acceptable (average 0.43 Å), and below the standard threshold of 1 Å [2, 3].

Breaking down the conventional secondary structure into SSBBs suggested that the 12 SSBBs form five clusters A, A′, B, B′, C. The SSBBs in cluster A contain the largest fraction of residues classified as α-helices (91.6%) while those in cluster B are mostly β-strands (78.7%). Concurrent with their helical form, the two SSBBs from cluster A form the most compact structures, with a positive orientation. In contrast, the two SSBBs from cluster B correspond to more extended structures, β_1 describes four-residue segments that are nearly flat while those from β_2 have a negative orientation. Clusters A′ and B′ contain residues classified as both regular structures and coils, while the SSBBs in cluster C contain residues mainly classified as coils. The SSBBs in clusters A′, B′ and C involve a limited number of four-residue segments (about 4%), whereas those in clusters A and B account for many more segments (about 38% in A and 24% in B).

The HMMs approach allows one to describe different degrees of variation within each SSBB and thus to identify SSBBs that are well defined, with little variation, such as those in clusters A (RMSD < 0.2 Å) and B (RMSD < 0.35 Å), as well as others that are less precise, particularly, those in cluster C.

3.2 Organization of the protein structure in terms of the SSBB sequence

Our stochastic model is based on the local dependence of structural conformations on the neighboring conformations. HMM produced a direct estimation of the transitions between the identified SSBBs and thereby quantified the paths by which these elementary blocks are connected. Figure 2 illustrates this transition matrix between the SSBBs and shows a limited number of connections between SSBBs. Only some of the estimated transition probabilities are greater than 10%, most are few. Globally the only path between some clusters is one-way, suggesting that the connections by which the blocks form the protein structure are well organized. In particular, cluster A, which is clearly associated with the α-helices, and cluster B, linked to the β-strands, have no direct transitions and only a few possible pathways between them. But there are many connections within cluster A. There is an 84% probability that a protein remains in SSBB α_1 (average number of repeats: 6.3), while SSBB α_2 seems to be the preferred block by which connections are made to other clusters. This suggests that α_1 forms the core of α-helices, while α_2 describes their ends. Cluster A′ is the main path between cluster A and cluster B′. Cluster B′ connects cluster B to other clusters. Cluster B is the second most repeated cluster, with many switches. Cluster C is sometimes a gateway to cluster A, but goes

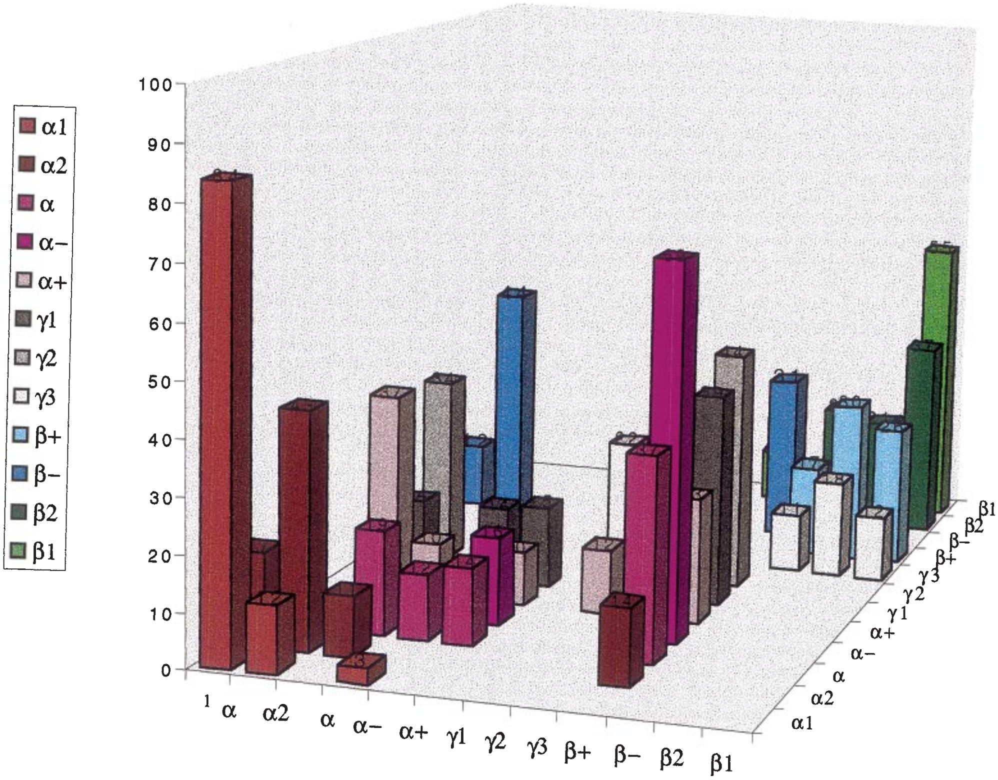

Fig. 2. Transition matrix between short structural building blocks (SSBBs) estimated by the hidden Markov model (HMM). Each probability of transition from SSBB *i* to the 12 different SSBBs is colored according to the SSBB category *i*

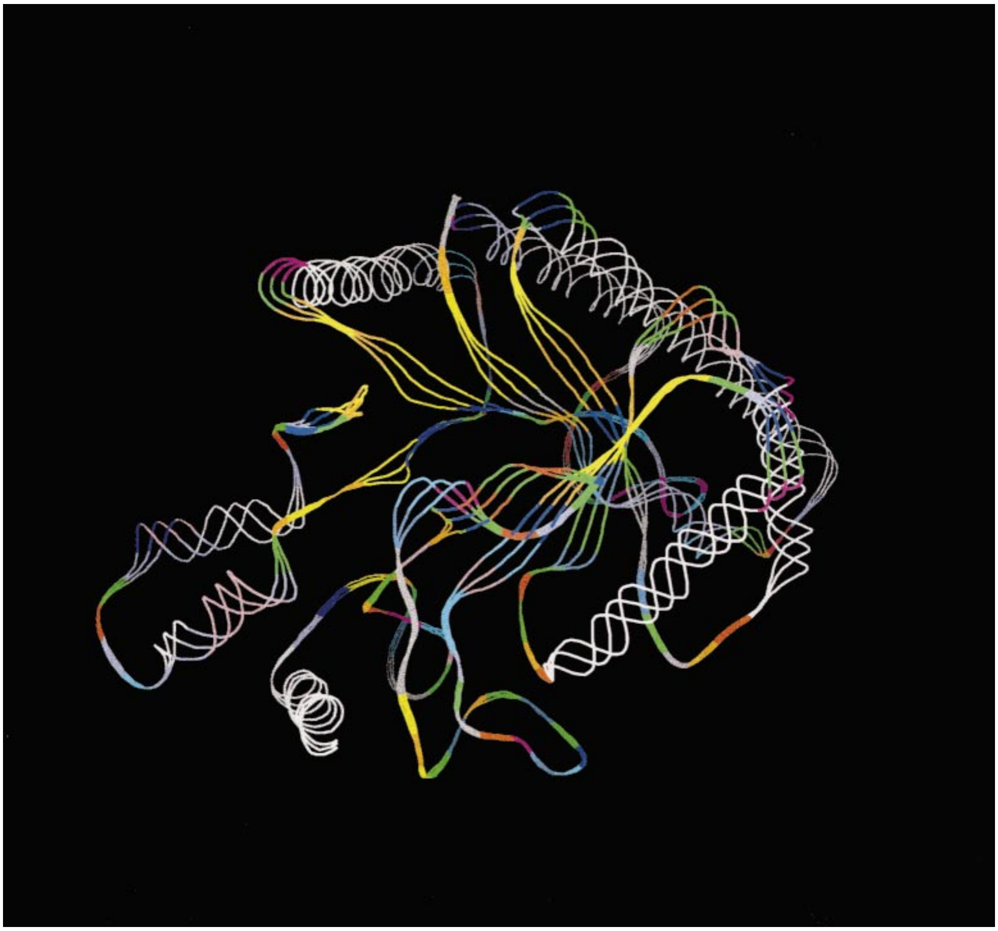

Fig. 3. Structure of the beta amylase (1tml.pdb), a β-barrel protein, colored according to its assignment by the HMM, showing the SSBB category of the third residue in each four-residue segment, α_1 (white), β_1 (yellow), β_2 (orange). The figure was generated using XmMol [20]

mainly to SSBB β'_{-}. Finally, cluster A' is found consistently at the N-end of helices and B' at the C-end of strands, while cluster C is typically found in regions identified as "random coil".

3.3 Protein backbones labeled in terms of SSBBs

All four-residue segments of the database were labeled in terms of these 12 blocks, so that each protein chain could be described as a succession of SSBBs. The structure of a β-barrel protein (1tml.pdb) is shown in Fig. 3, colored according to its assignment by this model, that is, the SSBB category of the third residue in each four-residue segment. The α-helices are classified as SSBB α_1 (white), except for their extremities. Cluster B also divided regular β-strands into SSBB β_1 and β_2 (yellow and orange). Coil regions or irregular strands were broken up into different successions of SSBBs.

3.4 Connecting fragments between secondary structures and repeated patterns

The distribution of the different types of connecting fragments as a function of their length (i.e., number of SSBBs) is illustrated in Fig. 4. In this database, the connecting fragments between two β-strands ($X\beta$) are longer than the other connecting fragments (average length: 9.1 SSBBs in 305 type $\beta\beta$ connecting fragments and 7.0 SSBBs in 251 type $\alpha\alpha$ fragments, 6.9 SSBBs in 283 type $\alpha\beta$ fragments and 6.7 SSBBs in 277 type $\beta\alpha$ fragments). The fragments followed by a β-strand are longer (two fragments $\beta\beta$ and one fragment $\alpha\beta$ contained more than 60 SSBBs, while type $X\alpha$ fragments contain fewer than 45 SSBBs). We then focused on the distribution of SSBBs in the four connection types for different lengths (the numbers of fragments in each length are given in Fig. 4). The equivalent number of SSBBs (denoted eSSBBs) in the various types of connecting fragments was computed from the Shannon entropy, as described above. Short connecting fragments (with fewer than 3 SSBBs) use only a limited number of SSBBs (value smaller than 4.5 eSSBBs), specific for the structure of the subsequent regular secondary structure. The equivalent number of SSBBs used for each length is close to 7.9 eSSBBs, with the average value for type $X\beta$ being slightly greater (8.3 ± 2.0 eSSBBs for type $\alpha\beta$ and 8.1 ± 1.9 eSSBBs for type $\beta\beta$ versus 7.3 ± 2.3 eSSBBs for type $\beta\alpha$ and 7.8 ± 1.7 eSSBBs for type $\alpha\alpha$). The SSBBs structure of the connecting fragments is influenced more by the structure of the following regular secondary structure than by that of the preceding one (cluster A before an α-helix and cluster B before a β-strand), for all fragment lengths. Fragments broken down into different SSBBs show more similarity between fragments with identical following structures (fragments $X\alpha$, or fragments $X\beta$) than between fragments with identical preceding structures (fragments βX, or fragments αX), see Fig. 5. More than 50% of the SSBBs in the $X\alpha$ fragments are mainly composed of SSBBs from clusters B and B', in particular SSBB β'_{-}. Fragments $\beta\beta$ involve more SSBBs from clusters C and A', in particular SSBB γ_3, while fragments $\alpha\beta$ involve more SSBBs from cluster A and SSBB β'_{+}.

3.5 Extraction of repeated patterns

Finally, we examined a series of SSBBs to determine whether the "SSBB categories" obtained by the HMM

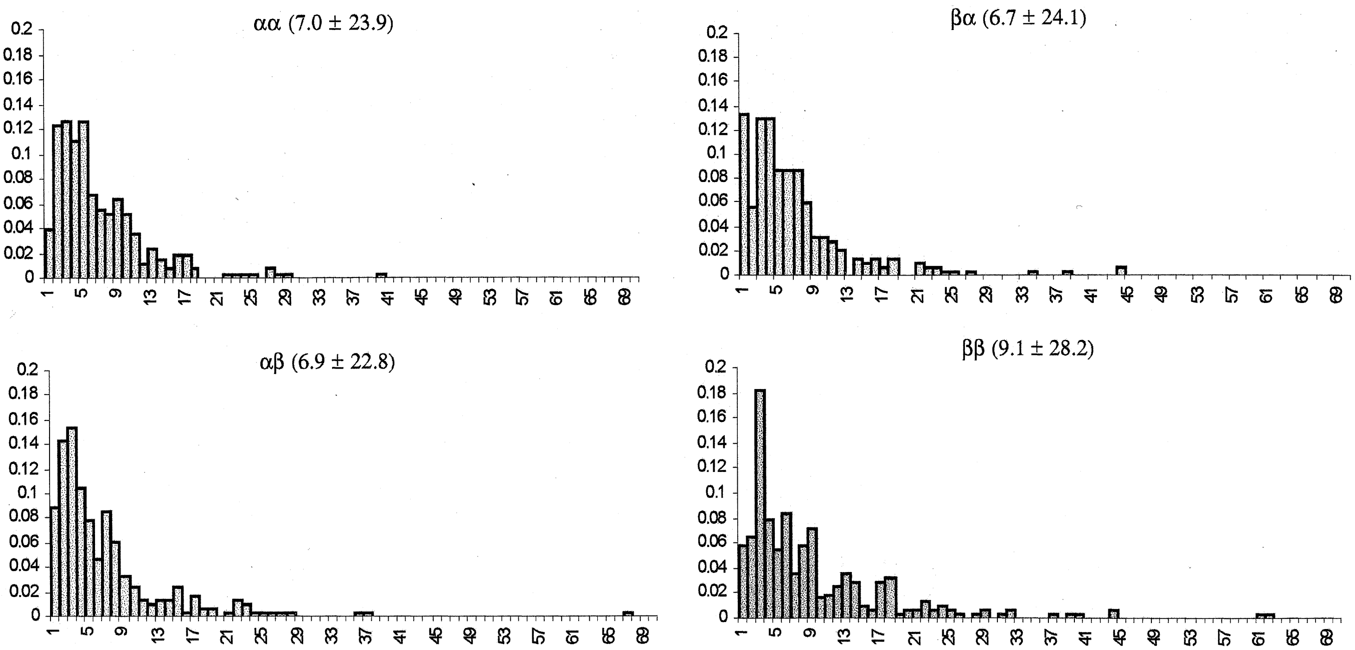

Fig. 4. Normalized distribution of the number of fragments as a function of their length (i.e., number of SSBBs) for the four types of connecting fragment

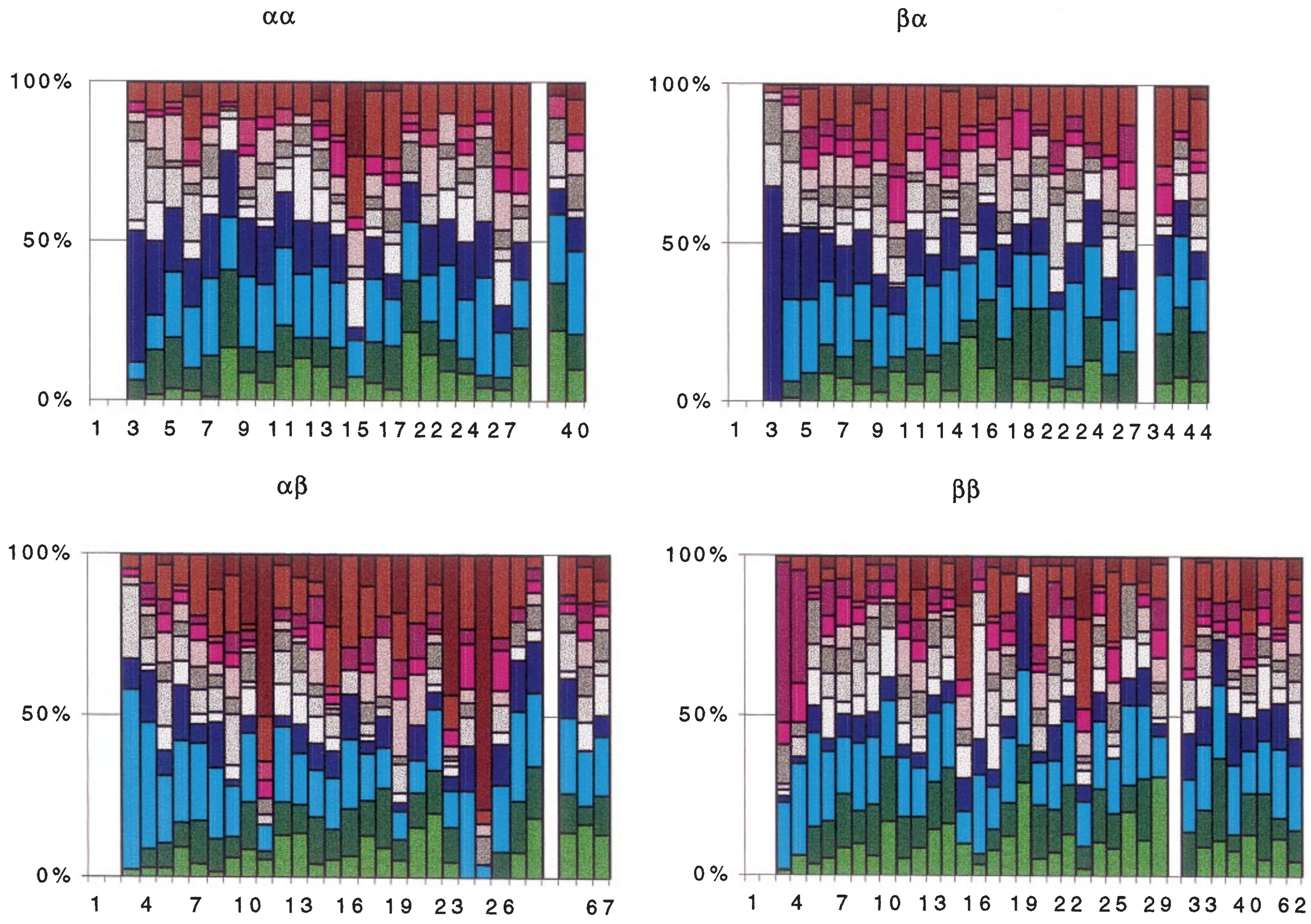

Fig. 5. Cumulative distribution of the 12 SSBBs [%] in the four types of connecting fragments, $\alpha\alpha$, $\alpha\beta$, $\beta\alpha$, $\beta\beta$, versus the length of these fragments

made sense in terms of the extraction of repeated series of SSBBs, or patterns, in different connecting fragments. Table 2 shows some patterns in different types of connecting fragments. These patterns were repeated more than twice, with a minimum length of 11 residues (8 SSBBs) and a maximum of 17 residues. However, these patterns were extracted using very restrictive criteria. First, we only look for exact matches between patterns, second, connecting fragment selection excluded residues flanking regular secondary structures and extracted only "coil patterns".

The patterns identified are mainly repeated in different proteins, but there are also some repeats within the same protein (8abp, for instance). Some proteins have patterns in different types of connections. And some proteins have several patterns common to another protein: proteins 2cyp and 1lgA have similar patterns in both $\alpha\alpha$ and $\beta\beta$ type of connections.

A few series of SSBBs or words (three or five SSBBs corresponding to six or eight residues) are present in several patterns for different types of connecting fragments. For instance, the six patterns identified in $\alpha\alpha$ type connections include four words that are repeated in different proteins, include a few SSBBs $(\alpha'\beta^+\beta_2)$, and two of them include longer series $(\alpha'-\beta^+\beta_2\beta^-\alpha_2)$. The last two patterns are found in a series of 14 SSBBs from protein 1lgaA, whose first part is shared by protein 1l18, and whose second part is shared by protein 2cyp, with two SSBBs in the two patterns. The five patterns identified in connecting fragments $\alpha\beta$ also show this phenomenon, the first two patterns involve a series of 13 SSBBs from protein 1mctA, whose first part is the same as that of protein 1omp and the second part is the same as in protein 1lgaA with four SSBBs in the two patterns. The last three patterns include the same word $(\beta^+\beta_2\beta^-\alpha_2\alpha_2\beta^+\beta_1)$. This word is repeated three times in protein 8abp and once in proteins 2acq and 2rslb. The $\beta\beta$ connecting fragments have 7 repeated patterns, 4 of which include the word $(\beta^-\alpha_2\beta^+\beta_2\beta^-)$. Again, two patterns are present in a long series of SSBBs from protein 1rdh whose first part is the same as in protein 1bfg, and whose second part is identical to protein 1add. The last pattern is noteworthy because it includes mainly SSBBs from cluster C. In the last type of connecting fragment $\beta\alpha$, 11 patterns are found, several of them repeated in protein 2aaiB. Six of them include small words $(\alpha^+\beta^+\beta_2)$, or $(\beta^+\beta_2\beta_2)$, or the longest one $(\alpha^+\beta^+\beta_2\beta_2\beta_1)$. Another word of this type $\beta\alpha$ is $(\beta_1\beta^-\beta^+\beta_1)$.

Finally, these results suggest that there are short series of SSBBs specific for different types of connecting fragments and show that there are relatively long series of SSBBs (11–14 SSBBs corresponding to 14–17 resi-

Table 2. Patterns identified in the database of nonhomologous proteins. The descriptors are: type of connecting fragments, length of the pattern (L), full SSBB series, number of repeats (Nb) and the protein PDB codes, followed by the position of the first residue in the PDB entry. These patterns include more than 8 SSBBs ($L \geq 8$) and are repeated more than twice ($Nb \geq 2$). Short series of SSBBs or words, specific to different types of connection are indicated in italics. Proteins with a long series of SSBBs whose two parts are identical to those of other proteins are underlined. Proteins having several patterns are indicated in italics

Types of connection	L	Pattern (series of SSBBs)	Nb	PDB code (position)
$\alpha\alpha$	8	$\alpha^+\alpha^-\beta^+\beta^-{}_2\beta^-\alpha_2\alpha_2\alpha_2$	3	1avhA(199, 277), 3gapA (167)
	8	$\alpha\alpha^-\beta^+\beta_2\beta^-\alpha_2\alpha_2\alpha'$	2	*2scpA* (113), 1csh (147)
	8	$\alpha_2\alpha^-\beta^+\beta_2\beta^-\alpha_1\alpha_1\alpha_1$	2	2acq (229), 1cus (67)
	11	$\gamma_1\alpha_2\alpha^-\alpha'\alpha^-\beta^+\beta_2\beta_1\beta^-\alpha_1\alpha_1$	2	*2scpA* (96, 128)
	8	$\gamma_1\beta^+\beta^-\beta^+\beta^-\beta^+\gamma_1\alpha_2$	2	1lgaA (179), 1l18 (272)
	9	$\gamma_1\alpha_2\beta^+\beta^-\beta^+\alpha'\alpha_2\alpha_2$	2	1lgaA (185), 2cyp (184)
$\alpha\beta$	9	$\beta^+\beta_1\gamma_3\alpha^-\beta^+\beta_2\gamma_1\beta^+\beta_2$	2	1mctA ($_1$77), 1omp (89)
	8	$\beta_2\gamma_1\beta^+\beta_2\beta^-\beta^+\gamma_2\beta_1$	2	1mctA (182), 1lgaA (199)
	8	$\beta^+\beta_2\beta^-\alpha_2\alpha_2\beta^+\beta_1\beta_1$	4	2acq (61), 8abp (154, 209), 2rslb (23)
	8	$\alpha^-\beta^+\beta_2\beta^-\alpha_2\alpha_2\beta^+\beta_1$	4	8abp (120, 153, 208), 2rslb (22)
	10	$\alpha^-\beta^+\beta_2\beta^-\alpha_2\alpha_2\beta^+\beta_1\beta_1$	3	8abp (153, 208), 2rslb (22)
$\beta\beta$	11	$\beta_2\gamma_3\alpha'\alpha_2\beta^+\beta_2\beta_2\beta^-\alpha_2\beta^+\beta_2\beta^-$	2	2cyp (131), 1lgaA (133)
	8	$\beta^-\alpha_2\beta^+\beta_2\beta^-\alpha_2\alpha_1\alpha_1$	3	2cyp (137), *1ppt* (5, 33)
	10	$\beta^-\beta^+\beta^-\alpha_2\beta^+\beta_2\beta^-\alpha_2\alpha_1$	2	*1ppt* (3, 31)
	10	$\beta^-\alpha_2\beta^+\beta_2\beta^-\alpha_1\alpha_1\alpha_1$	2	2pia (236), 1trkA (256)
	8	$\beta^-\alpha_2\alpha_2\alpha_2\beta^+\beta_2\beta^-\alpha_2$	2	1rdh (200), 1add (12)
	10	$\beta^-\alpha_2\alpha^-\beta^+\beta_2\beta_2\beta_2\beta^-\alpha_2\alpha_2$	2	1rdh (206), 1bfg (90)
	8	$\gamma_2\gamma_2\gamma_2\gamma_3\alpha'\alpha^-\alpha'\alpha_2\alpha_1\alpha_1$	2	1tndA (5), 5p21 (5)
$\beta\alpha$	8	$\beta^+\beta_1\gamma_2\alpha^+\beta^+\beta_2\beta_1\beta_1$	2	1gpr (88), 2stv (144)
	14	$\beta_2\beta^-\beta^+\beta^-\alpha'\alpha_2\beta^+\beta_2\beta_2\gamma_3\alpha^-\beta^+\beta_1\beta_1$	2	1ast (1, 242)
	11	$\beta^+\beta_1\beta_1$	2	*2aaiB* (12, 214)
	9	$\alpha^-\beta^+\beta_2\beta_2\beta_1\beta_2\beta_2\gamma_3\alpha^-\beta^+\gamma_1$	2	*2aaiB* (60, 100)
	9	$\beta^+\beta^-\beta^+\beta_2\alpha^+\beta^+\beta_2\beta_2\beta_1$	2	*2por* (159), 1pyaB (94)
	9	$\beta^-\beta^+\gamma_1\alpha_2\alpha^+\beta^+\beta_2\beta_2\beta_1$	2	2chsA (73), 3gapA (25)
	8	$\beta_1\beta_2\alpha^+\beta^+\beta_2\beta_2\gamma_3\beta^+\beta_1$	2	1cdh (25), 1hbq (114)
	8	$\beta^-\alpha_2\alpha^-\beta^+\beta_2\gamma_1\beta^+\beta_1$	2	*1alkA* (355), 1arb (127)
	10	$\beta^-\beta^+\beta_2\beta_1\beta^-\beta^+\beta_1\beta_2$	2	*2aaiB* (170), 1bfg (80)
	10	$\beta^-\alpha_2\alpha^+\gamma_1\alpha_2\beta^+\beta_1\beta^-\beta^+\beta_1$	2	*2aaiB* (69) *1alkA* (6)
	8	$\beta_2\beta_1\beta_2\beta^-\alpha_2\alpha_2\beta^+\beta_2\beta^-\alpha_2$	2	*2por* (99), 1tml (255)
	10	$\alpha^-\alpha'\alpha^+\beta^+\gamma_2\gamma_2\gamma_3\beta^+$ $\beta_2\gamma_1\beta^+\alpha^-\beta^+\beta^-\alpha_2\alpha_2\beta^+\beta_1\beta_1$	2	*2aaiB* (112), 2acq (233)

dues) in proteins whose different parts are shared by other proteins.

4 Discussion and conclusion

HMM analysis identified 12 distinct structural blocks with different roles without any a priori knowledge of the secondary structure. It is thus possible to identify the blocks corresponding to classic regular secondary structures and coils, and also to subdivide the α- and β-bounding regions. The HMMs approach therefore provides a more robust description of protein conformation than do algorithms based on predefined templates [21–24] or usual clustering approaches [2, 25–28]. The HMMs also automatically quantify the connections between the SSBBs and thus describes the preferred pathways by which blocks are assembled to form the 3D structure of a protein. Four types of coil fragments ($\alpha\alpha$, $\alpha\beta$, $\beta\beta$, $\beta\alpha$) between regular secondary structures were obtained from proteins and labeled in terms of SSBBs. The observed SSBB preferences for these connecting fragments suggest that they depend more on the regular secondary structure of the subsequent fragment than on that of the preceding one. The patterns formed by series of SSBBs repeated more than twice were extracted from different types of connecting fragments. Longer patterns, containing at least 11 residues (8 SSBBs) and less than 17 residues (14 SSBBs) were found. However, these patterns were extracted using an exact match between words, which is very restrictive. Hence, the only structural diversity of patterns came from the differences in SSBBs. The connecting fragments also do not contain any residues that flank regular secondary structures, unlike those of Ref. [28]. This definition results in the extraction of purely coil patterns and limits the number of residues in patterns. The extracted patterns suggest that the types of connecting fragments have a structural specificity and show that there are series of SSBBs in one protein whose different parts are identical to those of other proteins. These results confirm that SSBBs can be used as building blocks for analyzing protein structures.

Rackovsky [29, 30] demonstrated the existence of a local inverse protein folding code in proteins from four C_α structural fragments obtained in a sample of 114 protein structures. This suggests that it would be interesting to investigate the sequence specificity of SSBBs or series of SSBBs by studying the inverse protein folding code using the SSBB code. Further information could be obtained by applying HMMs to specific families of proteins, or by increasing the number of proteins. It may be possible to build an enhanced SSBB catalog that would permit a more detailed description of coil regions.

References

1. Orengo CA, Flores TP, Taylor WR, Thornton JM (1993) Protein Eng 6: 485
2. Unger R, Harel D, Wherland S, Sussman JL (1989) Proteins 5: 355
3. Unger R, Sussman JL (1993) J Comput Aided Mol Des 7: 457
4. Baum LE, Petrie T, Soules G, Weiss N (1970) Ann Math Stat 41: 164
5. Churchill GA (1989) Bull Math Biol 51: 79
6. Brown M, Hughey R, Krogh A, Mian IS, Sjölander K, Haussler D (1993) Proc Workshop AI Mol Biol 47
7. Stultz CM, White JV, Smith TF (1993) Protein Sci 2: 305
8. Krogh A, Brown M, Mian S, Sjölander K, Haussler D (1994) J Mol Biol 235: 1501
9. White JV, Stultz CM, Smith TF (1994) Math Biosci 119: 35
10. Sonnehammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Nucleic Acids Res 26: 320
11. Asai K, Hazamizu S, Handa K (1993) CABIOS 9: 141
12. Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C (1997) Protein 1: 134
13. Di Francesco V, Garnier J, Munson PJ (1997) J Mol Biol 267: 446
14. Berstein FC, Koetzle TG, Williams G, Meyer EF, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M (1977) J Mol Biol 112: 535
15. Hobohm U, Scharf M, Schneider R, Sander C (1992) Protein Sci 1: 409
16. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J (1993) Protein Eng 6: 377
17. Schwartz G (1978) Ann Stat 6: 461
18. Rabiner LR (1989) Proc IEEE 77: 257
19. Camproux AC, Saunier F, Chouvet G, Thalabard JC, Thomas G (1996) Biophys J 71: 1125
20. Tuffery P (1995) J Mol Graph 13: 67
21. Levitt M, Chothia C (1976) Nature 261: 552
22. Kabsch W, Sander C (1983) Biopolymers 22: 2577
23. Richards FM, Kundrot CE (1988) Proteins 3: 71
24. Zhu Z (1995) Protein Eng 8: 103
25. Rooman MJ, Rodriguez J, Wodak SJ (1990) J Mol Biol 213: 327
26. Fechteler T, Dengler U, Schomburg D (1995) J Mol Biol 253: 114
27. Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C, Saviano M (1996) Biopolymers 38: 705
28. Wintjens RT, Rooman MJ, Wodak SJ (1996) J Mol Biol 255: 235
29. Rackovsky S (1993) Proc Natl Acad Sci USA 90: 644
30. Rackovsky S (1995) Proc Natl Acad Sci USA 92: 6861