Theor Chem Acc (2001) 106:28–35
DOI 10.1007/s002140100261

Theoretical
Chemistry Accounts

Regular article

Exploring the use of a structural alphabet for structural prediction of protein loops

A. C. Camproux, A. G. Brevern, S. Hazout, P. Tufféry

INSERM U436, Université Paris 7, case 7113, 2 place Jussieu, 75251 Paris, France

Received: 17 July 2000 / Accepted: 5 January 2001 / Published online: 3 April 2001

Abstract. The prediction of loop conformations is one of the challenging problems of homology modeling, owing to the large sequence variability associated with these parts of protein structures. In the present study, we introduce a search procedure that evolves in a structural alphabet space deduced from a hidden Markov model to simplify the structural information. It uses a Bayesian criterion to predict, from the amino acid sequence of a loop region, its corresponding word in the structural alphabet space. The results show that our approach ranks 30% of the target words with the best score, 50% within the five best scores. Interestingly, our approach is also suited to accept or not the prediction performed. This allows the ranking of 57% of the target words with the best score, 67% within the five best scores, accepting 16% of learned words and rejecting 93% of unknown words.

Key words: Loop conformation – Conformation prediction – Proteins

1 Introduction

One of the most challenging problems in homology modeling remains the prediction of loop conformations. Being the less conserved regions of protein structures, they often cause serious errors in protein models because of their flexibility and the preferred occurrence of insertions and deletions. They are, however, often known to play an important role in protein function and stability [1]. Loop regions are organized as nonrepetitive conformations connecting regular secondary structures. They represent, on average, close to 30% of a protein. Although the conformations of these regions are, by essence, irregular, many preferred conformations have been identified [2, 3, 4, 5]. Some, authors have also suggested a relationship between loop conformation and sequence [6, 7].

Several attempts have been made to automate the prediction of the conformation of the loops. Owing to the number of possible conformations, the prediction of the conformation of loops has often been considered using conformational sampling techniques [8, 9, 10, 11]. A possible limitation of the size of the combinatorial is to look for conformations existing in the protein structures [12, 13, 14].

Finally, with the increasing number of structures available, several attempts have been carried out to classify the loop conformations and to extract some relationship with their associated sequences to perform prediction [15, 16, 17, 18, 19]; however, doing so, the authors are confronted with the problem of defining the different representative conformations used as templates and of establishing a relationship with some sequence signature.

Here, we explore whether a structural alphabet, composed of structural building blocks (SBBs), learned by applying a hidden Markov model (HMM) [20] from a collection of known structures, can be used to discretize the loop conformational space and to perform conformational prediction from loop amino acid sequence. Previous work has shown that the distribution of SBBs differs according to loop type [21].

The advantage of using a structural alphabet is that it simplifies the structural information, hence the combinatorial problem associated with the conformational search. Also, using such representation, it is, in theory, possible to perform a fast search for classes of "words" that could represent the conformation of a given loop. Finally, such a representation is well suited for automated search. The aim of our work consists of searching for words of fixed size characterizing exhaustively the different three-dimensional configurations of coils and predicting these words from the sequence windows (encompassing these series of structural blocks) by a Bayesian approach using probability estimations deduced from Dirichlet functions. A criterion of

Correspondence to: A. C. Camproux

Contribution to the Symposium Proceedings of Computational Biophysics 2000

predictability which indicates the ability of discriminating the words learned (i.e. those present in the training set of coils) and the new words (i.e. the configurations newly appearing in the assessing set of coils) is introduced.

2 Methods

2.1 Definition of the structural alphabet

The structural alphabet used in this study was obtained by fitting a HMM on a collection of proteins of known structure [20]. The structures were described as consecutive overlapping blocks of four residues. Each block was described by a four-distances vector: the three distances between the nonconsecutive α-carbons ($d_{C_\alpha 1-C_\alpha 3}$, $d_{C_\alpha 1-C_\alpha 4}$, $d_{C_\alpha 2-C_\alpha 4}$) and the oriented distance of the last α-carbon to the plane formed by the first three. Given such data, HMM then produces a short SBB description of the structures. The dependence between the successive SBBs is taken into account by a first-order Markov chain. The geometry associated with each block is reported in Table 1. Note that SBBs are not only described by their geometry but also by their transitions with others. For example, SBBs α_1 and α_2, describing α-helices, close in terms of geometry, are distinguishable by their transitions, while β_1 and β_2, strongly connected, both decompose β-strands. The variability of each SBB is less than 1 Å. Since in this study, we are interested in the loops connecting some elements of secondary structure, the distribution of each SBB in the three usual secondary structure types (helix, coil or β-strand) is also reported. The transition matrix associated with the Markov process is described in Ref. [20].

2.2 Encoding of the protein structures in the structural alphabet space

Knowing both the average geometry associated with each SBB and the transition matrix associated with the first-order Markov process, it is possible to translate from protein three-dimensional coordinates into the SBB space, or "alphabet space", by using the Viterbi algorithm [22]. This algorithm directly estimates the most probable series of SBBs underlying a structure. Its advantage is that it is, in theory, much more accurate than a simple step-by-step procedure. Hence, the use of the transition matrix between blocks is implicitely taken into account in the present study.

2.3 Collection of protein structures

The encoding into the alphabet space was performed for a collection of nonredundant protein structures taken from the "culled PDB" (http://www.fccc.edu/research/labs/dunbrack/culled-pdb.html). In order to keep a balance between the largest number of proteins selected for learning and the representativity of the dataset, we used the non redundant set presenting less than 50% sequence identity. Since loop sequences are known to be less conserved than core sequences [23], sequence identity of the loops is expected to be lower. We removed the proteins for which some ambiguity occur in the coordinates, such as missing residue or the presence of alternative conformations. This resulted in a collection of 878 proteins, representing after encoding a total of 195, 421 SBBs.

2.4 Identification of loops in the alphabet space

Given a protein description according to the structural alphabet, each loop is identified by a "structural alphabet word". For example, for $\alpha\alpha$ loops, we search for words of given length, l, delimited on both sides by two occurrences of SBBs $\alpha 1$ or $\alpha 2$. The pattern is thus: $2\{\alpha_1, \alpha_2\} - l(X) - 2\{\alpha_1, \alpha_2\}$, where l is the length of the loop and X is any character (no series of α_1 or α_2) apart from α_1 or α_2 at the two first and two last locations. In this study, we considered $\alpha\alpha$ and $\beta\beta$ loops using $\{\alpha_1, \alpha_2\}$ and $\{\beta_1, \beta_2\}$, respectively, as bounds, from three to 13 residues long ($3 \leq l \leq 13$) and their associated words. This results in a bank of structural alphabet words noted word$_l$ describing loops of length l for $\alpha\alpha$ loops or $\beta\beta$ loops type. For each l value, classes of words are defined, those differing by at least one SBB. We give the label class$_{k,l}$ to a particular class of words among a collection of N_l words found in the learning set to describe a given type of loop of length l.

2.5 Scoring function

To predict words of length l starting from a sequence in the 20 amino acid sequence space, we use a score based on the a posteriori probability calculated using Baye's theorem:

$$p(\text{class}_{k,l}/\text{sequence}_l) = \frac{p(\text{sequence}_l/\text{class}_{k,l}) \times p(\text{class}_{k,l})}{p(\text{sequence}_l)} , \tag{1}$$

where "sequence$_l$" is related to a sequence of length l in the 20 amino acid description, "word" to a series of l letters in the structural alphabet space and class$_{k,l}$ is a class of words.

$p(\text{sequence}_l)$ can be estimated according to an independence model of the l consecutive amino acids as $\prod_i f_i$, where f_i is the occurrence frequency of the observed amino acid i in the database. We preferred to learn a contingency matrix specific for each type of loop of length l on a window size of $4 + l + 4$. The enlargement of four residues both sides was done to take into account some specificity of the flanking sequences. The probabilities $p_{i,j/l}$ of the occurrence of each amino acid type i in position j of a window of size $4 + l + 4$ are obtained as

$$p_{i,j/l} = \frac{n_{i,j/l}}{N_l} , \tag{2}$$

Table 1. Description of the 12 sorts of structural building blocks (SBBs). d_1, d_2, d_3, d_4: mean and standard deviation of the four-distance values (see Methods) for the average conformation (in angstroms). *rmsdw*: similarity index within each SBB, as estimated from the average root-mean-square deviation obtained on a sample of its associated segments. %: the proportion of corresponding four-residue segments. α, coil, β: distribution of SBB segments on the usual secondary structures. A four-residue segment is classified in one secondary structure when its third central residue carbon is assigned to it

SBBs	d_1	d_2	d_3	d_4	rmsdw (Å)	%	α (32.9%)	Coil (47.2%)	β (19.9%)
α_1	5.45 ± 0.11	5.13 ± 0.16	5.45 ± 0.11	2.92 ± 0.17	0.09	23.07	90.3	9.3	0.4
α_2	5.48 ± 0.21	5.43 ± 0.35	5.53 ± 0.21	3.00 ± 0.40	0.20	14.84	62.9	36.7	0.4
α'	5.81 ± 0.33	5.59 ± 0.47	5.91 ± 0.28	1.66 ± 0.60	0.26	3.52	31.5	67.3	1.2
α'_-	5.57 ± 0.27	7.40 ± 0.98	5.65 ± 0.26	−3.18 ± 0.48	0.56	2.98	0.4	97.4	2.3
α'_+	5.64 ± 0.30	7.46 ± 0.83	5.67 ± 0.38	3.38 ± 0.44	0.38	3.63	31.0	67.9	1.1
γ_1	6.65 ± 0.38	6.71 ± 1.18	5.61 ± 0.27	−0.24 ± 1.71	0.59	2.98	5.3	93.3	1.4
γ_2	6.20 ± 0.41	9.10 ± 0.32	5.67 ± 0.26	−0.18 ± 0.97	0.38	4.2	1.6	84.1	14.3
$\gamma_{\alpha\beta}$	6.68 ± 0.31	8.57 ± 0.47	5.55 ± 0.26	−2.54 ± 0.53	0.33	5.62		80.2	19.8
$\gamma_{\beta\alpha}$	5.69 ± 0.27	8.25 ± 0.94	6.74 ± 0.32	1.60 ± 1.54	0.73	11.44	2.1	80.1	17.8
γ_β	6.81 ± 0.32	9.13 ± 0.89	6.71 ± 0.40	−0.61 ± 1.68	0.84	3.81	1.1	90.2	8.7
β_2	6.74 ± 0.31	9.40 ± 0.47	6.46 ± 0.26	−2.36 ± 0.48	0.32	8.77		52.0	48.0
β_1	6.65 ± 0.31	10.11 ± 0.34	6.74 ± 0.30	−0.65 ± 0.67	0.34	15.05		23.6	76.4

where $n_{i,j/l}$ is the number of occurrences of amino acid type i at location j of the window, among N_l words obtained from the database. Thus, we have

$$p(\text{sequence}_l) = \prod_{j=1}^{4+l+4} p_{i,j/l} \ . \tag{3}$$

Similarly, for each class k, we could estimate

$$p(\text{sequence}_l/\text{class}_{k,l}) = \prod_{j=1}^{4+l+4} p_{i,j/k,l} \ , \tag{4}$$

with

$$p_{i,j/k,l} = \frac{n_{i,j/k,l}}{N_{k,l}} \ , \tag{5}$$

where $n_{i,j/k,l}$ is the number of occurrences of amino acid i in position j of the window for the different occurrences of the class k, and $N_{k,l}$ is the number of occurrences of the class k. Since $N_{k,l}$ may be small, we use a different estimation of $p_{i,j/k,l}$, based on Dirichlet functions:

$$p_{i,j/k,l} = \frac{\alpha n_{i,j/l} + n_{i,j/k,l}}{\alpha N_l + N_{k,l}} \ , \tag{6}$$

where the coefficient α conditions the estimation. Low values of α lead to an estimation of $p_{i,j/k,l}$ close to values obtained for class k using Eq. (5), while large values result in values close to that obtained for the whole set of words obtained for length l with Eq. (2).

2.6 Criterion of acceptance of the prediction

The high variability of loops results in a possible large number of words describing loops of the same length. Thus, it is possible that words not learned in the learning set, called new words, appear in the validation set. It is desirable to have some indicator of whether the score associated with a given word from the amino acids sequence using Eq. (1) can be related to some correct prediction. To accept or not the prediction associated with a score, we use an acceptance criterion based on the difference between the two highest scores:

$$\begin{aligned}\Delta_{1-2} &= p(\text{class}_{k,l}/\text{sequence}_l)_{\text{rank1}} \\ &\quad - p(\text{class}_{k',l}/\text{sequence}_l)_{\text{rank2}} \ . \end{aligned} \tag{7}$$

We accept the prediction if Δ_{1-2} is larger than a given threshold, T, i.e. when a large difference between the first and the second highest scores is observed.

2.7 Quantification of the results

First we distinguish the rate of correct prediction (RCP) as the fraction of words learned correctly predicted in the validation set. A correct prediction will be either to obtain the best score for the class the word belongs to, corresponding to a correct prediction at the first rank noted RCP(1), or to obtain the class within the five best scores, corresponding to a correct prediction at the fifth tank, noted RCP(5).

For a given length l, the validation set consists of $N_{s,l}$ sequences tested. It can be decomposed into two parts: sequences associated with classes occurring in the learning set ("predictable classes"), denoted as $N_{s,l,\text{p}}$, and those associated with "new" classes, denoted as $N_{s,l,\text{np}}$, ("not predictable classes"). In terms of the criterion of acceptance of the prediction (Eq. 7), one can distinguish sequences ($N_{s,l,\text{a}}$) for which the prediction is accepted and sequences ($N_{s,l,\text{na}}$) for which it is not.

We define

1. The "sensitivity" associated with the threshold T as the fraction of the number of predictable sequences for which the prediction is accepted ($N_{s,l,\text{a}} \bigcap N_{s,l,\text{p}}$), among the number of predictable sequences $N_{s,l,\text{p}}$.
2. The "specificity" associated with the threshold T as the fraction of the number of unpredictable sequences for which the prediction is not accepted ($N_{s,l,\text{na}} \bigcap N_{s,l,\text{np}}$), among the number of nonpredictable sequences $N_{s,l,\text{np}}$.

3 Results

3.1 Loop distribution

Figure 1 shows, for different lengths, the number of words extracted from the database. Overall, 8,792 words were extracted (3,750 for $\alpha\alpha$, 5,042 for $\beta\beta$). Other studies found a mean loop number of 7.13 per protein in a

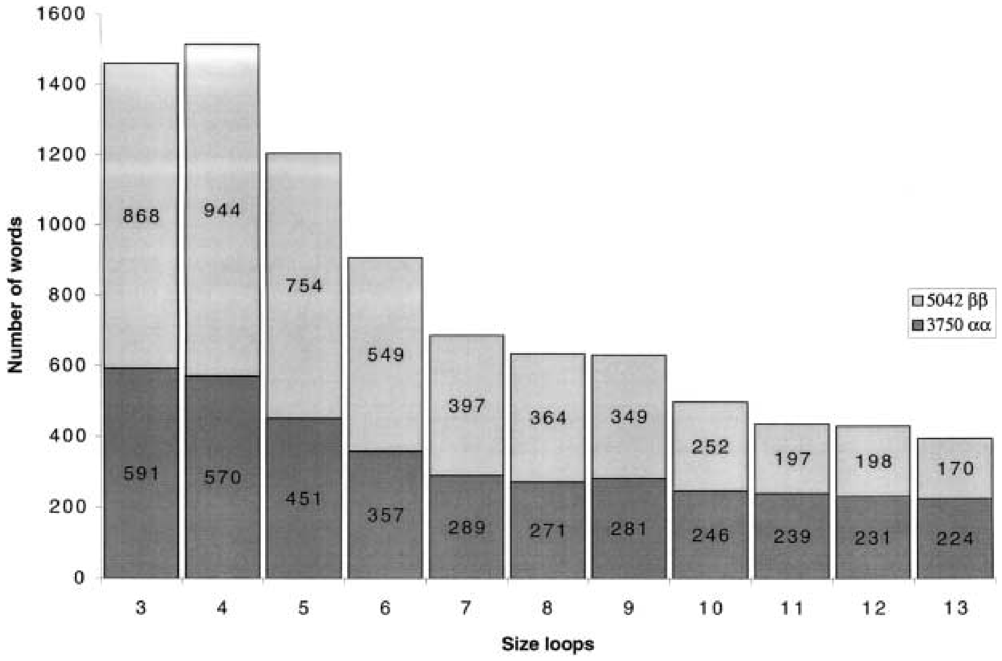

Fig. 1. Distribution of $\alpha\alpha$ and $\beta\beta$ loops according to loop size

comparable database (50% identity, 1,411 proteins) [15, 19]. Here we have a mean loop number of 7.05. Also the distribution of the loops as a function of their length is similar ($p < 0.001$). We observe more $\beta\beta$ loops, according to the observed presence of more β-strands in known protein structures. For a length larger than 8, this is not observed.

3.2 Distribution of words

The distribution of the number of words and classes for different sizes is shown in Fig. 2. Since we chose a ratio of 50% of the database as the learning set, we find a number of words similar in the learning and validation sets for different lengths. 50% of the 8,792 words appear in the learning set and 56.4% of the 5,029 classes are learned in it. Only 30% of the classes correspond to short loop lengths (less than 7) with more than 67% in the learning set.

The databank of words appears representative only for sizes less than 7 ("short" sizes): the ratio classes/words remains low (38% for $\alpha\alpha$ and 25% for $\beta\beta$). Since the complexity of the conformational space increases exponentially with word size, the number of repeated words decreases when the size increases. For lengths equal to or more than 7 and less than or equal to 13 ("medium" sizes), the number of classes is close to the number of words detected (i.e. a number of occurrences close to 1 for each class). Thus, for large sizes, the database is not representative enough to obtain statistical significance of the results; however, we kept this data since we wished to check the existence of some specificity inherent to the relationship SBB–amino acid sequence.

Finally, considering the two types of loops $\alpha\alpha$ and $\beta\beta$, we note that, despite the number of $\beta\beta$ words being larger for short sizes, the number of classes identified is of the same order as for $\alpha\alpha$ (2,564 versus 2,465). This implies a reduced variability in the alphabet space for $\beta\beta$ loops compared to $\alpha\alpha$ loops and a better representativity of the learning set. For instance, for a size of 3 and for $\beta\beta$ loops, more than 90% of the overall number of classes occur in the learning set.

3.3 Assessment of the effectiveness of the prediction

A preliminary step was to optimize the Dirichlet weight. The optimum (in terms of correct prediction) was reached for a value of 0.1 for the α parameter. The results are reported in Table 2.

For this value, the self prediction score (learning set) is between 90 and 100% for all sizes. For the validation set, we first focus on the prediction of classes occurring in the learning set (occurrences of new classes of the validation set are not considered). For short words (3–6), for which the number of occurrences of each word is more than 1.2, the mean prediction rate based on the best rank score is 28.4% (29.1% for $\alpha\alpha$, 27.8% for $\beta\beta$). This was obtained for a mean number of words of 236 for each size. For medium sizes (7–13) the score is 66.6% (67.6% for $\alpha\alpha$, 53.7% for $\beta\beta$), but with a number of predicted words between 1 and 25. Considering the five best scores instead of only the first one, the rates are 50.4% for short sizes (52.8% for $\alpha\alpha$, 48.0% for $\beta\beta$) and

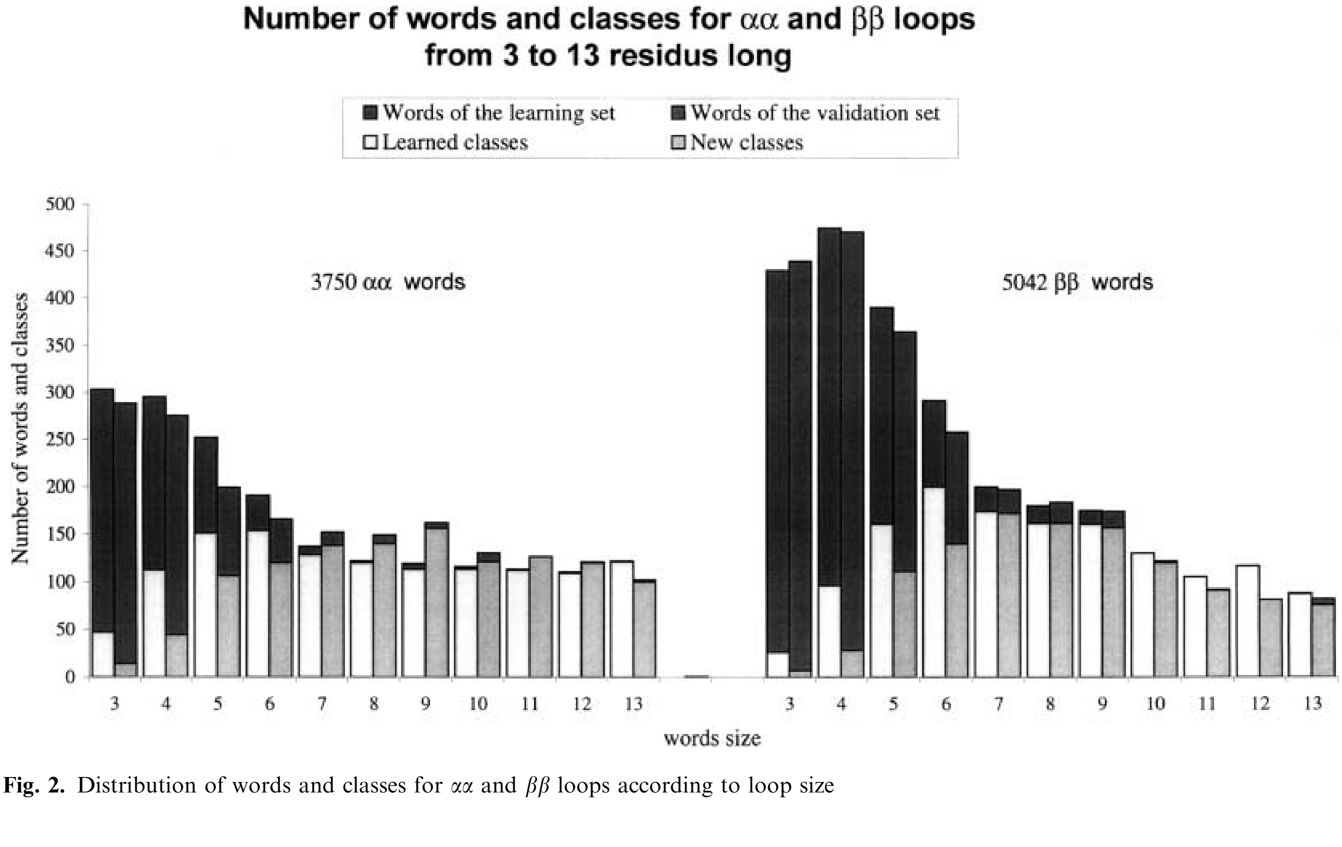

Fig. 2. Distribution of words and classes for $\alpha\alpha$ and $\beta\beta$ loops according to loop size

Table 2. Prediction achieved for the learning and validation sets. *RCP*(1): rate of correct prediction using best-ranked class (new classes not considered); *RCP*(5): rate of correct prediction using the five best-ranked classes (new classes not considered); *RCP*(1)*: equivalent to RCP(1), but new classes considered (systematic bad prediction)

Types of loops	Loop length	Learning set			Validation set				
		Number of words	Number of learned classes	RCP(1)	Number of words	New classes	RCP(1)	RCP(5)	RCP(1)*
$\alpha\alpha$	3	303	47	97.67	274	14	36.86	64.60	35.07
	4	295	112	98.62	231	44	29.87	51.52	25.09
	5	252	151	99.19	93	106	23.66	38.71	11.06
	6	191	154	100	46	120	26.09	56.52	7.23
	Mean (3–6)	260.25	116	98.87	161	71	29.12	52.84	19.61
	Mean (7–13)	120	121	100	5.86	128.71	67.59	83.86	2.78
$\beta\beta$	3	429	26	90.38	432	7	40.97	75.23	40.32
	4	474	96	91.43	442	28	25.11	45.70	23.62
	5	390	161	98.14	253	111	17	31.23	11.81
	6	291	200	100	118	140	27.97	39.83	12.79
	Mean (3–6)	396	120.75	94.99	311.25	71.50	27.76	47.99	22.13
	Mean (7–13)	142	134	99.92	22.14	122.71	53.68	60.52	4.59
$\alpha\alpha$, $\beta\beta$	Mean (3–6)	328.13	118.38	96.93	61.03	72.43	88.19	57.10	32.55
	Mean (7–13)	229.30	127.36	99.96	14	125.71	60.64	72.19	3.68

72.2% for medium sizes (83.9% for $\alpha\alpha$, 60.5% for $\beta\beta$). These results show that the procedure has a relatively good ability to predict learned words.

On introducing occurrences of new classes, not represented in the learning set, the scores of correct prediction at the first rank are much lower, 20.9% for short sizes, and decrease to 3.7% for medium sizes. These results are simply due to the increasing complexity of loops with length and thus to an increasing number of new classes.

3.4 Validation of the acceptance criterion

Since for a real prediction test, one only knows the sequence of the loop and one does not know a priori whether the SBB word describing it was learnt, we now analyze our results using an acceptance criterion.

The objective is here twofold: the identification of predictable words (learnt) and the optimization of the rate of correct prediction. For the first goal, we are interested in discarding unpredictable words, i.e., to obtain good specificity. For the latter, we prioritize the correctness of the prediction, even if this results in only a few words being predicted (low sensitivity).

Two examples of words and associated predictions in loops of type $\beta\beta$ for a length of 5 are given in Table 3. For instance, class 1, corresponding to word $\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$, is repeated 15 times in the learning set and is observed 14 times in the validation set. It is correctly predicted at the first rank on ten of 14 occasions (71.4%) and at the fifth ranks in all cases. Using the criterion of acceptance ($T = 1.28$), all sequences not correctly predicted at the first rank are considered as unpredictable. Class 2 corresponds to a word $\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$, which is not present in the learning set but is observed five time in the validation set. It is always considered as unpredictable using the criterion of acceptance.

Global results are reported in Table 4. We have considered different thresholds for $\alpha\alpha$ loops and $\beta\beta$ loops, and the results are presented for two sets of thresholds. First, we focus on short sizes. For the first thresholds (0.5 for $\alpha\alpha$, 0.64 for $\beta\beta$), we have a mean sensitivity of 15.6% and a specificity of 92.9% for short words, meaning only a few predictable words were extracted, but almost all nonpredictable words have been discarded. Interestingly, among predictable words, the score of well-predicted words is for the first rank only 57.1% (compared to 28.4%) and increases to 66.9% for the fifth-best scores. For the second series (1.5 for $\alpha\alpha$, 1.28 for $\beta\beta$), the sensitivity is lowered, but results in a better score of well-predicted words (62.1%). The specificity is slightly better (97.7%). For larger values of the threshold, the sensitivity decreases, and the set of predictable words is not representative any longer; hence, the weight of each failure becomes larger.

For medium words, we always obtain both good sensitivity (53%, 48.6%) and good specificity (91.4%, 93.8%) but owing to the low number of occurrences of the words (hence a poor learning) we only predict 24.1 and 28% of good predictions.

4 Discussion

4.1 How effective is the use of a structural alphabet?

In homology modeling, the structure of the backbone of the flanking regions as well as the sequence of the query region are assumed known. This study meets these requirements, and we have focused on two particular types of loops ($\alpha\alpha$, $\beta\beta$).

We perform loop conformation prediction in a structural alphabet space by accepting the equivalence between this space and the three-dimensional space. By

Table 3. Examples of words and associated predictions using the acceptance criteria. *OLS*: occurrence of class in the learning set; *OVS*: occurrence of class in the validation set; criterion of acceptance: $T = 1.28$, (refer to Methods), (1 if the sequence is considered as predictable, otherwise 0)

Classes	OLS/OVS	Words description			Corresponding predictions		
		Observed words	Corresponding sequences	Proteins: first residue	Predicted words (1)	Rank of correct prediction	T
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	FRELVFDDEKGTVDF	1obpA: 41	$\gamma_{\alpha\beta}\alpha_2\alpha'\gamma_{\alpha\beta}\gamma_2$	3	0
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	FGKFKLNLGTREMFR	1opc: 6	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	KKTAAWNSGTSTLTI	1ospO: 191	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	HICVSWESSSGIAEF	1sacA: 95	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	DGVWTYDDATKTFT	2igd: 47	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	NVVIAFNAATNVLTV	2ltnA: 163	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	KVTVIYDSSTKTLSV	2pelA: 164	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	TVALKADAANKQCRL	2sli: 114	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	EEFMRFNPRTGNWSG	3fruA: 119	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	DRVDEVDHTNFKYNY	1bv1: 77	$\gamma_2\alpha'\_\alpha_2\gamma_{\alpha\beta}\gamma_2$	4	0
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	MTVTRFDSMTGAHFV	1bw9A: 14	$\gamma_{\alpha\beta}\alpha_2\alpha'\gamma_{\alpha\beta}\gamma_\beta$	3	0
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	TAISKVNSDTNSLLY	1a59: 18	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	QSTRIYDRETGEIHY	3tdt: 212	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	1	1
1	15/14	$\gamma_2\alpha_2\alpha'\alpha'\_\gamma_{\alpha\beta}$	GDLVTYDKENGMHKK	7ahlA: 25	$\gamma_{\beta\alpha}\gamma_{\beta\alpha}\alpha'\gamma_{\alpha\beta}\beta_2$	4	0
2	0/5	$\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	TVEVELTTEKGVFRS	1oneA: 20	$\gamma_{\beta\alpha}\alpha_1\alpha'\_\gamma_{\alpha\beta}\beta_2$	–	0
2	0/5	$\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	QLVNFQCKEDGIIAQ	1plq: 26	$\beta_2\beta_1\gamma_\beta\alpha'\gamma_{\alpha\beta}$	–	0
2	0/5	$\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	RGEALIQTAYGEMK	1wapA: 53	$\beta_2\gamma_{\beta\alpha}\alpha_2\alpha'_1\gamma_{\alpha\beta}$	–	0
2	0/5	$\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	IVGIAVVNEHGRFFL	1xwl: 38	$\beta_2\gamma_{\beta\alpha}\alpha_2\alpha'_1\gamma_{\alpha\beta}$	–	0
2	0/5	$\gamma_\beta\gamma_1\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	DGTVFLSGAFGKIEM	2por: 60	$\gamma_{\beta\alpha}\gamma_{\alpha\beta}\alpha_2\gamma_{\alpha\beta}\gamma_\beta$	–	0

Table 4. Effect of the acceptance criterion. Sensitivity, specificity, *T*: refer to Methods. *T*1: $T = 0.5$ for $\alpha\alpha$; $T = 0.64$ for $\beta\beta$. *T*2: $T = 1.5$ for $\alpha\alpha$; $T = 1.28$ for $\beta\beta$

Types of loops	Loop length	Validation set							
		Sensitivity		Specificity		RCP(1)		RCP(5)	
		*T*1	*T*2	*T*1	*T*2	*T*1	*T*2	*T*1	*T*2
$\alpha\alpha$	3	19.34	9.12	92.86	100	77.78	96	88.89	100
	4	23.81	11.26	93.18	100	56.90	69.23	72.41	88.46
	5	9.68	6.45	97.17	98.11	66.67	75	75.00	75
	6	10.87	6.52	94.17	97.50	41.67	50	41.67	50
	Mean (3–6)	15.92	8.34	94.34	98.90	60.75	72.56	69.49	78.36
	Mean (7–13)	66.40	60.58	91.98	94.72	22.22	28.52	25.25	32.98
$\beta\beta$	3	22.92	9.95	85.71	100	67	83.72	84	90.70
	4	13.57	4.30	89.29	92.86	57.14	52.38	71.43	66.67
	5	7.91	5.53	96.40	97.30	37.50	23.53	45.83	29.41
	6	16.10	9.32	94.29	95.71	51.85	47.06	55.56	47.06
	Mean (3–6)	15.12	7.27	91.42	96.47	53.37	51.67	64.20	58.46
	Mean (7–13)	39.56	36.67	90.76	92.97	25.99	27.53	25.99	27.53
$\alpha\alpha$, $\beta\beta$	Mean (3–6)	15.52	7.81	92.89	97.68	57.06	62.11	66.85	68.41
	Mean (7–13)	52.98	48.63	91.37	93.84	24.10	28.02	25.62	30.26

using a limited number of "characters" to reproduce best the three-dimensional space, we avoid part of the difficulty inherent in using full three-dimensional space, which usually leads to a preliminary definition of classes of conformations. Using a discrete space is, in general, easier than considering a continuous space, and it offers the perspective of better understanding the continuity from one conformation to another by analyzing the changes in the characters. In the present study, we do not tackle the problem of going back from alphabet space to three-dimensional space. We focus exclusively on our ability to predict words.

Concerning the representativity of the alphabet space, it is conditioned by the limited number of characters (SBBs) used to describe protein conformations. Here, only ten different characters are combined to summarize loop conformations. However, we observe, for various lengths, a number of words detected very similar to the number of loops reported in other studies using a three-dimensional criterion [15, 19].

Finally, one main interest of using a structural alphabet is that it allows a large simplification of the combinatorial of the search, but we are still not able to reach a satisfactory representativity for each class when the size increases. For short words, the number of classes remains low; for medium words, this number increases greatly, and the number of occurrences per class is close to 1. Thus, the goal of preserving conformational complexity seems to be reached.

4.2 Efficiency of the prediction procedure

In terms of prediction, we use the relationship between the structural alphabet space and the amino acid space. Different studies have shown that there exists some amino acid sequence specificity for certain types of loops [16, 17, 19]. One major interest of the approach described herein is to combine both the specificity of the sequence inherent in each type of loop and the specificity inherent in each class of words. Using only words, we should be confronted with the problem of the representativity of the dataset. Using Dirichlet functions, we can reach an equilibrium between the weak representativity of words while preserving information depending on each type of loop. A classical Bayesian procedure could not be directly applied owing to a lack of representativity of certain words. Interestingly, we can study the sequence specificity associated with each word by observing the effect of the weight used in the Dirichlet function. Here, our best results are obtained for a weight value of 0.1, which is low, and suggest that the sequence specificity of each word is important, even for long sizes, as already suggested by Efimov [4] and Martin et al. [13].

How efficient can one expect the loop conformation prediction to be? Our scores are difficult to compare with the results of other approaches, usually in terms of root-mean-square (rms) deviations between the predicted conformation, i.e., one prototype of a cluster of conformations, and the target. Using a discrete space, an adapted metrics is rather in terms of change of characters. Here, "successful prediction" consists, from the amino acid sequence, in predicting the exact word in the structural alphabet space. For small sizes (up to 7) the mean prediction of words belonging to known classes, and using the only the best score, is close to 30%. This score increases up to more than 50% if one considers the five best scores. Note that the value of 5 seems particularly small facing the average number of classes (118 for short loops). Lessel and Schomburg [24] evaluated the quality of their prediction on the basis of a knowledge-base method by calculating the rms deviation to their target loops on the best 20 proposal target loops. A prediction was marked as successful, if at least one of the first three proposals had an rms deviation to the target below 1 Å. Their best results are for short fragments, with the percentage of correct predictions of 30%, which is comparable to our first rank results.

Finally, we also introduced the concept of "predictability" of a given sequence to face the problem of unknown words. Such a concept seems important since the existence of a large enough database to reach representativity for each class is far from being reached. Using such a criterion, we are able to reject as many as 93% of unknown small words; however, we accept that only 16% of known words are predicted, which is weak, among which 57% are scored at the first rank and 67% are within the five best ranks. Hence, for such cases, the procedure will only propose five different conformations among which the correct word is present. Rufino et al. [25] made an attempt to quantify the predictability of the loops using a score based, per class, on the frequency of the amino acids at each location. Their results showed a sensitivity larger than ours. For short words, they obtained as many as 75% of predictable loops accepted for prediction, with a correct prediction rate of 57%; however, their specificity was only close to 50% and decreased when the correct prediction of known classes increased. We did not observe such a fact: the specificity of our procedure is always more than 90%.

5 Conclusions and perspectives

In the present study, we investigated how plausible the use of a structural alphabet deduced from a HMM could be to perform conformational searches for loops. Our results are still incomplete since the whole study was performed within the alphabet space and since the conversion from such an alphabet space back to the three-dimensional space was not considered. However, before considering such a step, we need first to assess whether the simplification introduced by the use of such an alphabet reaches both the goal of describing loop conformational complexity and the goal of encompassing some specificity between amino acid sequence and loop conformation. In this respect, the present results are encouraging. First, the distribution of loops observed in the structural alphabet space is comparable to that of other studies. Second, the prediction rates of the words describing loop conformations in the structural alphabet space, as well as the fact that we are able to reject the prediction for most sequences associated with words not learned, suggest strongly that our procedure is able to capture the specificity of the sequences.

Interestingly, it is possible to extend this work in different ways. Considering the Bayesian criterion used, the results presented here were obtained by using the same sets of parameters whatever the lengths. It could be of interest to fit the Dirichlet weight and the predictability threshold for each loop type. In particular, the Dirichlet weight could be dependent on the number of occurrences and the length of each word. The rank used for the correct prediction could be a function of the criterion of acceptance of the prediction. Moreover, we did not investigate the influence of the size of the amino acid sequence window on the prediction rate.

Also, using a detailed structural alphabet to ensure a good description of the conformational complexity of the three-dimensional structures leads to the problem of a weak representativity for classes when the word size increases. We studied the effectiveness of the procedure considering no fuzziness of the words. A further direction to improve the prediction accuracy could be to consider a "fuzzier space", by accepting some equivalence between some of our characters defining the alphabet. One could search for the best equivalences either starting from analyzing the sequence signatures associated with each SBB or starting from the geometric proximity of their conformations.

Finally, if is also possible to extend the methodology to establish a direct relationship between amino acid sequence and structural alphabet sequence without considering the classes of conformation learned. The combination of such a "class-independent" approach with the methodology described here could lead to significant improvements.

References

1. Fetrow JS (1995) FASEB J 9: 708–717
2. Sibanda BL, Thornton JM (1985) Nature 316: 170–174
3. Rooman MJ, Rodriguez J, Wodak SJ (1990) J Mol Biol 213: 327–336
4. Efimov AV (1993) Prog Biophys Mol Biol 60: 201–239
5. Wintjens RT, Rooman MJ, Wodak SJ (1996) J Mol Biol 255: 235–253
6. Efimov AV (1993) Curr Opin Struct Biol 3: 379–384
7. Wilmot CM, Thornton JM (1990) Protein Eng 3: 479–493
8. Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C (1986) Proteins 1: 342–362
9. Collura V, Higo J, Garnier J (1993) Protein Sci 2: 1502–1510
10. Bruccoleri RE, Haber E, Novotny J (1988) Nature 335: 564–568
11. Moult J, James MN (1986) Proteins 1: 146–163
12. Jones T, Thirup S (1986) EMBO J 5: 819–822
13. Martin AC, Toda K, Stirk HJ, Thorton JM (1995) Protein Eng 8: 1093–1101
14. Van Vlijmen HW, Karplus M (1997) J Mol Biol 267: 975–1001
15. Kwasigroch JM, Chomilier J, Mornon JP (1996) J Mol Biol 259: 855–872
16. Donate LE, Rufino SD, Canard LH, Blundell TL (1996) Protein Sci 5: 2600–2616
17. Rufino SD, Donate LE, Canard LH, Blundell TL (1997) J Mol Biol 267: 352–367
18. Olivia B, Bates PA, Querol E, Aviles FX, Sternberg MJ (1997) J Mol Biol 266: 814–830
19. Wojcik J, Mornon JP, Chomilier J (1999) J Mol Biol 289: 1469–1490
20. Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S (1999) Protein Eng 12: 1063–1073
21. Camproux AC, Tuffery P, Buffat L, Andre C, Boisvieux JF, Hazout S (1999) Theor Chem Acc 101: 33–40
22. Baum LE, Petrie T, Soules G, Weiss N (1970) Ann Math Stat 41: 164–171
23. Chelvanayagam G, Roy G, Argos P (1994) Protein Eng 7: 173–184
24. Lessel U, Schomburg D (1999) Proteins 37: 56–64
25. Rufino SD, Donate LE, Canard L, Blundell TL (1996) Pac Symp Biocomput 570–589