Regular article

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

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

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
© Springer-Verlag 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
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 $\alpha$-carbons ($d_1$, $d_2$, $d_3$, and $d_4$) and the oriented distance of the last $\alpha$-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 SBBSs 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 $x_1$ and $x_2$, describing $\alpha$-helices, close in terms of geometry, are distinguishable by their transitions, while $\beta_1$ and $\beta_2$, strongly connected, both decompose $\beta$-strands. The variability of each SBBS is less than 1 A. Since in this study, we are interested in the loops connecting some elements of secondary structure, the distribution of each SBBS in the three usual secondary structure types (helix, coil or $\beta$-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 SBBS and the transition matrix associated with the first-order Markov process, it is possible to translate from protein three-dimensional coordinates into the SBBS 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 implicitly 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/culledpdb.html). In order to keep a balance between the largest number of proteins selected for learning and the representativeness 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 SBBSs.

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 $xz$ loops, we search for words of given length, $l$, delimited on both sides by two occurrences of SBBS $x_1$ or $x_2$. The pattern is thus: $2 \{x_1, x_2\} - l(X) - 2 \{x_1, x_2\}$, where $l$ is the length of the loop and $X$ is any character (no series of $x_1$ or $x_2$ apart from $x_1$ or $x_2$ at the two first and two last locations. In this study, we considered $xx$ and $\beta\beta$ loops using ($x_1, x_2$) 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: describing loops of length $l$ for $xz$ loops or $\beta\beta$ loops type. For each $l$ value, classes of words are defined, those differing by at least one SBBS. We give the label class$_{ij}$ 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:

$$p(\text{class}_{ij} / \text{sequence}) = \frac{p(\text{sequence} / \text{class}_{ij}) \times p(\text{class}_{ij})}{p(\text{sequence})},$$

where “sequence” 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$_{ij}$ is a class of words.

$p(\text{sequence})$ can be estimated according to an independence model of the $l$ consecutive amino acids as $\prod f_i$, 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 $p_{\text{ij}}$ of the occurrence of each amino acid type $i$ in position $j$ of a window of size $4 + l + 4$ are obtained as

$$p_{ij} = \frac{n_{ij}}{N_l},$$

Table 1. Description of the 12 sorts of structural building blocks (SBBSs). $d_1$, $d_2$, $d_3$, $d_4$ mean and standard deviation of the four-distance values (see Methods) for the average conformation (in angstroms). rmsdsv: similarity index within each SBBS, as estimated from the average root-mean-square deviation obtained on a sample of its associated segments. $\%$: the proportion of corresponding four-residue segments, $\alpha$, coil, $\beta$: distribution of SBBS 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.