Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.

Alexandre de Brevern, Catherine Etchebest, Serge A. Hazout

Proteins 15/11/2000; 41(3): 271-87

Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks.

A.G. de Brevern[†], C. Etchebest[*†] and S. Hazout[†+]

†:Equipe de Bioinformatique Génomique et Moléculaire , INSERM U436 ,
Université Paris 7, case 7113,
2, place Jussieu, 75251 Paris cedex 05, France.

* : Laboratoire de Biochimie Théorique,
UPR 9080 CNRS, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie,
75005 Paris, France.
present adress :†.

+ : Correspondence to : Alexander de Brevern,
E-mail : debreven@urbb.jussieu.fr
phone : 33 - 1 - 44 27 77 31 fax : 33 - 1- 43 26 38 30

short title : Prediction of structures in protein blocks.

key words : protein backbone structure, unsupervised classifier, structure-sequence relationships, structure prediction, protein block, Bayesian approach, prediction strategies.

May 31, 2000

Abstract

Using an unsupervised cluster analyser, we have identified a local structural alphabet composed of 16 folding patterns of five consecutive C_α ("protein blocks"). The dependence that exists between successive blocks is explicitly taken into account. A Bayesian approach based on the relation protein block-amino acid propensity is used for prediction and leads to a success rate close to 35%. Sharing sequence windows associated with certain blocks into "sequence families" improves the prediction accuracy by 6%. This prediction accuracy exceeds 75% when keeping the first four predicted protein blocks at each site of the protein.

In addition, two different strategies are proposed: the first one defines the number of protein blocks in each site needed for respecting a user-fixed prediction accuracy and alternatively, the second one defines the different protein sites to be predicted with a user-fixed number of blocks and a chosen accuracy. This last strategy applied to the ubiquitin conjugating enzyme (α/β protein) shows that 91% of the sites may be predicted with a prediction accuracy larger than 77% considering only 3 blocks per site. The prediction strategies proposed improve our knowledge about sequence-structure dependence and should be very useful in *ab initio* protein modelling.

Introduction

The protein sequence contains the whole information of the protein 3D structure. Proteins can not fold into unlimited number of structural motifs [1, 2]. Yet our lack of understanding of the physicochemical and kinetic factors involve in folding prevent us from advancing from knowledge of the primary sequence to reliable predictions of the biologically-active 3D structure. The first level of the protein structure is the secondary structure characterised in terms of α-helix, β-strand and unrepetitive coil. Thousand different predictions algorithms have been developed, statistical methods like the pioneer GOR [3, 4] or neural networks like the well-known PHD [5] and the more recent work of Chandonia and Karplus [6, 7]. The accuracy of these works were strongly increased with the addition of the multiple sequences alignment in the neural networks [8], probabilistic approach [9], or computational informative encoding [10]. The increase in the entries in the biological databases may permit an increase in the prediction rate [11].

Concerning the 3D structure, the *ab initio* protein folding algorithms, using only energetic or physicochemical parameters, were limited to small proteins [12, 13, 14]. Numerous studies describe the *ab initio* modelling of a 3D structure from the sole knowledge of its primary structure. However due to actual weakness of the prediction rate, this determination is still an open field.

The results obtained in the recent CASP III meeting are the best witnesses of such tentative [15]. The compatibility of the sequence versus known structures is an alternative approach to find the best approximation of the protein fold [16, 17]. Most of the methods for finding the folding state of a protein are mainly based on the use of the 3D structure of homologous proteins combined with simplified spatial restraints, statistical analysis and physico-chemical constraints [18, 19].

Recently, the use of fragment library [20] more detailed than 3-states and based on the most frequent local structural motifs (in terms of polypeptide backbone) encountered in the ensemble of 3D structure protein database, had led to improved results [21, 22] within a knowledge-based *ab initio* method [23].

Clearly, the main difficulty to overcome resides along the pathway going from the secondary structure prediction to the tertiary structure prediction. In this spirit, the study of the local conformations of proteins had a long history principally based on the study of the classical repetitive structures. We can notice interesting works such as those based on the geometrical and sequential characterisation of α-helices [24] or discrimination between the different types of β-turns [25]. Most algorithms which described global conformations of the proteins used this simple structural alphabet [26, 27, 28]. Recently, with the constant augmentation of the Protein Data Bank, automatic researches designed to determinate

families of specific coils have been carried out [29, 30, 31].

Among the different works concerning the definition of a structural alphabet (the consensus structural patterns will be labelled Protein Blocks or PBs), two main types of libraries of PBs can be distinguished : those composed of a high number (around 100) of protein blocks for describing protein structures or those characterized by a limited number of fold prototypes (4 to 13).

In the first type, the use of small blocks (fragments of six amino acids) for rebuilding a protein structure had begun with the work of Unger *et al.* [32] using the *rmsd* (root mean square deviation) as criterion. The authors have identified about 100 building blocks which could replace about 76% of all hexamers with an error of less than 1 Å. Schuchhardt *et al.* [33] similarly obtained a library of 100 structural motifs by an unsupervised learning algorithm from the series of dihedral angles. These libraries are adequate for approximating a 3D protein structure, however they are not easily usable for prediction.

In the second type of approaches, Rooman *et al.* define recurrent folding motifs by a clustering algorithm using the *rmsd* on distances between selected backbone atoms [34]. They described 16 motifs (embedded by groups of 4) of different lengths (from 4 to 7 residues). This small alphabet is directly related to the the four classes of secondary structure (α-helix, β-strand, turn and coil), and permits distinction between β-bulges and β-strands. Fetrow *et al.* developed an autoassociative artificial neural network (autoANN) to define 6 clusters corresponding to supersecondary structures encompassing the classical secondary structures [35].

Bystroff and Baker have generated a high number of similar short folds of different lengths and then grouped them into 13 clusters for a prediction approach [20]. A recent approach performed by Camproux *et al.* [36] takes account of the succession of the folds in the training by Hidden Markov Model (HMM) [37] and has allowed the definition of a library of 12 blocks of 4 C_α. This approach as expected allows the assessing of the transition frequencies between the blocks.

In this paper, the aims consist (i) in building a set of structural blocks able to approximate at best the different structural patterns observed along the protein backbones, and (ii) in predicting the local 3D-structure of the backbone in terms of PBs from the knowledge of the sequence. Identification of the different structural blocks is performed by an unsupervised cluster analyser taking account the sequential dependence of the blocks, this point is also considered in HMM [37].

After this phase of training performed on a given non-redundant protein database, we can tackle the problem of the prediction of these structural blocks from the knowledge of the protein sequence. From a given library of PBs, amino acid preferences for different

positions along the fragment can be extracted for each fold pattern. So using Bayes theorem, these probabilities may be further used to predict the structural motifs able to be adopted by a given protein chain. Bayesian probabilistic approach is largely applied in this type of study, for instance predicting solvent accessibility [38], secondary structure [9] or characterization of biological pathways from sequences to functions [39].

In this study, we have worked in different aspects to improve the PB prediction :

(i) *1 protein block - n sequences*: Associating one protein block with one class of sequences is a restrictive point of view. A same fold pattern (or PB) may be associated with different types of sequences, So, we have built a procedure for splitting the set of sequences (or "windows") encompassing a given PB into a fixed number of subsets showing (called "sequence families") at best different amino acid distributions in each window site.

(ii) *1 sequence - n protein blocks*: It is the inverse concept. Similar sequences are not always associated with the same fold [40], but with different "possible" folds. So we can devise a "fuzzy model" in which we have a certain probability for finding the true PB (this one that approximate at best the local structure of the backbone) among the proposed PBs. Concerning the existence of this "fuzzy model", we ought to check that the true PB is present among the solutions of the r first ranks (i.e. having the best prediction scores) provided by the Bayesian approach.

We will show the interest of an entropy-based index to discriminate different zones of the protein with high probabilities of prediction. With the scoring schema and the index control, two main directions have been explored. The first one called "global strategy" consists of locally determining the optimal number of protein blocks to be selected after fixing the prediction rate

for the whole protein sites. So, the number of selected solutions per position may be variable. In contrast, the second direction called "local strategy" scans the protein sequence with a fixed number of solutions (i.e. a constant number of protein blocks per position) and determines the regions able to be predicted with this given number and with a fixed prediction accuracy. In this way, the prediction only concerns these protein regions. Consequently, two prediction strategies are available; they both provide information complementary and enough for a former use in *ab initio* modelling.

Materials and Methods

Protein database

342 proteins are selected in a database of non-homologous protein structures (less than 25% of sequence similarity) [41, 42]. For each protein, we have stored the series of dihedral

angles and the primary sequences. Each protein backbone is transformed into a signal corresponding to the series of the dihedral angles (ϕ_i, ψ_i). So the database is composed of 342 signals. For the analysis, the proteins are splitting up in fragments of 5 consecutive residues to define the protein blocks. We have shared the set of proteins into two subsets, one of 228 proteins used for the training stage (i.e. the step allowing the definition of the protein blocks and the relationships between the protein blocks and the amino acid composition), the other one of 114 proteins for the stage of prediction accuracy assessing.

The proteins are classified according to the nomenclature based on the criteria of Michie and co-workers [27] which allows to share the protein set into four classes all α, all β, α/β and unclassified. The secondary structures are defined by a consensus assignment based on three algorithms [43].

Coding of 5-residue chains

A conventional approach for describing the backbone of a protein consists in converting the peptide coordinates into a series of backbone dihedral angles ϕ, ψ and ω. In the study, we will neglect the variation of the ω angle whose values vary around 0^o or 180^o [44]. We have limited the analysis to fragments of 5 residue length since it is sufficient to describe more than an short helix α (4 residues [24]) and a minimal β structure (3 residues [43]). A set of 5 consecutive peptides is an acceptable structural pattern to compute locations of hydrogen bonds between them. The link between the two successive carbons ($C\alpha_n$, $C\alpha_{n+1}$) located at the nth and (n+1)th positions in the protein sequence is defined by the dihedral angles ψ_n of $C\alpha_n$ and ϕ_{n+1} of $C\alpha_{n+1}$. A series of M peptides is defined by a signal of 2(M-1) values. So a fragment of 5 residues (M=5) centred at the alpha-carbon $C\alpha_n$ is displayed by a vector of 8 dihedral angles: $\mathbf{V}(\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2})$ associated with the consecutive carbons $C\alpha_{n-2}, C\alpha_{n-1}, C\alpha_n, C\alpha_{n+1}, C\alpha_{n+2}$, respectively [33]. The fragments used are overlapped. Hence, a protein of length L is described by L-4 fragments. This leads to a database of 86 628 dihedral vectors corresponding to the 342 protein signals.

Training by an unsupervised cluster analyser

The goal is to define a structural alphabet for coding the local 3D structure of protein backbones. This alphabet is composed of "Proteins Blocks" (PBs) which represent average patterns of the backbone fragments extracted from the database. Each one is defined by a vector of 2(M-1) values of dihedral angles like the fragments of the database.

Principle of the unsupervised cluster analyser

The method uses the principle of the self-organized learning of the Kohonen network (or Self-Organized Maps often noted SOM [45, 46]), i.e. by reading a certain number of times (called "cycles"), the totality of the vector database in order to define the "weights" of the neurons. In the terminology of the SOMs, the neurons and the weights correspond to a class of objects (in our study, protein blocks) and the associated information (herein, the average dihedral vectors) respectively. We have defined two steps in the training procedure. The first one consists of learning the protein blocks, by only considering the local protein structure (i.e. series of 5 carbons α), and the second one, by introducing constraints on the transitions between the protein blocks to favour a Markovian process of order 1, as the Hidden Markov Models (HMM [37]) to use the natural sequence of the protein structure. Our approach does not set any hypothesis of *a priori* distributions of the data (herein, the dihedral vectors).

Dissimilarity measure

The choice of the measure of dissimilarity between the M-residues fragments is essential for defining PBs strictly different in the training phase. In our study the vector associated with a fragment (called "dihedral vectors") is a series of dihedral angles. So the chosen dissimilarity measure between two vectors $\mathbf{V}_1$ and $\mathbf{V}_2$ of dihedral angles is defined as the Euclidean distance among the M-1 links, the *rmsda* (root mean square deviations on angular values [33]):

$$rmsda(\mathbf{V}_1, \mathbf{V}_2) = \sqrt{\frac{\sum_{i=1}^{i=M-1}[\psi_i(\mathbf{V}_1) - \psi_i(\mathbf{V}_2)]^2 + [\phi_{i+1}(\mathbf{V}_1) - \phi_{i+1}(\mathbf{V}_2)]^2}{2(M-1)}}$$

where $\{\phi_i(\mathbf{V}_1), \psi_{i+1}(\mathbf{V}_1)\}(resp.\ \phi_i(\mathbf{V}_2), \psi_{i+1}(\mathbf{V}_2))$ denotes the series of the (2M-1) dihedral angles for $\mathbf{V}_1(resp.\ \mathbf{V}_2)$. The angle differences are computed modulo 360^o.

So in the training, this distance is used for assessing the dissimilarity of any fragment of the database with the different PBs.

Description of the training approach

Each protein block PB_k considered as a neuron is initially defined by a vector $\mathbf{W}(k)$ of 8 dihedral angles (k denoting a given PB), either defined by a series of 4 dihedral angle pairs drawn in a Ramachandran diagram or randomly drawn in the database of vectors. Initially the number B of PBs is arbitrarily fixed. We assume that, at the beginning of the training, no transitions exist between these structural blocks.

α. *First Training.* In the first step of the training described in Figure 1, we consec-

utively read the dihedral vectors from a signal (representing a given protein). For every dihedral vector $\mathbf{V}(m)$ (m denotes the mth vector in the signal). The sequence of the dihedral vectors observed along the proteins is kept. We search for, among the B possible PBs, this one, PB_k, whose vector $\mathbf{W}(k)$ is the closest one according to the dissimilarity measure previously defined (minimal *rmsda*). Then the vector $\mathbf{W}(k)$ of this closest PB is changed into:

$$\mathbf{W}(k) + (\mathbf{V}(m) - \mathbf{W}(k)).\nu(c)$$

where $\nu(c)$ is a coefficient (initially taken low, ν_0 =0.02) decreasing during the training. $\nu(c)$ is a function of c denoting the number of vectors read during the training:

$$\nu(c) = \nu_0/(1 + \tau.c)$$

where τ is arbitrarily fixed to 1/N (i.e. the number of dihedral vectors, here N = 56442) so that $\nu(c)$ is reduced of half after one entire database reading. This procedure is conventionally used for the training of a Kohonen network.

The training is iterative: a certain number C of cycles of readings of the vector database is needed for defining the optimal vectors $\mathbf{W}$ associated with PBs. At each cycle, the proteins are randomly drawn for the treatment.

β. Refinement of the training. After C cycles of reading, we have obtained a first training of the protein blocks. They are used to encode the protein structures of the training set into series of PBs. Then, we compute the transition matrix between PBs by counting the occurrences of the pairs of PBs observed consecutively in the series and then transforming it in frequencies.

We have carried out again C cycles of database reading, but now we forced the transitions between protein blocks during the readings of the consecutive vectors in a protein (see Figure 1). This step consists of looking for the n blocks structurally close to a concerned vector $\mathbf{V}(m)$ among the whole set of PBs, and selecting in this subgroup this one having the highest transition frequency with the previous block defined for the vector $\mathbf{V}(m\text{-}1)$. So we forced the transitions between PBs when the structural similarity is observed. The number n of elements in each subgroup is a user-defined parameter.

γ. Determination of the optimal number of PBs by a shrinking procedure. We also have introduced a shrinking procedure to define an optimal number of PBs (or neurons). We start with a number B of PBs, then after each cycle, we test the "structural similarity" and the "transition similarity" between two PBs and we delete one PB between two PBs

considered as similar. The procedure is stopped when no deletion can be performed. The method allows to obtain an optimal number of PBs structurally dissimilar. The "structural similarity" between two PBs is defined by : PB_1 and PB_2 are structurally similar when the rmsda($\mathbf{W}_1$,$\mathbf{W}_2$) between the corresponding dihedral vectors $\mathbf{W}_1$ and $\mathbf{W}_2$ is less than an user-fixed threshold r_0. The "transition similarity" between two blocks PB_1 and PB_2 is obtained when the transitions probabilities of PB_1 and PB_2 to the other blocks are close. When the two criteria are verified, the least observed PB is deleted.

So, at the end of the process, each PB is represented by an average dihedral vector.

Propensities of amino acids to be located in a given PB

After the training, the whole proteins of training set are encoded according to the structural alphabet (i.e. the B blocks finally found) by using the minimal *rmsda* as a criterion. So each PB is associated with a set of sequence windows. It allows to compute one occurrence matrix of amino acid residues per PB. From the central position of a given PB, we examine the amino acid composition of the positions varying from -w to +w. The number of occurrences n_{ij}^k of a given amino acid (indexed by i = 1,2,...,20) located in a given position j (j varying in the range [-w,+w]) in the window is computed. We deduce the probability P(a_i in j / PB_k) by the ratio n_{ij}^k / N_k where N_k denotes the number of PB_k observed in the training set. P(a_i in j / PB_k) is the conditional probability of the amino acid a_i located at position j in a window encompassed PB_k. In our study, the length w has been fixed to 7, i.e. a sequence window of 15 residues.

Analysis of the occurrence matrix associated to PBs

We can analyse the relationships between a protein block PB_k and the amino acids present in the associated sequence windows :

(i) by assessing globally the specificity of each window location, i.e. to see which positions in each PB are the most informative in terms of amino acid distribution ,and ,

(ii) by determining which amino acids in certain location in the window are specific to this block.

To deal with the first point, we have used the relative entropy or Kullback-Leibler asymmetric divergence measure [49]:

$$K(\mathbf{p},\mathbf{q}) = \sum_i p_i \; ln\left(\frac{p_i}{q_i}\right)$$

This quantifies the contrast between the observed amino acid frequencies $\mathbf{p}$: $\{p_i\}_{i=1,...,20}$ and a reference probabilistic distribution $\mathbf{q}\{q_i\}$. We have applied this expression for as-

sessing the divergence $K_k(\mathbf{p}_j,\mathbf{q})$ of observed amino acid distribution p_j in a given position j of the window relative to the one observed in the database taken as reference distribution $\mathbf{q}$ for PB_k. The divergence profile, denoted by KLd profile, displaying the divergence measure function of the position j (value varying between -w and +w) allows to detect the "informative" locations for a given protein block.

The relative entropy $K(\mathbf{p},\mathbf{q})$ (value non negative) multiplied by 2N (N is the number of observations) follows a chi-square of 19 degrees of freedom (since we analyze the amino acid distributions). So for defining the informative locations for a PB type, we can threshold the KLd profile, the lower limit being $\chi^2_{19}/2N$, χ^2_{19} denoted the chi-square value obtained for a given type I error α and 19 degrees of freedom.

Concerning the second point, we have normalized the amino acid occurrences of each position into a Z-score $= \left(n^k_{ij} - n_{ib}\right)/\sqrt{n_{ib}}$ where n_{ib} is the expected number of the ith amino acid ($n_{ib} = \mathrm{N}_k.\mathrm{f}_i$ where N_k and f_i denote respectively the number of PB_k and the observed frequency of the amino acid i in the database). The positive Z-scores (respectively negative) correspond to over-represented amino acids (respectively under-represented) in the block PB_k, a threshold value of 4.4 had been chosen, i.e. a probability p less than 10^{-5}.

Prediction by an Bayesian probabilistic approach

Principle

For every site s of a protein, i.e. the central position of the PB and the sequence window, we would calculate for a given amino acid chain X_S, the probability of observing a given protein block PB_k, $P(PB_k/X_S)$.

From the information given by the conditional probabilities previously defined, it is possible to compute this probability by using the Bayes' theorem. It accomplishes the inversion for the sequence X_S and the structure PB_k:

$$P\left(PB_k/X_S\right) = \frac{P\left(X_S/PB_k\right).P(PB_k)}{P\left(X_S\right)}$$

where $P(PB_k)$ is the probability of observing the block PB_k in the database, and $P(X_S)$ is the prior probability of observing the chain X_S of residues without structural information, i.e. the product of the frequencies of the amino acids assessed from the database. A similar approach was described by Thompson and Goldstein [9] for the secondary structure prediction.

The term $P(X_S/PB_k)$ is the conditional probability of observing the given chain X_S (a_{-w},...,a_{+w}) of amino acid residues in the window given the particular type of protein

block PB_k. It can be computed as the product of the probabilities of observing each amino acid of the chain in the positions of the window. This leads to the equation:

$$P(X_S/PB_k) = \prod_{j=-w}^{j=+w} P(a_j/PB_k)$$

To define the optimal protein block PB^* for a given amino acid fragment X_S around a site s in a protein, we use the ratio R_k (or its logarithm) defined by,

$$R_k = \frac{P(PB_k/X_S)}{P(PB_k)} = \frac{P(X_S/PB_k)}{P(X_S)}$$

From the Bayes' theorem, we compute R_k defined by the ratio $P(X_S/PB_k)/P(\mathrm{X}_S)$ which is easily computed from the occurrence matrices. By this ratio, we compare the probability of observing a given protein block PB_k given the sequence X_S with the prior probability of observing PB_k given no sequence information. So, when $ln(R_k)$ is positive, the knowledge of the sequence X_S favours the occurrence of the block PB_k, and conversely when it is negative.

The rule for defining among the B possible blocks the optimal structural block PB^* for X_S consists of selecting the protein block PB_k for which the ratio R_k is maximum.

Consequently, for every sequence window, we define an ordered list of B protein blocks according to the computed ratios, the optimal protein block corresponding to the first rank.

So we can assess the prediction by the percentage Q(1) of correct predictions at the first rank, and Q(r) when the true block is among the r first solutions.

Improvement of prediction

Bayesian approach implies the use of one occurrence matrix by PB, however, sequences significantly different may be associated with the same fold. So, we introduce the concept of "sequence family". Relative to the fuzzy model, this notion is related to the first concept *1 fold - n sequences*, n denoting the possible number of sequence families associated with a given block. To define the sequence families, a procedure similar to the protein block learning is used. For each PB_k, the corresponding set of sequences is arbitrarily divided into f groups. In the first step, an occurrence matrix is computed for each of the f families, called PB_k^l with l varying from 1 to f. In the second step, for each sequence X_S the conditional probability $P(X_S/PB_k^l)$ is computed for the f different occurrence matrices. So f probability scores are calculated. Each sequence is then reallocated to the corresponding subgroup with the maximum probability. At the end of the first step the f matrices are computed again. Once all the sequences have been tested, the procedure

restarts from step one. The training is stopped when the reallocation weakly modified the matrices between two consecutive cycles.

The optimal number f of sequence families for each PB (we check a number varying from 2 to 6) is determined on the basis the increase of prediction rate Q(1).

Optimising protein block prediction

Our purpose is to predict the local 3D structure of the protein backbone encoded in protein blocks. We introduce the second concept *1 sequence - n folds* of the fuzzy model in which a given sequence has a probability distribution to be associated with the different blocks. The true PB may be in the most probable PB (i.e. at the first rank) but not always, it may be among the first selected blocks. Consequently, we want to define the optimal number of protein blocks to be selected in each site of a protein, i.e. the rank r of the ordered PB list given by the Bayesian approach, in order to ensure a given percentage Q(r) of correct predictions. In the following section, we describe two strategies of prediction based on the Shannon entropy function.

A large homogeneity of the scores R_k at a given site would show poor sequence specificity of the chain X_S and would lead to a prediction weakly accurate at the first rank. Conversely, a high score at the first rank would be associated to a good prediction. So in this case, it is necessary to keep r first protein blocks (according to the scores R_k) in order to obtain a certain level of correct predictions, i.e. the probability of observing the true block among the r selected PBs. To quantify the "uncertainty" with regard to the prediction, we have calculated an entropy over the scores R_k transformed into probabilities $S_k = R_k \,/\, \sum_l R_k$, with l for all the PBs. The expression of the entropy is

$$H = -\sum_k S_k \ln (S_k)$$

where k is an index over types of protein blocks.

We transform H into $N_{eq} = \exp [H]$ (called equivalent number of protein blocks). This quantity varies between 1 (i.e. when an unique block is predicted) and B (when the B blocks are equiprobable).

We have extracted the sites whose entropy vary within a given range, and we have built the corresponding distribution of the rank of the true PBs found in the ordered PB lists given by the Bayesian approach in every site. From this distribution (associated with a given N_{eq} interval), we determine the optimal rank r corresponding to a fixed Q(r). This step has been done extensively for all ranks of solutions possible, i.e. for 1 to B per site.

Two different prediction strategies are defined from the distributions previously defined:

(i) *a global approach* to define the number r_s of blocks to be selected among the possible protein blocks for each position s of a protein sequence, the prediction accuracy Q_g being

a user-fixed parameter. In this case, the number of selected number of selected PBs is variable along the protein.

(ii) *a local approach* to search for the positions along the sequence for which we can find again the true block with a given prediction accuracy Q_l by taking the r first protein blocks defined by the Bayesian approach. r and Q_l are fixed by the user. In this approach, the prediction is limited to certain regions of the protein sequence.

Results

In a first section, we describe the different protein blocks obtained by the unsupervised cluster analysis. The following characteristics are calculated : the dihedral vectors, the *rmsda* and *rmsd* (the conventional root mean square deviations computed from the C_α coordinates), the occurrence frequencies in the secondary structures (α-helix, β-strand and coil) and transition frequencies between consecutive blocks.

In a second section, we assess the prediction accuracy of the Bayesian strategy with or without the presence of several sequence families per block. Also, we discuss the possible effect of the protein size or the protein type on the prediction accuracy.

In a third section, we detail the results of the two prediction strategies applied to a protein, the ubiquitin conjugating enzyme (code name PDB : 2aak). This protein is taken as an example since its 3D structure shows both α-helices and β-sheets.

Description of protein blocks

In our study, we have selected an alphabet of 16 PBs which gives a good angular approximation with an average rmsda of 30^o. The least represented PB is associated with 1% of the database. This choice of the number of PBs is explained in section Discussion.

Figure 2 shows fragments superimpositions (MOLSCRIPT software[50])for the 16 PBs (denoted by the letters a, b,...,p) obtained by the unsupervised cluster analysis. The PBs are ordered on the basis of their transitions and their locations frequencies in the secondary structures. These information are given in Table I.

Within variability

The quality of the PBs is assessed through a variability measure. Using the dihedral vector $\mathbf{V}_k$ representative of each PB_k, the corresponding $C\alpha$ coordinates of blocks named $C\alpha^k$ are constructed. The *rmsd*s of the set of the $C\alpha$ series belonging to PB_k with the average $C\alpha^k$ is computed. Globally, the mean approximation of 21^o of the local backbone structure is convenient. The PBs show a good average local *rmsd*, as seen in Table I, less

than 0.74 Å, except for PB*j*. It must be noted that the PBs specific of the secondary structure, PB*m*, central α-helix and PB*d*, protein block for β-sheet, are not the only well approximated PBs. Several other PBs have their average *rmsds* close to 0.5 Å, as PB*p* (0.46 Å).

Structural difference between blocks

The computed *rmsd*s between the average locations of $C\alpha^k$ are distributed from 0.21 to 2.07 Å. PBs *m* and *n* (0.21 Å), *f* and *h* (0.23 Å), *n* and *o* (0.24 Å), *c* and *d* (0.25 Å) are the closest ones. However, these small values do not reflect the diversity of the block shape.

The *rmsda* between the pairwise protein blocks varies between 19.2^o and 47.8^o. The closest PBs are PBs *m* and *n* (19.2^o), *f* and *h* (19.5^o) and *c* and *d* (19.8^o). The observation of the differences of the angular values for these three pairs show that 5 to 6 angles are very close (less than 10^o) and only 1 to 3 angles entirely different (more than 100^o). It gives to these PBs their structural specificity. So, the *rmsda* more sensitive to the difference between BPs is a more appropriate measure to quantify dissimilarity between blocks than the *rmsd*.

Reproduction of the structure

Each angle of the protein structure, due to the use of a sliding window, is associated with 4 PBs with exceptions of N- and C-terminal. The resulting angle is defined as average of the angles of the 4 PBs. This procedure leads to a good approximation. Only 3 % of the angles are badly approximated (more than 90^o of difference with the reality), and more than 50% of the protein angles is approximated with less than 21^o. Moreover, we have checked that the attribution of the Portein Blocks is partically insensitive to the variation in the temperature factors of the bond lengths and valence angles. So 16 PBs is a convenient number to approximate all the protein structures.

Other works have described alphabets from various lengths [34, 35, 20]. To study the coding quality on motifs (i.e. series of PBs) of different lengths, we have extracted the motifs connecting two consecutive repetitive secondary structures PB*m* and/or PB*d*. Motifs ranging from 1 to 6 block length are examined. For instance, *mm*(*xyz*)*dd* is a motif *xyz* connecting two PB*m* and two PB*d*.

Among them the most representative motifs for each length are *mm*(*cc*)*dd* (30 observations, *rmsd* 0.70 Å), *dd*(*fkl*)*mm* (414 obs., *rmsd* 1.26 Å), *dd*(*fbdc*)*dd* (121 obs., *rmsd* 1.43 Å), *mm*(*nopac*)*dd* (215 obs., *rmsd* 0.76 Å), *dd*(*fkopac*)*dd* (64 obs., *rmsd* 1.05 Å).

For short motifs of one or two blocks length, the occurrences are low (less than 40

observations). While for larger motifs, the global number of PBs' combinations grows up. For instance, for a length of 4 (i.e. 5 and 6), an average of 20 different motifs are computed (i.e. 22 and 30). The number and type of motifs are strongly different and inhomogeneous, depending exclusively of the types of the secondary structures located at extremities. In the case of length 3 delimited by the motif connecting *dd* and *mm* extremities, the motif *fkl* represents 98% of the motif, and, connecting *mm* and *mm* extremities, *nop* represents 82% with only 24 occurrences in the database. In all the other cases and whatever the lengths there are no motifs which represents more than 75% of the structure examined.

As a result, the structural approximation of the 3D structure by means of protein blocks stay correct with an important number of PBs.

Transition between PBs

A large diversity of transition is observed. Table I gives the output frequencies (i.e. $\pi_{ij}/(1-\pi_{ii})$, π_{ii} and π_{ij} denoting the transition frequencies of the ith PB toward the ith and jth protein block respectively).

The three main transitions for the non-repetitive PBs correspond to at least 76% of the possible transitions (apart PBj). For instance, the transitions towards PBd (62.2%), PBf (24.4%) and PBe (5.6%) represent 92.2% of all the transitions from PBc. In the same way, more than half of the possible transitions does not appear with a frequency less than 1.0%. The number of transitions, which have a frequency more than 5%, is generally 3 and reaches 5 at most.

Figure 3 shows the coding of the protein 2aak in terms of PBs and the variation of the *rmsd*s associated with each PB along the protein structure. The succession of PBs d and m are easily pointed out. Various repeated motifs are observed such 4 chains (*cfkl*) located in the coils leading to the $\alpha-helix$, 4 (*dfk*), 2 (*ehia*), 2 (*bccd*) between β-strands and 2 (*opacd*) located in the coils. Small rmsd values less than 0.46 Å(average *rmsd*) are not only associated with repetitive PBs.

It must be noted than the PBj is the only PB badly designed. All the *rmsds* are less than 0.74 Å, apart it (1.03 Å). This due to its low occurrence frequencies (0.96%) and its absence of high transition frequencies (less than 17%).

Relationship with secondary structures: the repetitive PBs

The PBs can be characterised by their secondary structure composition. We note that they do not correspond exactly to classical secondary structures (as noted in the last column of Table I). The PBm is a central α-helix ($\psi = -47^o$ and $\phi = -57^o$). The PBd is structurally an "ideal" protein block for β-sheet ($\psi = 135^o$ and $\phi = -139^o$).

So, the PBs labelled from *a* to *c* and *d* to *f* are grouped around *d* due to high propensities to go in or out it. The third C_α of this block show a propensity to be located in the β sheets. The PBs labelled (*k*, *l*) and (*n*, *o*, *p*) are the local structures concerning the α helix N-cap or C-cap respectively; they show propensity to be located in α helix. The last group composed of PBs labelled from *g* to *j* mainly concerns the coils with a frequency more than 81.5% for the third C_α.

The average number of repeats (*anr*), i.e. the size of series composed by the same block, is estimated by the quantity $1/(1\text{-}\pi_{ii})$ where π_{ii} is the transition frequency of the *i*th block toward itself. It afford us a confirmation of the repetitive 3D structures :

(i) PB*m* (i.e. π_{ii}=85.2%) with an *anr* of 6.74 blocks corresponds exactly to the regular α-helices. In the same way, 78.1% of the third C_α for the learning database present in α-helices belongs to this block and 86.7% of the third C_α of this block is found in α-helices.

(ii) PB*d* (i.e. π_{ii}=63.5%) specifies β sheets with an average size of 2.74 blocks.

(iii) PB*c* and PB*e* have an *anr* higher than 1.1 since they corresponds to distorted β state C- or N-cap.

The labels of Table I help one to make a relationship with conventional 3-states alphabet with only three states (α-helix, β-strand and coil). For instance, PB*b* goes to PB*f*, a labelled "N-cap β" to a labelled "C-cap β" directly with 13.7% rate. In the same way, PB*m*, the labelled "α type" goes to PB*b*, a labelled "N-cap β" directly with 9.2% rate. The flexibility of the alphabet is higher than the label given shows it.

Dependency between protein blocks and sequences

The relationship of PBs with the amino acid sequence can be assessed by the occurrence matrices (i.e. the observed amino acid distribution in a given location of the sequence window associated with a PB).

Example

Figure 4 shows the 3D structures of four PBs (PB*p*, PB*b*, PB*d* and PB*m*) by superposition of backbone fragments extracted from the database with XmMol [51], the associated occurrence matrices, the Kullback-Leibler asymmetric divergence profiles (i.e. KLd profiles).

The blocks PB*m*, PB*p* and PB*d* have the lowest rmsd values (0.43Å, 0.46Å and 0.48 Å respectively) as shown by the backbone superposition. The PB*b* is slightly more variable (rmsd=0.51 Å) mainly due to extremities of the fragments.

The propensities of amino acid to be located in a given position in the window are accurately represented by the Z-scores (see Methods section). The dark rectangles (re-

spectively white) indicate the over-represented amino acids, i.e. Z > 4.4 (respectively under-represented, i.e. Z < - 4.4), the threshold corresponds to a type I error p less than 10^{-5}. Grey zones correspond to intermediate Z-score.

The analysis of the KLd profile (i.e. the dissimilarity between the observed amino acid distribution observed in a given position and that in the database) enables us to define the informative locations for a given block. The four KLd profiles are representative of the obtained profiles. They have been ordered according to the decrease of their KLd maxima.

The PB*p* is characterized at its central position, with an over-representation of Gly and Asn. As important is the number of under-representations like aromatic and hydrophobic residues. The KLd profile is a sharp peak at the central particular position (KLd = 0.55). The PB*b*'s KLd profile is a bell curve five times smaller than the previous one (KLd maximum = 0.08). Reversibly, we notice that the number of informative positions is increased in the range [-2:+2] and not only at the central position. The sequence specificity is faint due to a lower KLd magnitude nevertheless we observe opposite representations of Pro and Gly along the window.

For the block *d* corresponding to a regular β strand, the dissimilarity profile is different from the previous one: a maximum (KLd = 0.06) for the central C_α and symmetric decreases from this position. We observe strong differences of the Z-scores between inside and outside the structural block [-2:+2]. The over-representations concern mainly certain hydrophobic residues Ile and Val and slightly Phe, Tyr, Trp and Thr within the block, and under-representations mainly the polar residues (Lys , Arg, Asp, Gln and Asn) and outside the block for Gly and Pro [25].

PB*m* like PB*d* has specific KLd profile. Mainly the central positions contribute to the sequence specificity of PB*m*, a regular α-helix, and correspond to the five C_α (positions -2 to +2, KLd maximum = 0.05) of the block. We observe an over-representation of aliphatic amino acids Leu, Met, Ala and polar residues Gln, Glu, Arg, Asp and Lys. The under-representation of well-known α-helix breakers Pro and Gly is strong, as well for His and Asn on all the positions of the window [48, 47, 24].

The only other characteristic pattern (not shown) is a bimodal profile for certain PBs, PB*c* (positions -2 and +2), PB*e* (positions -1 and +1), PB*l* (positions -2 and 0), PB*k* (positions -1 and +1) and PB*p* (positions 0 and +2). Those informative positions are essentially breaks of the repetitive structures.

PBs Z-scores

Table II gives for each PB and each position in the window [-4;+4] the amino acid which have a Z-score computed >4.4 (noted +) or <-4.4 (noted -). A large number of

amino acids per position have significant Z-scores. To focus on the largest specificity, a rate of 4.4 had been chosen as in Figure 4. It could be noted than outside this window only one position over all the PBs is associated with a KLd value larger than 10% of their KLd maxima. The more informative positions have been defined using a threshold of 300 / $2N_k$, N_k denoted the observed number of PB_k (see section Propensities).

The main transitions between the PBs (see Table I for the transition rate) are found again in the amino acid compositions. For instance, the sequence specificity [Gly, Asn] is observed for PB *n*, *o*, *p* and *a* in positions (+2), (+1), (0) and (-1). In the same way, PB*e* and PB*g* in position +2, go to PB*h* (at the position +1) and PB*i* (at the position 0). However differences can be noticed like for the PB*d* which goes towards PB*f* with a rate of 51.9%, its position +1 has over-representation slightly different compared to PB*f* in position 0, for instance Val and Ile are over-represented in PB*d* at the position+1 and under-represented in PB*f* at the position 0. In fact, the propensities of the amino acids to be located in certain positions are not always conditioned by the transitions between PBs. For instance, the under-representation of Pro in position +1 of PB*b* is followed by PB*c* with a rate transition of 17.9% an over-representation of Pro in this latter block is observed, in position 0.

The over- and under-expressions are mainly concentrated in the central window [-2;+2]. Figure 4 shows an illustration of the real importance of each position.

We note that the repetitive structures show classical over- and under-expressions: $[AEL]^+$ / $[GPST]^-$ for PB*m* and $[IV]^+/[ADEGN]^-$ for PB*d* [25, 47, 24]. The over-expression of Gly is often accompanied by an Asn's over-expression in the coils.

Sequence families

The set of sequence windows encompassing a given protein block allows the computing of occurrence matrix used in the Bayesian approach. On the basis of the concept of *1 block - n sequence*, we have tried to split the sequence window set into subsets called "sequence families" showing differences of amino acid propensities. Some PBs have been divided in sequence families, mainly the most frequent (the integer between parentheses indicates the number of families): *m* (6), *d* (3), *c* (2), *f* (2) and *b* (2). The split PBs give families with an effective close to the other PBs (less than 5% of the database). We ought to control that the sequence families do not created a new fold pattern, i.e. they are always associate to one PB.

The average dihedral vectors associated with these new sequence families could be computed. We not they do not provide significant differences with the initial dihedral vectors associated with a given PB. The angular differences do not exceed 5^o As a result,

sequence families do not create new type of 3D structures, and occurrence matrices can be used for the PB prediction.

Figure 5 concerns the splitting PB*b*. The KLd profile of the original occurrence matrix is given with those of the two families (PB*b1* and PB*b2*).

The original KLd profile was low (values less than 0.1), PB*b*'s family *1* and *2* reach to 0.3; the information providing from each family is more sequence specific. By thresholding theses profiles at a level of 0.08, we observe that PB*b1* have the most sequence specific positions in the range [-3;+2], and for positions (-7) and (+4), and the second in [-2;+2]. The KLd profiles are different, the modes are in positions (+1) and (0) respectively.

Comparing the associated occurrence matrices assesses the amino acid composition differences : PB*b1* relative to PB*b2* shows an over-representations of Ala in position (-7), Asn (-2), Pro (-1), His and Asp (0), Pro (+1) and Phe (+6), and under-representations of Lys (-2), Gly (+1) and Cys (+4). It must be noticed that the main characteristics of the PB*b*'s occurrence matrices are conserved in the both families, as the over-expression of Pro in position (+2). We point out that the differences shown correspond to different reallocations of amino acids. For instance, Ala is over-represented in position (-7) in PB*b2*, though its frequency in PB*b1* is same as in the whole database.

Bayesian prediction of local structure

Protein block prediction

The prediction is carried out using the occurrence matrices with the Bayes' theorem. Here, the under- and over-representations given in Table II play the major role because they determine the quantity R_k for each PB. R_k is the product of frequencies of the amino acid observed at each position of the sequence window. Whereas, the "true PB" is geometrically defined by the dihedral vector, the "predicted PB" is simply defined by the rule of the highest R_k (or ln R_k), this only depends on the sequence window centred in a given site of the protein sequence. So the accuracy prediction is assessed by comparing the proportion of sites where the predicted and the true PBs are identical. This calculation is performed with the 1/3 of the database not used in the training step. The prediction rate is initially of 30.0% by using the sequence windows of 5 amino acids, i.e. the structural window of 5 C_α. The use of flanking sequences (a total window of 15 residues, i.e. a symmetrical elongation of 5 amino acids encompassing the structural window) allows a prediction rate of 34.4% (i.e. an gain of 4.4%). The gain is observed for all the PBs and is not specific of the repetitive secondary structure. For example, PB*b* increases from 11.0% to 13.5%, PB*e* from 33.0% to 43.2%, PB*i* from 32.9% to 42.2%, PB*p* from 26.9%

to 33.5%,

Prediction with family sequences

The prediction is then applied with the sequence families. Figure 6 shows the relationship between the initial prediction rate Q(1) (X-axis) and the difference Q(1)* - Q(1) between the initial prediction rate and new prediction rate when using the sequence families (Y-axis). The main result is: the gain is positive for more than 95% of the proteins and does not depend on the size of the proteins (i.e. the gain follows a Gaussian distribution, figure not shown), The average prediction rate increases to 40.7%. We notice that the prediction rate between the sets of proteins used in the learning and in the validation step only differs of 0.5%.

In addition, more 51.4% of the proteins had a prediction rate more than 40% instead of 20.5% in the first prediction. By using the protein classification rules of Michie *et al.* [27], we note that a 9.1% gain for the protein all-α (from 37.3% to 46.4%), the all-β had a lower gain (i.e. 3%, from 30.2% to 33.2%), the mixed α-β had a 4.9% gain (35.7% - 40.6%), and the unclassified a 4.8% gain (33.9% - 38.7%).

Four proteins have an high prediction rate : serine proteinase inhibitor (PDB code : 1cseIh, initial prediction rate : 47.9%, with sequence families : 64.6%), electron transfer (5rxnh, 56.4%, 59.0%), translational regulator protein (1regX, 59.8%, 63.5%), lipoprotein apolipoprotein*E3 (1lpe, 60.4%, 74.4%). Three first ones are all-α and the last one all-β.

The homogeneity of the prediction was another important point observed in the cutting out of the sequence families. When splitting the most frequent PBs, it is possible to obtained a better global prediction rate, but with regards to other PBs their rate will drop significantly. For instance, the addition of one supplementary family to PBm or PBd allows a global gain of more than 1%, but in parallel the prediction rate of PBb drops under 25.0%. So, after introducing sequence families, we found a range of prediction rates more reduced: initially from 13.1% (PBb) to 60.3% (PBa), and, then from 27.0% (PBb) to 53.2% (PBa).

Moreover, the gain is not correlated with the PB frequency in the database. Certain PBs of low occurrence frequency such as PBa (3.9% of the database), PBj (1.0%) and PBo (2.6%) are highly predicted with rates of 53.2%, 47.3% and 45.7% respectively.

The most frequent PBs which correspond to the cores of the secondary structures have a prediction rate of 50.6% for PBm (α-helix)and 34.6% for PBd (β-strand).

Prediction example

Table III gives the PB predictions of the 18 first sequence windows (15-residues window encompassing the protein block of 5 C_α) of the protein 2aak, ubiquitin conjugating enzyme. A regular α-helix of 10 PBm followed by a coil of 7 PBs leading to a β-sheet (see figure 3) is studied. This example includes the concept of sequence families previously described. Each line of the table corresponding to a sequence window. For instance, the 5th window centred on the motif MRDFK is assigned to PBm. The first three solutions (obtained after ordering the predictions scores R_k) given by the Bayesian approach are PBm, PBf and PBb, the respective scores are 22.13, 1.25 and 0.40. So the first score indicates that the probability of observing the PBm is 22.13 higher than the probability of observing this block without sequence information. For this position, the prediction is correct. The two first solutions can be explained by taking into account the proportions of amino acids to be located in certain positions of protein blocks (see Table II). The elevated scores of the first solutions are justified by the presence of amino acids Leu, Met, Arg, Lys, Arg and Leu respectively in position (-3), (-2), (-1), (+2), (+3) and (+4). In the same way, the block BPf is classified at the second rank because of a Asp as central position. By considering the first ranks, 10 protein blocks are correctly predicted among the 18 first blocks. For the whole protein, the prediction rate Q(1)* is 40.8%. Without taking into account sequence families, it was 30.4%, hence an appreciable gain. Usually, the prediction accuracy is only assessed from the solutions of the first rank. But by observing the solutions given by the three first ranks in Table III, we found again 17 of the 18 true PBs. The only misprediction corresponds to an end of α-helix which has a unusual amino acid composition in the central window [-4;+4], i.e. KRLQQDPPA. So rather than only considering the first ranks, an interesting approach consists of examining the accuracy Q(r) for a given rank r. The index *Neq* quantifies the dispersion of the scores. In the first part of the α-helix, it varies in the range [2.06;3.78] and it is correlated to an optimal prediction. Reversibly, so the probability of finding the true BP decreases at the end of the helix, while the *Neq* increases (more than 4.82), this reflects the presence of less informative sites. Moreover intermediate *Neq* values are observed for the 7 last residues, the optimal rank is mainly 2. Consequently, a strategy based on a multiple choice of PBs in each site should be informative.

Prediction strategies

Multiple choices per site

In the database, we observe a positive relationship between the highest score (R_k) and the prediction rate (Q(1)*). We have established that the true PB is often among the first selected PBs, i.e. the highest scores.

The number of selected ranks will be defined by the *Neq* index that reflects the dispersion of the 16 scores. Globally the prediction rate Q(4) for the 4 first solutions, i.e. the 4 PBs with the highest R_k, was 71.4% with the initial Bayesian approach, and then by using the family sequences, it increases to 75.8%. The prediction rates Q(4)* vary according to the PB type from 57.2% (PBg) to 82.8% (PBm). The repetitive PBd and PBm have prediction rates more than 80% with this strategy. Therefore, prediction strategies based on a multiple choice per site should be able to improve the accuracy.

The analysis of the prediction results allows one to point out that the true PB is generally present among the predicted PBs showing the highest scores. So two prediction strategies can be elaborated :

(i) a "global strategy" : to define the number r_s of blocks to be kept in every sites of a protein in order to obtain an user-fixed prediction rate Q_g, i.e. the true protein blocks is found again among the r_s selected PBs with a probability Q_g.

(ii) a "local strategy" : to characterize a set of sites for which the user-fixed prediction rate Q_l is obtained for a fixed number r of PBs.

In the first strategy, the entire protein is predicted in terms of protein blocks, but the combination of selected blocks is variable along the protein. In the second strategy, the protein is restricted to certain sites whose the prediction should be correct with a given probability by taking r PBs.

Global strategy

With the use of the learning database, we have established a relationship between the probability of finding the true PB among the r selected solutions and the diversity of score values assessed by *Neq* (from an entropy assessed from scores normalized into probabilities).

First, we have selected the blocks for which the *Neq* values vary in a given range. Then the distribution of the rank of the true PB in the solutions has been calculated. So, for a given prediction rate Q_g, we have determined from those distributions how many solutions must be selected for every *Neq* range (figures not shown). For instance, for a *Neq* value in the range [1.0;6.32], we selected the 3 first PBs given by the Bayesian approach in order

to obtain a 70% prediction rate.

Figures 7 show the results of this strategy for the studied protein 2aak. The *Neq* profile (Figure 7a) gives the variation of the index (values between 1.06 et 9.79). Figure 7b gives the rank of the true PB in every site of the protein. We note that the true PB is found again in the 3 first solutions for most of the sites (77.8%). Certain restricted zones of the protein need a higher number of selected PBs, such as the two coils separating the 3 first β-sheets (positions 22 to 46), and the large coil (positions 82 to 90) containing a small α-helix.

The profile of Figure 7c indicates the number of blocks to be selected with an average prediction rate Q_g of 75%, the two dot series below this graph correspond to the sites where the PB is correctly predicted at the first rank and when the true PB is found again in the selected blocks. The maximal number of PBs is 4. The prediction rate is initially 40.7% with only the first rank; for Q_g=75%, we keep 1 to 4 PBs for 8, 17, 37 and 72 sites respectively. The repetitive structures and the blocks close to them (cf. Figure 3) are correctly delimited. On the other hand, the coils are more difficult to be defined. The comparison between the two series of points shows that the zones with some true PBs found at the first rank can be spread easier than the zones without true PBs at the first rank.

This strategy leads to an excess of selected blocks in each position. However, it selects in every site a number of blocks so that a given prediction accuracy is ensured.

Local strategy

The second strategy consists of defining limited zones in the protein where the prediction is guaranteed by a number r of selected blocks with a rate Q_l. Figure 8 shows the evolution of the prediction rate computed function of the *Neq* index and for different values of r (r varies from 1 to 5). For building these curves, we selected in the whole proteins of the learning database the sites where the *Neq* index is less than a given value, so the corresponding proportions of sites is assessed. We then computed the proportion of sites in which the true PBs are found again among the first solution (r=1), the two first solutions (r=2) and so on. So for example, we limit our prediction to 70% of protein sites thus *Neq* value must be less than 4.8 (see Figure 8). If we only selected one (i.e. 2, 3 and 4) rank(s), we should obtain an average prediction rate of 46.8% (respectively 63.4% 73.1% and 79.6%). We could note too for a given prediction rate, for instance 80%, that we will have a *Neq* index of 1.28 and 5.5% of the population for the first rank (i.e. 1.66 and 11.5% for the second rank, 2.61 and 26.9% for the third, 4.64 and 66.3% for the fourth).

Figure 7d indicates for the protein 2aak, the zones for the prediction rate Q_l of 75% and by taking the 3 first solutions. The corresponding *Neq* is less than 5.11. As 62 sites had be selected, the dot series indicates the 49 positions where a true PB had been found again in the selected positions. So, the observed prediction rate is 79% for 46.3% of the protein positions. By comparison with the first approach, it is clear that take 3 possibilities is sometimes an excess. In the same manner, with r=4 and Q_l=75% (data not shown) 72 sites (52% of the protein) are concerned.

Figure 7e shows the same strategy with Q_l=70% and r=3 ranks. By taking a *Neq* maximal of 6.32, 122 sites (91% of the proteins) were selected and 95 sites give the true PBs among the three proposed blocks, which corresponds to an observed prediction rate of 77.9%.

So, this strategy allows one to locate the sites of high predictability, however a critical research must be made in the choice of the number r of ranks to be selected. For example, for a prediction accuracy Q_l=70%, the proportion of selected sites in protein rises dramatically with the change of selection, 2 into 3 blocks. This produces a site increases of 49%. In a further application of this strategies, in the field of *ab initio* modelling, the choice of the number selected blocks per site raises certain problems such as: increasing the rank r leads to a larger covering of the protein, but also to a higher combinatory between blocks for building a molecular model from protein sequence.

Discussion

In our study, we have defined a structural alphabet, which allows the local approximation of the 3D protein structure. We have used this library of fragments (PBs) in a new Bayesian probabilistic prediction approach. We have then developed two types of strategies, consisting not only in looking for the most probable block for every protein sequence position, but also in searching for the optimal blocks to be conserved per site with a given prediction accuracy.

Structural alphabet

The first important parameter involved in this study is the length of the PB. Clearly depending on the authors, the length maybe variable or fixed. A constant size is a simple approach to perform the prediction step. Long segments must need much more blocks to have the same structural description. Different studies have been carried out using various segment sizes: from 4 to 7 for Rooman *et al.* [34], from 7 to 19 for Bystroff and Baker [20], or fixed at 8 for Unger *et al.* [32, 52], 9 for Schuchhardt *et al.* [33], 6 for Fetrow *et al.* [35], or 4 for Camproux *et al.* [36]. We have chosen a size of 5 C_α that is

enough convenient to conserve the local contacts within the regular structures: positions (i, $i+2$) for the β-strand and (i, $i+3$, $i+4$) for the α-helix.

The second parameter, partially related with the first one, is the number of PBs. It is critical to learning and prediction steps. Relative to the works of Unger *et al.* [32, 52] and Schuchhardt *et al.* [33], we have determined a limited number of small 3D local structures. Our choice is guided by two facts : (i) the precision of the protein 3D-structure description (expressed into *rsmda*) which follows directly from this choice, and (ii) the prediction accuracy which is likewise dependent on this number. In fact, smaller the number of PBs is, more the average *rmsda* increases (i.e. the dihedral angles are less well approximated). Inversely, using a higher number of PBs for prediction should result in an decreasing prediction accuracy. To assess the levels of relationship between the parameters : number of PBs, *rmsda* and Q(1)-value (i.e. prediction accuracy at the first rank), we have carried out a training allowing a large reduction of PBs by taking a high threshold r_0 in the structural similarity criterion. With an initial number of 34 PBs and after 8 shrinking processes, we have obtained 10 PBs. The four first shrinking show a fast decreasing of the number of PBs with a slow reduction of *rmsda*: it dropped from 34 PBs (with an average *rmsda* of 25.4^o) to 22 (28.5^o), then 19 (29.0^o) and 16 (30.0^o). The following processes permit to obtain successively 14, 12, 11, and 10 PBs at the end, the angular approximation remains stable between 30^o-32^o in these last steps.

By using a Bayesian prediction without the splitting into sequence families and only based on a structural window of 5 amino acids, the accuracy decreases dramatically from 34.6% to 30.0% and 22.7% with 10, 16 and 34 BPs respectively. As expected, the prediction accuracy is largely controlled by the number of PBs.

It must be emphasised that for a lower number of PBs (less than 12), the training only yields classical secondary structures completed by their caps. As the coil regions of the protein are more variable than the classical α-helix and β-strand secondary structures, and associated with fold patterns less frequent, it is necessary to select a higher number of PBs. Consequently, the choice of 16 BPs is consistent with a suitable balance between a correct approximation of the 3D structures (i.e. an average *rmsda* = 30^o) and an acceptable initial prediction accuracy.

Training approaches

Different types of methods have been used for classifying the 3D segments into a limited set of fold patterns. Various approaches in the field of the training have been applied: hierarchical clustering [20, 32, 34, 52], neural networks [35], self-organized maps (SOM) [33] or hidden Markov model (HMM) [36]. Apart the last work, the sequential dependence between the protein blocks does not take into account as constraints in the training.

Another advantage of our unsupervised classifier must be highlighted relative to the approach HMM [37]: no hypothesis about the probability laws of parameters is needed in the model. In fact, it is a nonparametric model. More we have introduced a shrinking process allowing a fast selection of the optimal number of PBs according to a given threshold of structure similarity. The algorithm is suited to a fast and efficient training of multiple signals (in our study, the dihedral vectors along the proteins). A definition of the protein blocks partially embedded for ensuring the continuity of the protein backbone is of interest in light of its potential application in the building of molecular models from protein sequence.

Prediction et Strategies

The Bayesian probabilistic approach has been frequently used [9, 38, 39]. Relative to neural network approach, the scores directly reflect the sequence information content expressed from the amino acid composition in every site of the sequence window. It is not only limited to a PBs ordering. It gives the possibility to compute the Kullback-Leibler asymmetric divergence profiles (i.e. KLd profiles), which point out the most informative amino acid positions in the sequence window.

Concerning the prediction approach, only rare works in the literature have been carried with such a structural alphabet. Two values for the prediction accuracy are actually available. The first one in the range 65 - 75 % [4, 5, 7] is obtained with a 3-states alphabet (classical secondary structure prediction). The second ones is the value given by Bystroff and Baker [20], a prediction close to 50% with a 13-states alphabet (in reality 13 clusters of fragments of various lengths). It is trivial to see that a difference between a prediction with 3 or 16 blocks cannot give the same level of accuracy. Bystroff and Baker [20] have used a method which finally gives 13 clusters of fragments of different length which are longer than our PBs of 5 C_α, 50% accuracy for a 13-states alphabet is close to our results with a 16-states alphabet.

The developed concept of a fuzzy sequence / 3D-structure model (*1 fold - n sequences / 1-sequence - n folds*) is really important in the elaboration of a structural model. The two parts of the fuzzy model have employed "*1 fold for n sequences*" in the definition of the "sequence families" and "*1 sequence for n folds*" in the both strategies.

The success of the *ab initio* approach are actually very limited to only local prediction [15] or limited to polypeptide folds prediction [14]. Our approach that leads to a global prediction solution has to be considered as a very powerful initial step coupled with a more elaborated method using for instance realistic physical force. For summarising, two strategies have been built: the first gives a set of potential blocks along the sequence. The second one consists of giving a constant number of blocks for limited zones. The only

assumption is that the true PB is among the few selected ones, (this is globally true). It must be noticed as an important point than for each site, the most 4 likely PBs are sufficient to provide a high level of prediction rate (more than 75%).

References

[1] Govindarajan S, Goldstein RA. Why are some proteins structures so common ?. Proc Natl Acad Sci USA 1996;93:3341-3345.

[2] Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins 1999;35:408-414.

[3] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97-120.

[4] Garnier J, Gibrat J-F, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540-553.

[5] Rost B. PHD : predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266:525-539.

[6] Chandonia J-M, Karplus M. The importance of larger data sets for protein secondary structure prediction with neural networks. Protein Sci 1996;5:768-774.

[7] Chandonia J-M, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293-306.

[8] Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. J Mol Biol 1997;268:31-36.

[9] Thompson MJ, Goldstein RA. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. Protein Sci 1997;6:1963-1975.

[10] Kawabata T, Doi J. Improvement of protein secondary structure prediction using binary word encoding. Proteins 1997;27:36-46.

[11] Frishman D, Argos P. The future of protein secondary structure accuracy. Fold Des 1997;2:159-162.

[12] Defay T, Cohen FE. Evaluation of current techniques for *ab initio* protein structure prediction. Proteins 1995;23:431-445.

[13] Yue K, Dill KA. Folding proteins with a simple energy function and extensive conformational searching. Prot Sci 1996;5: 254-261.

[14] Derreumaux P. A diffusion process-controlled Monte Carlo method for finding the global energy minimum of a polypeptide chain. I. Formulation and test on a hexapeptide. J Chem Phys 1997;106:5260-5269.

[15] Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;suppl.3:149-170.

[16] Bowie JU, Lűthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253;164-169.

[17] Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. Protein Sci 1996;5:947-955.

[18] Sali A, Blundell TL. Compartative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779-815.

[19] Jaroszewski L, Rychlewski L, Zhang B, Godzik A. Fold prediction by a hierarchy of sequence, threading, and modeling methods. Prot Sci 1998;7,1431-1440.

[20] Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motif. J Mol Biol 1998;281:565-577.

[21] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82-95.

[22] Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;suppl.3,171-176.

[23] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209-225.

[24] Kumar S, Bansal M. Geometrical and sequence characterisitics of α-helices in globular proteins. Biophysical Journal 1998;78,1935-1944.

[25] Hutchinson E, Thornton JM. A revised set of potentials for β-turn formation in protein. Prot Sci 1994;3,2207-2216.

[26] Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Protein Eng 1993;6:485-500.

[27] Michie AD, Orengo CA, Thornton JM. Analysis of domain structural class using an automated class assignement protocol. J Mol Biol 1996;262:168-185.

[28] Boutonnet NS, Kajava AV, Rooman MJ. Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. Proteins 1998;30,193-212.

[29] Wintjens RT, Rooman MJ, Wodak SJ. Automatic classification and analysis of alpha alpha-turn motifs in proteins. J Mol Biol 1996;255, 235-253.

[30] Kwasigroch J-M, Chomilier J, Mornon J-P. A global taxonomy of loops in globular proteins. J Mol Biol 1996;259,855-872.

[31] Wodjick J, Mornon J-P, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289,1469-1490.

[32] Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355-373.

[33] Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng 1996;9:833-842.

[34] Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990;213:327-336.

[35] Fetrow JS, Palumbo MJ, Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. Proteins 1997;27:249-271.

[36] Camproux AC, Tuffery P, Chevrolat J-P, Boisvieux J-F, Hazout S. Hidden Markov Model approach for identifying the modular framework of the protein backbone. Protein Eng 1999;12:1063-1073.

[37] Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc of the IEEE 1989;77:257-285.

[38] Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. Proteins 1996;25,38-47.

[39] Lathrop RH, Rogers Jr. RG, Smith T, White JV. A bayes-optimal sequence-structure theory that unifies protein sequence-structure recognition and alignement. Bulletin of Mathematical Biology 1998;60,1039-71.

[40] Sternberg MJ, Islam SA. Local protein sequence similarity does not imply a structural relationship. Protein Eng 1990;4,125-131.

[41] Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative protein data sets. Protein Sci 1992;1:409-417.

[42] Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522-524.

[43] Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon J-P. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 1993;6:377-382.

[44] MacArthur MW, Thornton JM. Deviations from planarity of peptide bond in peptides and proteins. J Mol Biol 1996;264:1180-1195.

[45] Kohonen T. An introduction to neural computing. Neural Networks 1989;1:3-16.

[46] Kohonen T. Self-Organizing Maps. (2nd edition) Springer series in information sciences vol.30. Berlin : Springer-Verlag; 1997. 376 p.

[47] Presta LG, Rose GD. Helix signals in proteins. Science 240,1632-1641.

[48] Richardson JS, Richardson DC. Amino acid preferences for specific locations at the end of α helices. Science 1988;240,1648-1652.

[49] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22:79-86.

[50] Kraulis JP. MOLSCRIPT: A Program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24,946-950.

[51] Tuffery P. XmMol: An X11 and Motif program for macromolecular visualization and modelling. J Mol Graph 1995;13,67-72.

[52] Unger R, Sussman JL. The importance of short structural motifs in protein structure analysis. J Comput Aid Mol Des 1993:7,457-472.