A Structural Alphabet for Local Protein Structures: Improved Prediction Methods

Catherine Etchebest, Cristina Benros, Serge Hazout, and Alexandre G. de Brevern*
Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U726, Université Denis DIDEROT—Paris, France

ABSTRACT **Three-dimensional protein structures can be described with a library of 3D fragments that define a structural alphabet. We have previously proposed such an alphabet, composed of 16 patterns of five consecutive amino acids, called Protein Blocks (PBs). These PBs have been used to describe protein backbones and to predict local structures from protein sequences. The Q_{16} prediction rate reaches 40.7% with an optimization procedure. This article examines two aspects of PBs. First, we determine the effect of the enlargement of databanks on their definition. The results show that the geometrical features of the different PBs are preserved (local RMSD value equal to 0.41 Å on average) and sequence–structure specificities reinforced when databanks are enlarged. Second, we improve the methods for optimizing PB predictions from sequences, revisiting the optimization procedure and exploring different local prediction strategies. Use of a statistical optimization procedure for the sequence–local structure relation improves prediction accuracy by 8% (Q_{16} = 48.7%). Better recognition of repetitive structures occurs without losing the prediction efficiency of the other local folds. Adding secondary structure prediction improved the accuracy of Q_{16} by only 1%. An entropy index (N_{eq}), strongly related to the RMSD value of the difference between predicted PBs and true local structures, is proposed to estimate prediction quality. The N_{eq} is linearly correlated with the Q_{16} prediction rate distributions, computed for a large set of proteins. An "expected" prediction rate Q^E_{16} is deduced with a mean error of 5%. Proteins 2005;59:810–827.**

Key words: structure–sequence relationship; probabilistic approach; Bayes' rule; secondary structure; protein blocks; ab initio

INTRODUCTION

Protein folds are often described as a succession of secondary structures. Their repetitive parts (α-helices and β-strands) have been intensively analyzed[1,2] since their initial descriptions by Pauling and Corey.[3,4] Because defining the rules for secondary structure assignments is not trivial, assignment methods based on different criteria have emerged. The greatest discrepancies are found mainly at the caps of the repetitive structures. These differences, even small, can result in different lengths for the repetitive structures, depending on the algorithm used.[5] In addition, a classification limited to three states (the classical repetitive secondary structures and coils) does not allow the protein structures to be described precisely at the 3D level, because it omits the relative orientation of connecting regions. The coil state, which represents 50% of all residues, corresponds to a large set of distinct local protein structures.

In the past few years, these observations have led to a new view of 3D protein structures. They are now thought to be composed of a combination of small local structures or fragments, also called prototypes. The complete set of prototypes defines "a structural alphabet."[6,7] Different teams have described these local protein structures according to different criteria. The clustering of distinct protein fragments is based on similarity measures that use different geometric descriptors (C_α coordinates, C_α distances, α or φ, ψ dihedral angles) and/or different algorithms (hierarchical clustering, empirical functions, Kohonen maps, artificial neural networks, or hidden Markov models). For example, Levitt's group[8,9] and Micheletti and coworkers[10] developed libraries optimized for the reconstruction of the global fold, while other groups have focused on interesting sequence specificities of these fragments that are useful for prediction.[11,12]

Choices depend on the objectives: a large set of fragments[12–16] is needed for precise description, while a limited set[11,17–23] is more useful for prediction purposes. Similarly, a delicate balance is required between a number of states sufficient for correct approximation of local backbone structures and a number small enough for extraction of relevant sequence–structure relations for

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

The first two authors contributed equally to this article.

Grant sponsor: the Ministère de la Recherche; Grant sponsor: "Action Bioinformatique inter EPST"; Grant numbers: 4B005F and 2003-2004 ("Outil informatique intégré en Génomique Structurale; Vers une prédiction de la structure tridimensionnelle d'une protéine à partir de sa séquence"); Grant sponsor: the Fondation de la Recherche Médicale (to A.d.B.); Grant sponsor: the Ministère de la Recherche (to C.B.).

*Correspondence to: Alexandre G. de Brevern, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U726, Université Denis DIDEROT—Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France. E-mail: debrevern@ebgm.jussieu.fr

Received 19 July 2004; Accepted 14 December 2004

Published online 8 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20458

adequate predictions. The optimal number of states also depends on the number of residues (residue length) of each letter composing the alphabet. For letters four to six residues long, an alphabet of 10 to 20 states can correctly both describe and predict structure (http://www.ebgm.jussieu.fr/~debrevern/).[19,21] The different structural alphabets developed recently have proved efficient in describing and predicting small loops[24–26] and long fragments.[27–29] One of the most interesting structural alphabets is that of Bystroff and Baker (I-Sites), which, together with a sophisticated procedure for 3D reconstruction, has been used with great efficiency to improve de novo prediction methods.[30–32]

Somewhat different but also interesting is the folding building blocks model proposed by Nussinov's group. Unlike alphabet approaches based mainly on 3D similarities between fragments, Nussinov and her collaborators focus on the elementary folding units that lead through a hierarchical process to the folded state.[33,34] These folding units are obtained from a progressive and hierarchical dissection based mainly on native 3D interactions in the folded state. The process concludes with fragments of variable length (at least 13 residues), and may be used to engineer new naturally occurring folds with low homology to existing proteins.[35] Such elementary folding units have been used in a homology prediction strategy. A target sequence is compared with and aligned to the building block sequences in the database. A graph approach then assigns the building block automatically to the query sequence. Before this method can be used for ab initio structure prediction purposes, further exploration needs to establish position-specific sequence–structure relations from these elementary folding units. This methodology, although it may constitute an alternative to structural alphabets, is clearly based on a distinct approach. Any direct comparison between these folding units and the fragments defined by a structural alphabet is less than straightforward.

Our structural alphabet is composed of 16 average protein fragments, five residues in length, which we call Protein Blocks. These PBs have been used both to describe 3D protein backbones and to predict local structures.[21,36,37] Karchin and coworkers have compared the features of this alphabet with those of eight other structural alphabets. Their results show clearly that our PB alphabet is highly informative, with the best predictive ability of those tested.[38]

These promising methods nonetheless need to be improved to strengthen their relatively weak prediction rates. These rates decrease rapidly when the number of states increases. For example, SSPRO,[39] a method using artificial neural networks coupled with sequence homology, has a predictive rate of 80% for three states and 62% for six states defined by DSSP.[40] Two other states with very few occurrences cannot be predicted.

This article explores two aspects of our structural alphabet: the features of the 16 PBs that compose it and the efficiency of prediction methods based on it after various improvements. First, the continuous growth of the Protein DataBank (PDB[41]) allows ongoing refinement of the sequence–structure relations of the PBs. To examine these effects, we compare the PBs with the classical secondary structures and with different turns. We also explore the presence of tight turns (by definition, short fragments) in some particular long PB series. Second, we propose, assess, and compare two possible improvements in our protein local structure prediction method. The first, based on Bayes' rule, uses a highly refined version of the concept of Sequence Families (SFs) that we defined earlier.[21] The second, which takes advantage of the high prediction rate for secondary structures, involves a novel strategy that combines information related to the secondary structure prediction and the PB prediction.

MATERIALS AND METHODS

Datasets

This study considers five sets of proteins. We have already used the first four in recent work:[26,36] *PAPIA* from the PDB-REPRDB database,[42] the *PDBselect* databank,[43] *PISCES*,[44] and *SCOP–ASTRAL*.[45,46] We preferentially used the *PAPIA* set, composed of 717 protein chains and 180,854 residues. The set contains proteins with no more than 30% pairwise sequence identity, X-ray crystallographic resolutions better than 2.0 Å, and an *R*-factor less than 0.2. Each structure selected has an RMSD value greater than 10 Å between every representative chain. An updated dataset is defined from the PDB–REPRDB database[42] with the same criteria as *PAPIA*. It comprises 1407 protein chains and 293,507 residues. The amino acid composition is not significantly different for these two *PAPIA* protein sets. Each chain was carefully examined with geometric criteria to avoid bias from zones with missing density.

PBs

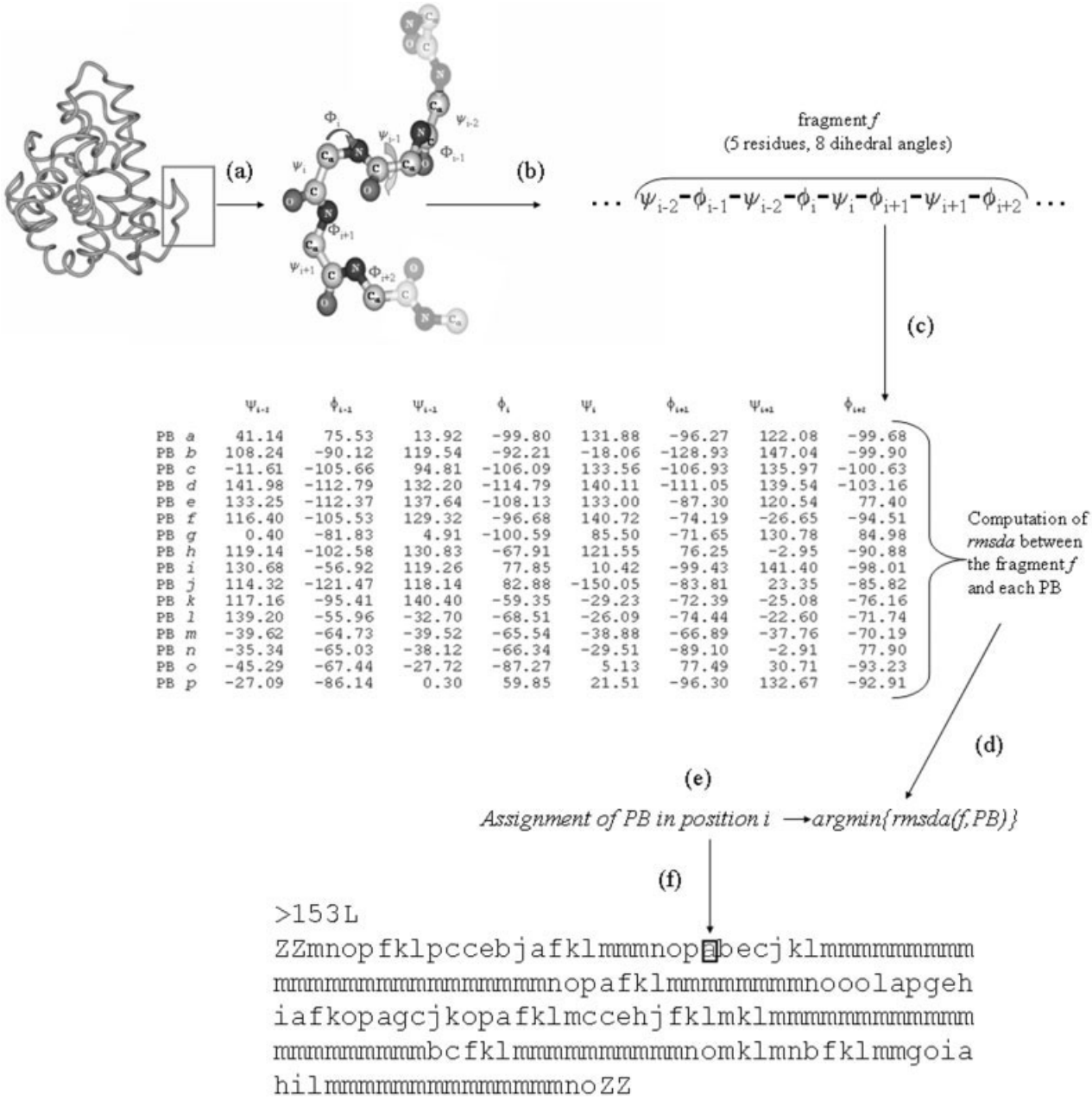
The structural alphabet is composed of 16 local prototypes called "Protein Blocks" (PBs, see supplementary data 1). They are overlapping fragments, $M = 5$ residues in length, encoded as sequence windows of 2 ($M - 1$) consecutive dihedral angles (ψ, ϕ), that is, 8 (values given in supplementary data 2). They were obtained by an unsupervised classifier similar to Kohonen maps[47,48] and hidden Markov models.[49] PBs *m* and *d* correspond to the prototypes for the central α-helix and the central β-strand, respectively. PBs *a* through *c* primarily represent β-strand N-caps and *e* and *f*, C-caps. PBs *g* through *j* are specific to coils, *k* and *l* to α-helix N-caps, and *n* through *p* to α-helix C-caps.[21,36]

Distance Criteria

The root-mean-square deviation (or RMSD) is computed as the average Euclidean distance between superimposed C_α.[50] The RMSD on angular values (or RMSDA) is defined as the Euclidean distance of dihedral angles (ψ, ϕ).[16]

Protein Coding

Protein structures are encoded as sequences of ϕ–ψ dihedral angles. They are cut into consecutive overlapping fragments, each M amino acids in length [Fig. 1(a)]. A fragment f is defined by a signal of $2(M - 1)$ dihedral angular values [Fig. 1(b)]. The fragment signal is compared with each PB signal with the RMSDA measure [Fig.



	ψ_{i-2}	ϕ_{i-1}	ψ_{i-1}	ϕ_i	ψ_i	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}
PB *a*	41.14	75.53	13.92	-99.80	131.88	-96.27	122.08	-99.68
PB *b*	108.24	-90.12	119.54	-92.21	-18.06	-128.93	147.04	-99.90
PB *c*	-11.61	-105.66	94.81	-106.09	133.56	-106.93	135.97	-100.63
PB *d*	141.98	-112.79	132.20	-114.79	140.11	-111.05	139.54	-103.16
PB *e*	133.25	-112.37	137.64	-108.13	133.00	-87.30	120.54	77.40
PB *f*	116.40	-105.53	129.32	-96.68	140.72	-74.19	-26.65	-94.51
PB *g*	0.40	-81.83	4.91	-100.59	85.50	-71.65	130.78	84.98
PB *h*	119.14	-102.58	130.83	-67.91	121.55	76.25	-2.95	-90.88
PB *i*	130.68	-56.92	119.26	77.85	10.42	-99.43	141.40	-98.01
PB *j*	114.32	-121.47	118.14	82.88	-150.05	-83.81	23.35	-85.82
PB *k*	117.16	-95.41	140.40	-59.35	-29.23	-72.39	-25.08	-76.16
PB *l*	139.20	-55.96	-32.70	-68.51	-26.09	-74.44	-22.60	-71.74
PB *m*	-39.62	-64.73	-39.52	-65.54	-38.88	-66.89	-37.76	-70.19
PB *n*	-35.34	-65.03	-38.12	-66.34	-29.51	-89.10	-2.91	77.90
PB *o*	-45.29	-67.44	-27.72	-87.27	5.13	77.49	30.71	-93.23
PB *p*	-27.09	-86.14	0.30	59.85	21.51	-96.30	132.67	-92.91

Fig. 1. Protein coding in terms of Protein Blocks. (**a**) Each protein residue is translated into dihedral angles (φ, ψ). (**b**) Fragment *f*, five residues in length and centered in position *i*, is selected as a succession of eight dihedral angles (from ψ_{i-2} to φ_{i+2}). (**c**) The RMSDA is computed between each PB and fragment *f*. (**d**) The minimal RMSDA value is selected, and (**e**) allows assignment of the PB in position *i*. (**f**) All the protein fragments are coded according to this approach.

1(c)]. The lowest RMSDA value for the $2(M - 1)$ angles [Fig. 1(d)–(e)] determines the PB assignment.[21,36] A PB is assigned to each position along the sequence except for the two residues on each end [Fig. 1(f)].

Secondary Structure and Turn Assignments

Two different techniques for assigning secondary structures (α-helices, β-strands, and coils) were used: STRIDE[51] and PSEA.[52] Turns were also assigned (according to the definitions in supplementary data 2). A turn is defined as n consecutive residues (denoted from i to $i + n - 1$) for which the distance between the C_α atoms of the ith and $(i + n - 1)$th residues is less than 7 Å and the two central residues are not helices. The turns are classified according to their ϕ–ψ dihedral angles. We explored γ-turns ($n = 3$),[53] β-turns ($n = 4$),[34] and α-turns ($n = 5$).[55]

Transitions

Transition frequencies $T(x,y)$ between successive PBs x and y are based on the conditional probability that PB y is

in sequence position $i + 1$ when PB x is in position i, that is, $P(PB_y|PB_x)$. We focused on the transition from one PB to specific other PBs. $T(x,y)$ is finally defined as follows:

$$T(x,y) = \frac{P(PB_y \mid PB_x)}{1 - P(PB_x)} \quad (1)$$

where $1 - P(PB_x)$ is the probability of a PB different from PB x.

Similarly, we defined the average number of repeats, $anr(PB_x)$, that is, the size of series composed by repetitions of the same block, as:

$$anr(PB_x) = \frac{1}{1 - P(PB_x \mid PB_x)} \quad (2)$$

where $P(PB_x|PB_x)$ is the transition frequency from PB x to PB x.

Position-Specific Sequence Matrices

Once the databank is encoded as PBs, sequence specificity may be computed. Each PB is associated with a set of sequence windows. We considered an enlarged sequence window $[-w; +w]$ of length l, with $w = 7$: sequence window length (l) thus equals 15. An amino acid frequency matrix for each PB of dimension $20 \times l$ is computed.

The amino acid occurrences for a PB have been normalized into a Z-score, defined as:

$$Z(i,x) = \frac{n_{\text{obs}}(i,x) - n_{\text{th}}(i,x)}{\sqrt{n_{\text{th}}(i,x)}} \quad (3)$$

with $n_{\text{obs}}(i,x)$ the number of observations or occurrences of amino acid i in PB x, and $n_{\text{th}}(i,x)$ the expected number of occurrences:

$$n_{\text{th}}(i,x) = N_x f_i \quad (4)$$

with N_x and f_i denoting, respectively, the number of occurrences of PB x and the frequency of amino acid i in the entire databank. This calculation assumes that the frequency of a given amino acid is not correlated with the PB type, that is, hypothesis of independence. Positive Z-scores that exceed a user-fixed threshold ε (respectively negative, are less than $-\varepsilon$) correspond to overrepresented amino acids (respectively underrepresented).

Classical Prediction

The goal is to use the local structure alphabet to predict the optimal PB for each position along a target protein sequence, with a probabilistic approach similar to that proposed in previous works (cf. Fig. 2).[21,36,37] The probability of observing a protein block x given a sequence window X_s centered in a site s is computed according to Bayes' rule as:

$$P(PB_x \mid X_s) = \frac{P(X_s \mid PB_x) \cdot P(PB_x)}{P(X_s)} \quad (5)$$

where $P(PB_x)$ is the probability of observing PB_x in the databank and $P(X_s)$ is the product of the frequency of each of the amino acids in the databank (assuming the residues are independent). $P(X_s|PB_x)$ is the conditional probability of observing the chain X_s of l amino acid residues, given a known protein block x. It is computed as the product of the frequency of the amino acids in chain X_s for a given PB_x:

$$P(X_s \mid PB_x) = \prod_{j=-w}^{j=+w} P(aa_i^j \mid PB_x) \quad (6)$$

with aa_i^j the amino acid i in position j of the sequence window X_s, l amino acid residues in length. The best PB, denoted PB*, for sequence window X_s is selected by maximizing the log likelihood ratio between $P(PB_x|X_s)$ and the probability of observing PB x without sequence information $P(PB_x)$:

$$R_x = \ln \frac{P(PB_x \mid X_s)}{P(PB_x)} = \ln \frac{P(X_s \mid PB_x)}{P(X_s)} \quad (7)$$

This ratio is easily computed with equation 5. PB* is given by $x^* = \text{argmax}\{R_x\}$. Each protein site is predicted independently.

Learning and Validation Sets

The different learning approaches were trained on a learning set composed of 2/3 of the nonredundant databank (450 proteins corresponding to 91,649 PBs and 97,949 amino acids). Using the occurrence matrices described above, we applied the different prediction strategies described below to each sequence in the validation set, which included 1/3 of the proteins in the databank (225 proteins corresponding to 45,154 PBs and 48,304 amino acids). To avoid possible bias between the learning and the validation sets that might significantly affect the subsequent learning procedures, we applied two different criteria to assess validation: (1) equivalent PB distributions and similar amino acid frequencies for both sets; (2) an equivalent and balanced representation of the structural protein classes in each set, that is, all-α, all-β, and other[56] (α + β and α/β are grouped together). The all-α proteins represent 13% of the databank, the all-β proteins 15% and the "other" group 72%. A third criterion was applied a posteriori: the difference in the Q_{16} value (see below) between the learning set and the validation set must not exceed 0.1%.

Prediction Assessment

Accuracy Q_{16}

To assess the predictions, we compute the accuracy Q_{16}, that is, the proportion of PBs correctly predicted. This value is equivalent to the Q_3 value for the secondary structures (only three states).

Confidence index

The prediction yields scores associated with each protein sequence position. These scores, R_x, are translated into probabilities, S_x, [see Eq. (8)] and then the Shannon entropy H[57] is computed [see Eq. (9)].

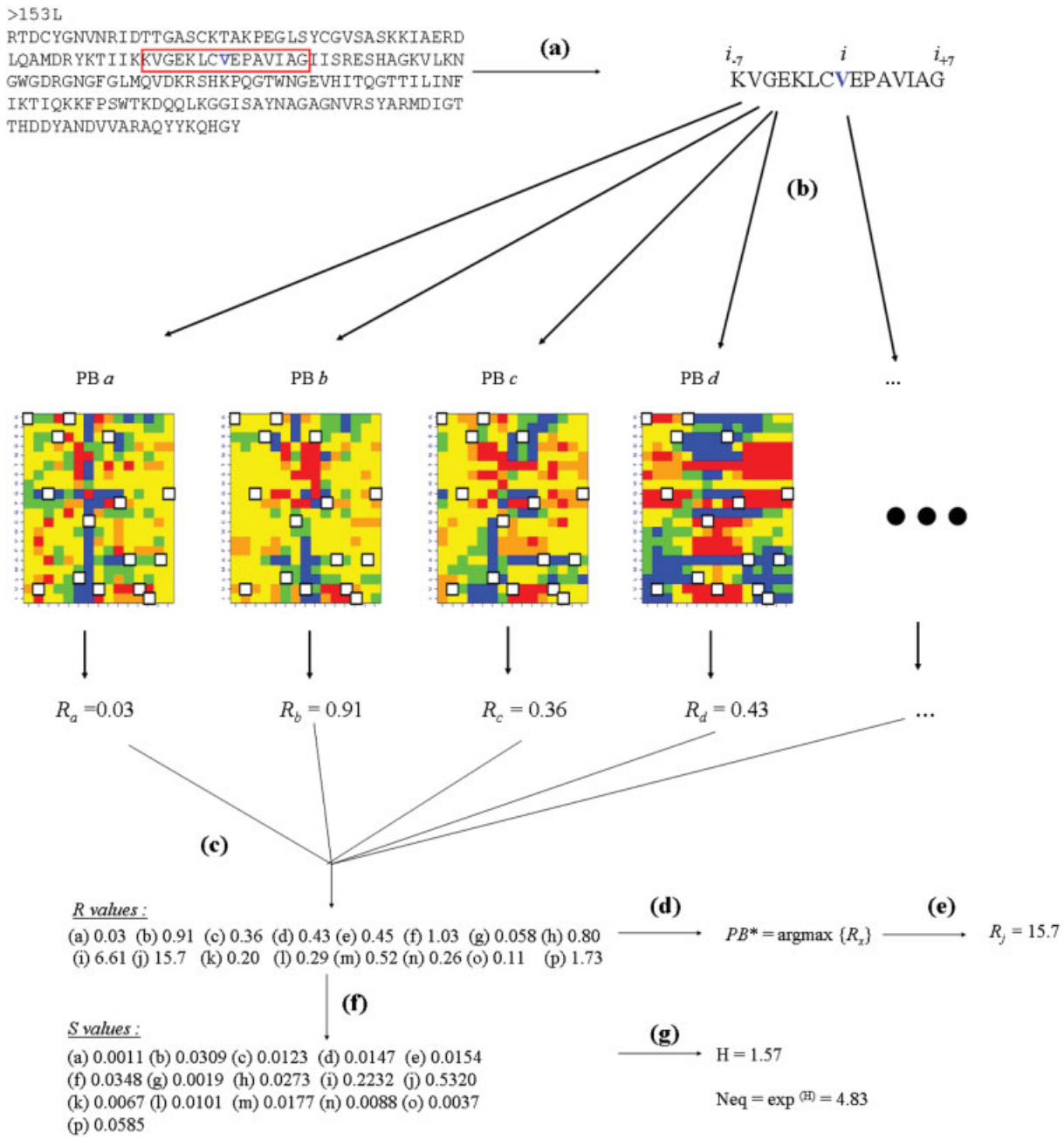


Fig. 2. Prediction. (**a**) The sequence window *f*, 15 residues in length and centered in position *i*, is selected. (**b**) Bayes' rule is used to compute scores for each PB. (**c**) The 16 *R* scores are analyzed. (**d**) The winning PB is associated to the maximal score. (**e**) In this example, it is the PB *j* with a score R_j equal to 15.7, that is, 15.7 times greater than random chance. (**f**) The scores are translated into probabilities, *S*. (**g**) This transformation allows the local entropy *H* and N_{eq} values to be computed.

$$S_x = \frac{R_x}{\sum_{u=1}^{u=16} R_u} \tag{8}$$

$$H = -\sum_{u=1}^{u=16} S_u \ln S_u \tag{9}$$

An index, N_{eq}, that is, an equivalent in number of PBs, is derived from H:

$$N_{eq} = \exp(H) \tag{10}$$

This measure varies between 1 (when a unique PB is predicted) and 16 (when all PBs are equally probable). It therefore assesses the dispersion of a given PB distribution. It is assumed that the smaller this measure, the higher the confidence in the prediction for this site.

The maximum deviation angle (MDA) criterion

This measure, previously proposed by Bystroff and Baker, has been widely used in similar studies. It allows

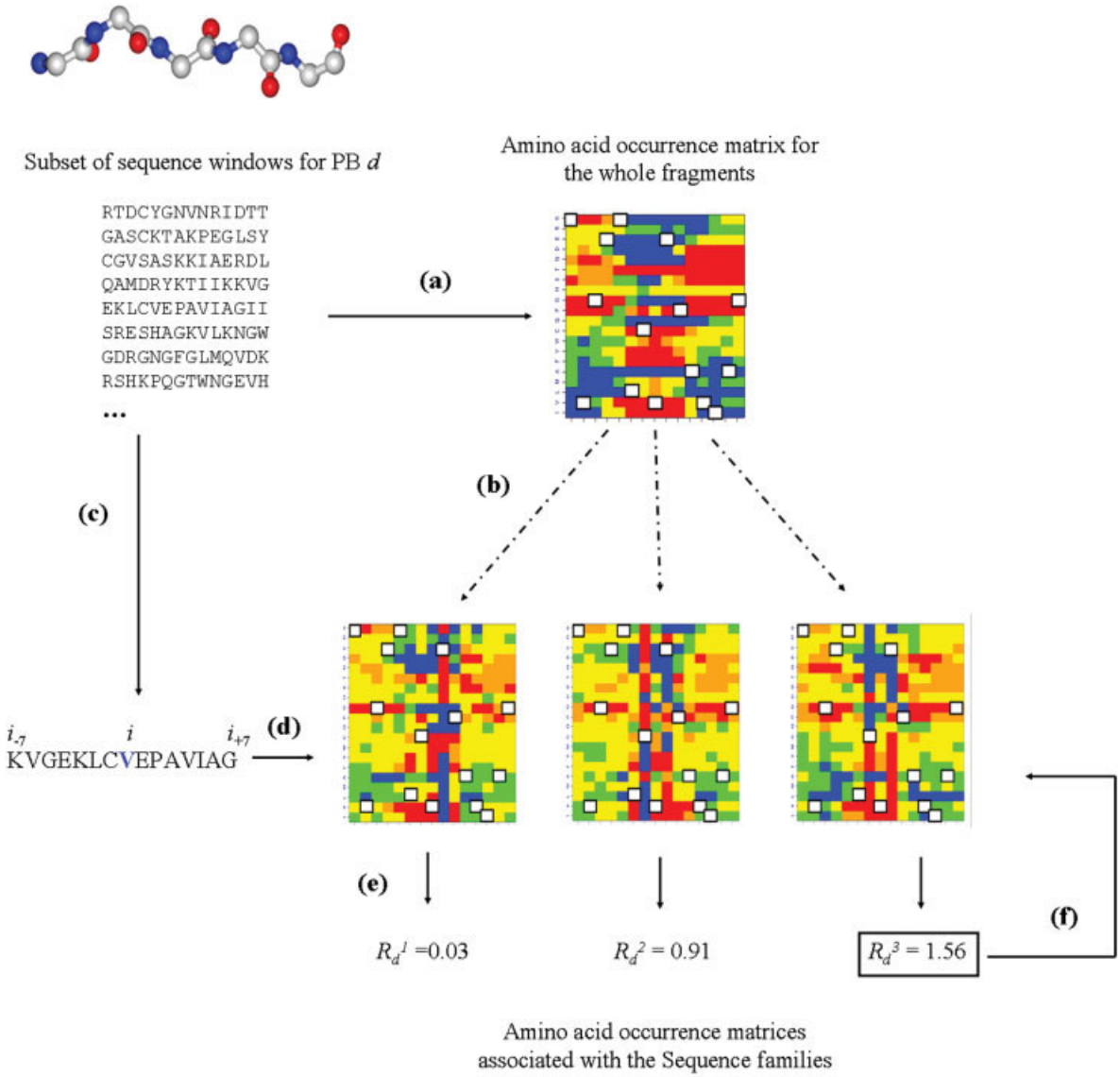


Fig. 3. Building sequence families. (**a**) An amino acid occurrence matrix is computed from a subset of sequence windows associated with a PB. (**b**) It is used to create the SFs. In this example, PB *d* is associated to three SFs. Initially then, the three SFs correspond to slight variations of PB *d*. (**c**) The training is done by randomly selecting a sequence fragment associated with the corresponding PB. (**d**) Three *R* scores are computed by a process similar to that used for predictions. (**e**) The scores are compared and the maximal one is selected. In our example, the highest score is found for SF d_3. (**f**) Its occurrence matrix is thus modified slightly to integrate the amino acid information of the fragment. The process is iterative from (**c**) to (**f**); all the sequence fragments are reexamined several times.

analysis of the local structure prediction on the basis of the dihedral angles.[19,58] The MDA quantifies the proportion of correctly predicted residues in a protein structure. A residue is considered correctly predicted if it is associated with at least one predicted fragment (in our case, a PB) that has no angle differing by more than 120° from the true 3D structure.

Sequence Families

Different sequence clusters may be associated with the same fold. In a previous work, we developed the concept of "*n* sequences for one fold," with *n* the number of sequence clusters associated with a given PB.[21] For each protein block, PB_x, the corresponding set of sequences is divided into *n* groups. Each is represented by one amino acid occurrence matrix and is noted PB_s^m with $m = 1, \ldots, n$ (see Fig. 3). SFs are computed in three steps:

1. Initialization of the SFs: the *n* occurrence matrices are initialized by the amino acid frequencies of PB_x. Then they are modified by adding weak random noise to differentiate the *n* matrices of dimension $20 \times l$ [see Fig. 3(b)]. At each position *j*, the probability sum equals 1. Hence, for the SF *m* of PB_x, $P(aa_i^j|PB_s^m)$ corresponds to the probability of observing amino acid *i* in position *j* in the sequence window of length *l*.

2. Search for the optimal sequence family: a sequence fragment associated with PB_x [see Fig. 3(d)] is chosen randomly and compared with each of the n matrices. To find the correct SF, an adequacy score R_{xm} [Eq. (11)] is computed for each SF. Its definition is similar to that of R_x [Eq. (7)]:

$$R_{x^m} = \ln \frac{P(PB_x^m \mid X_S)}{P(PB_x^m)} \tag{11}$$

Each sequence fragment is reallocated to the SF m^*, with $m^* = \text{argmax}\{R_{x^m}\}$.

3. Modification of the occurrence matrix m^*: The probability $P(aa_i^j)$ is modified as follows [cf. Fig. 3(f)]: (a) for amino acid i located at the jth position of sequence window X_S, we carry out the following transformation:

$$P(aa_i^j \mid PB_x^{m^*}) \leftarrow \frac{P(aa_i^j \mid PB_x^{m^*}) + \alpha}{1 + \alpha} \tag{12}$$

(b) elsewhere (i.e., for the other 19 amino acids):

$$P(aa_i^j \mid PB_x^{m^*}) \leftarrow \frac{P(aa_i^j \mid PB_x^{m^*})}{1 + \alpha} \tag{13}$$

This transformation allows us to conserve a probability sum that is always equal to 1 per position. Evolution of the training coefficient α is similar to that of the learning coefficient of Self-Organizing Maps (SOM):[47,48]

$$\alpha = \frac{\alpha_0}{1 + \left(\frac{t}{T}\right)} \tag{14}$$

with α_0 a user-fixed value ($\alpha_0 = 0.05$), t the number of fragments already used in the learning, and T the total number of fragments associated with PB_x. Every fragment is presented one time per cycle. In a previous work,[21] we used the expression with an initial value of $\alpha_0 = 0.01$ and conducted $N = 5$ cycles of learning [see Fig. 4(a)].

Sequence Family Improvement

In addition to the optimal adequacy score argmax $\{R_{xm}\}$, used to select the SFs, Q_{16} is introduced as an additional criterion during the learning step to define the SFs. Figure 4(b) explains the two successive training phases now used. The first aims at selecting the best solutions (those associated with high Q_{16} values) from the results of a "crude" training phase; it is very similar to the previous approach. The second phase is intended to increase the prediction rate Q_{16} by starting the training with the previously defined best solutions and then selecting the optimal SFs during the training.

In the first phase, a high value is chosen for the training parameter α_0 ($\alpha_0 = 0.05$) for one cycle, thus facilitating crude but rapid learning of the sequence windows associated with a given PB. The parameter α is then reduced according to Equation (14) for five cycles. The 20 best series of SFs, that is, those associated with the highest Q_{16} values, are selected from a large number of independent simulations (here, 2000).

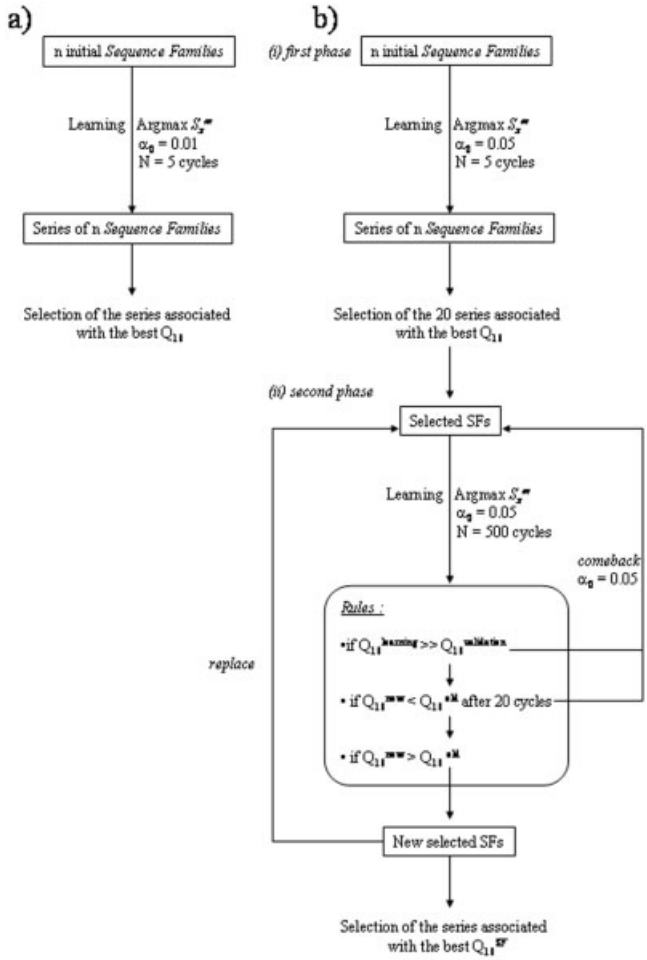


Fig. 4. Sequence family learning. (**a**) Classical approach.[20] (**b**) Improved approach (see text for details).

The second phase, a *refinement* phase, consists of a large number of training cycles ($N = 500$) from each series of selected SFs. The parameter α now decreases slightly. The Q_{16} value is assessed at each cycle. This process is stopped during the refinement phase if either of these criteria is met: (1) the Q_{16} value for the learning set is much higher than the Q_{16} value for the validation set, that is, overtraining has occurred, or (2) the prediction rate does not increase for 20 consecutive cycles after the best Q_{16} value is obtained. If either of these criteria is met, the last best series of SFs is used again and the process restarts with a high α value (equal to the initial α_0 value). Otherwise, if the Q_{16} value is better than the previous one, we select the new set of SFs.

This process is similar to conventional simulated annealing, with random trials and successive cycles with a decreasing control parameter T ("Temperature"). The Metropolis criterion is applied for each trial: a transformation is accepted either when the objective function (herein Q_{16}) increases or when the objective function does not increase, but a value randomly drawn in the range [0, 1] is less than

a given probability that depends on the differences observed in the objective functions.

PB Prediction by Adding Secondary Structure Information

To assess the influence of introducing secondary structure predictions into the prediction of PBs from sequences, we defined a *combined* prediction method that uses both: a well-known secondary structure prediction (PSIPRED)[59] software and PB prediction from sequences. This secondary structure prediction method is based on a two-stage neural network and can use either multiple sequence alignments ($Q_3 > 75\%$), or single sequences as input. The neural network yields a probability score, $P(Sec_t)$, associated with each residue for each predicted state t (t = 1, 2, and 3 for α-helices, β-strands, and coils, respectively). As we did for the prediction using Bayes' rule [cf. Eqs. (5) and (7)], we computed a prediction score, R_x^{sec}, for each PB_x:

$$R_x^{sec} = \ln \frac{P(PB_x^{sec})}{P(PB_x)} \tag{15}$$

where

$$P(PB_x^{sec}) = \sum_{t=1}^{3} P(PB_x \,|\, Sec_t) \times P(Sec_t) \tag{16}$$

The quantities $P(Sec_t)$ are provided by the PSIPRED program. PB x frequency in the secondary structures, $P(PB_x|Sec_t)$, can be defined in two different ways, with either the *true* secondary structures or the *predicted* ones. We assessed both by their prediction rates. The scores, R_x^{sec}, are translated into probabilities, S_x^{sec} [as in Eq. (8)]:

$$S_x^{sec} = \frac{R_x^{sec}}{\sum_{u=1}^{u=16} R_u^{sec}} \tag{17}$$

This *combined* prediction [Eq. (18)] uses both the S_x score from the prediction based on Bayes' rule [cf. Eq. (8)] and the S_x^{sec} score from the secondary structure prediction (Eq. 17):

$$S_x^{\beta} = (1 - \beta)S_x + \beta S_x^{sec} \tag{18}$$

β*, the optimal value of β, is associated with the maximum Q_{16} value. When β = 1.0, the prediction corresponds to the PSIPRED prediction, and when β = 0.0, to prediction with Bayes' rule [cf. Eq. (18)]. The optimization procedure varies β in small increasing steps (0.01) in the range [0.0, 1.0].

This approach also implies the use of a confusion matrix between the three states of the secondary structures and the 16 states of the PBs. Different confusion matrices have been computed: between the *true* secondary structures and the *true* PBs, *true* versus *predicted*, *predicted* versus *true*, and *predicted* versus *predicted*.

Expected Prediction Rate

Prediction rates are computed from a large set of sequences of known structures. Even when its average value is high, the dispersion of values may be substantial so that prediction efficiency is poor for some sequences. We propose an index that can distinguish between uninformative sequences, that is, those with poor prediction rates, and highly informative sequences, that is, those with high prediction rates. We previously defined for any given PB the measure N_{eq} [cf. Eq. (10)], which is correlated with the quality of local prediction. Accordingly, we use the mean N_{eq} value, $\overline{N}_{eq}$, to evaluate the prediction globally, for the entire PB series. It is defined as:

$$\overline{N}_{eq} = \frac{\sum_{i=1}^{L} N_{eq}^i}{L} \tag{19}$$

with L the total number of predicted fragments of the protein. The expected prediction rate, Q_{16}^E, is obtained by linear regression:

$$Q_{16}^E = a\overline{N}_{eq} + b \tag{20}$$

We then estimate the error on Q_{16}^E. This a priori observation can then be used to estimate the prediction rate.

RESULTS

We first examine the features of the PBs, in terms of PB distribution and of structural approximation, and compare the results with those of our previous studies. The relations between secondary structures, turns, and PBs are also described in depth to improve our understanding of the prediction results. The distribution of amino acids in PBs is described in the light of the database enlargement. Then we focus on predictions with PBs, evaluating and comparing the different prediction strategies. The quality of the prediction is assessed with various tools.

PB Features

New evaluation of PB features

In a previous study, we showed that a small 3D structure dataset was sufficient to capture the main geometrical features of protein fragments and to define relevant structural prototypes that can approximate the protein structures locally. In this study, we verified that these definitions remain valid after the size of the databank more than tripled (from 86,628 to 293,507 residues). Table I summarizes the principal characteristics of the PBs (for a detailed description of the earlier PB datasets see ref. 21 and supplementary data 1[60]).

Only few differences are observed between the new and previous datasets. The estimated number of PB repeats (*anr)* for PB *m* increased slightly, from 6.74 to 7.00, that is, the helices are a bit longer than previously. The frequencies of transitions between consecutive PBs in proteins are equivalent to those in the previous study with one notable exception, PB *j,* which has no preferential transition. The enlargement of the databank did not modify either the

TABLE I. Protein Blocks Characteristics

PB	Occurrence frequency (%)	Average number of repeats	Transitions				STRIDE			RMSDA (°)		RMSD (Å)	MDA 120 (%)	$d(C_1–C_5)$ (Å)
			First (%)	Second (%)	Third (%)	Sum[a]	α (%)	Coil (%)	β (%)	Mean	Dif[b]			
a	3.89	1.02	51.0(*c*)	16.9(*f*)	9.4(*d*)	77.3	0.1	75.8	24.1	45.2	29.3	0.46	97.65	10.6
b	4.41	1.01	48.4(*d*)	15.9(*c*)	12.9(*f*)	77.2	0.1	85.3	14.6	42.5	20.3	0.47	97.34	10.0
c	8.12	1.24	62.6(*d*)	23.5(*f*)	5.7(*e*)	91.8 / **93.7**	0.0	57.6	42.4	38.4	21.4	0.51	99.52	11.9
d	18.85	2.70	50.4(*f*)	26.3(*c*)	19.9(*e*)	96.6 / **98.7**	0.0	29.0	71.0	29.7	27.2	0.41	99.85	12.5
e	2.45	1.12	81.1(*h*)	8.6(*d*)		89.7 / **90.8**	0.0	45.5	54.5	40.9	23.5	0.71	98.28	11.9
f	6.68	1.01	61.5(*k*)	35.0(*b*)		96.5	0.0	73.3	26.7	37.5	22.1	0.40	99.21	11.3
g	1.15	1.04	37.5(*h*)	29.6(*c*)	16.1(*e*)	83.2	13.3	80.2	6.4	50.6	14.9	0.60	96.53	9.5
h	2.40	1.02	68.0(*i*)	13.8(*j*)	8.5(*k*)	90.3	2.0	76.2	21.9	47.0	20.9	0.46	96.22	8.5
i	1.86	1.01	82.8(*a*)	6.2(*l*)		89.0	2.0	90.3	7.7	43.4	25.0	0.41	96.05	8.6
j	0.83	1.01	21.7(*b*)	14.8(*a*)	14.7(*k*)	51.2	8.0	81.6	10.4	49.0	19.6	0.83	92.51	8.4
k	5.45	1.01	77.2(*l*)	10.5(*b*)	6.2(*o*)	93.9	49.3	50.2	0.5	35.9	25.4	0.30	98.61	7.5
l	5.46	1.01	68.2(*m*)	8.6(*p*)	7.1(*c*)	83.9	61.0	38.6	0.4	32.5	27.3	0.53	98.04	7.3
m	30.22	7.00	34.9(*n*)	15.7(*p*)	11.3(*k*)	61.9 / **94.6**	92.3	7.6	0.1	15.0	40.1	0.31	99.74	6.6
n	1.99	1.01	92.4(*o*)			92.4	75.7	24.0	0.3	26.8	31.2	0.31	96.96	6.5
o	2.77	1.02	78.2(*p*)	6.5(*m*)	5.6(*i*)	90.3	50.8	49.0	0.2	38.3	27.1	0.48	95.26	6.9
p	3.47	1.00	58.6(*a*)	23.7(*c*)	7.6(*m*)	89.9	17.1	81.3	1.6	43.8	25.9	0.47	97.30	9.4
Sum						89.3 / 94.3	37.7	39.8	22.5	30.1	29.5	0.41	98.82	9.2

[a]Sum of the three first transitions, For PBs *c, d, e,* and *m* that have a high repetition upon themselves an additional value is given that corresponds to the sum including the corresponding anr value (in bold).
[b]The difference between the smallest RMSDA and the second smallest RMSDA.

TABLE II. Protein Blocks and Turns

Name	%	Positions *i*	*i*+1	*i*+2	*i*+3	*i*+4	*Structural words*
γ-turns	*4.0*						
classic	0.6	h	i	a			*eehia, ehiac, hiacd*
inverse	3.5	—	—	—			
β-turns	*29.1*						
I	13.2	f	k,n	l,o	p		*fklmn, afklm, bfklm, fbfkl mmnop, mnopa, nopac, nopab*
II	2.3	e,g	h	i	a		*eehia, ehiac, hiacd*
VIII	2.0	f	b	h,l	c		no SW
I′	0.8	e	h	i	a		*eehia, ehiac, hiacd*
II′	0.6	h	j	b	c		no SW
VI b	0.3	a	c	f	b		no SW
IV	10.1	e,f	h,k,n	l,i,o	a,h		*eehia, ehiac, hiacd fklmn, afklm, bfklm, fbfkl*
α-turns	*4.0*						
I-RS	0.9	f	k	l,n	o	p	*fklmn, afklm, bfklm, fbfkl mmnop, mnopa, nopac, nopab*
I-LS	0.03	e,o	h	h	i,p	a	*eehia, ehiac, hiacd*
II-RS	0.1	e	h	i	—	—	*eehia, ehiac, hiacd*
II-LS	0.01	h,o	j	h	h,p	i,a	no SW
I-RU	0.1	h	j	k,l	l	—	no SW
I-LU	0.1	e	h	h	i	a	*eehia, ehiac, hiacd*
II-RU	0.04	h	h,o	i,p	a	—	*eehia, ehiac, hiacd*
II-LU	2.6	—	n	o	p	a	*mmnop, mnopa, nopac, nopab*

mean value of the RMSDA (30°) or the median value (26°). The mean RMSD, obtained by superposing the fragments and their closest prototype, equals 0.41 Å, and its median value is 0.34 Å. The PBs show only slight variability for all the geometric measures, except for the least frequent PB, *j*.

Secondary structures, turns, and PB series

The secondary structure assignment (defined by STRIDE) gives the same distribution as the previous work. PBs *k*, *l,* and *m* are associated mainly with α-helices, and PB *d* with β-strands. PB *m* is the only PB with a strong correspondence to a particular secondary structure—the central part of the helix—while PBs *k*, *l*, *n,* and *o* correspond to the helix caps. Eight of the 16 PBs are associated principally with the coil state and permit a more detailed description of it. Some specific relations between turns and PB types are summarized in Table II: two types of γ-turns (three residues), seven types of β-turns (four residues), and eight types of α-turns (five residues). These turns are all defined by a distance criterion and dihedral angle criteria[61] (see supplementary data 2).

Two points can be highlighted. First, one type of turn may be associated with only a few different PB series. For example, the β-turn I can be described by an *fkl** series or a

TABLE III. Prediction Results

PB	Number of sequence families	Classical prediction	Old SFs approach	Improved SFs	classical/ PSIPRED	PSIPRED with SFs	PSIPRED after SFs
a	1	59.2	53.5	57.4	47.8	58.3	56.6
b	2	12.7	27.0	23.3	5.4	19.4	20.9
c	2	26.4	32.9	35.8	6.2	29.4	32.9
d	3	28.3	34.8	47.3	56.7	36.4	54.0
e	1	40.1	35.9	38.2	22.6	37.9	38.6
f	2	29.7	36.2	33.0	14.8	33.3	30.9
g	1	30.3	35.1	29.8	24.9	28.5	30.1
h	1	42.6	42.7	40.9	35.9	40.2	40.9
i	1	37.7	41.0	37.5	41.3	36.7	38.1
j	1	49.1	47.2	48.5	50.9	48.8	49.7
k	1	38.5	35.2	34.9	18.9	35.6	33.4
l	1	37.5	32.1	36.7	14.8	38.5	35.5
m	6	39.7	50.8	68.3	67.0	63.3	70.6
n	1	51.2	44.7	51.7	38.4	52.2	50.0
o	1	49.2	45.8	47.9	16.1	48.4	48.1
p	1	30.5	33.9	31.1	20.5	31.9	29.2
Q_{16}		**35.4**	**40.7**	**48.7**	**41.2**	**44.7**	**49.9**
Q_{14}		**35.4**	**36.5**	**37.4**	**20.2**	**36.4**	**36.1**
β*					0.55	0.55	0.20

**nop* series, with * representing any PB. Conversely, one PB series can be associated with different turns. For example, the *hia* series is found in the γ-turn, β-turns II and I′, and α-turn I-LU. It is important to note that the PB series are longer than the length of any given turn. For example, the *hia* series corresponds to seven residues highly correlated with specific local protein structures, as seen in the Structural Word *ehiac* (Structural Words or SWs are the most frequent series of five consecutive PBs).[36] The results with α-turns, which have been analyzed less, show particular involvement of unusual *hh* series. Thus, we have pointed out several features that are encompassed in SWs (longer than classical turns). It may thus be possible to extend the turn classification to take their local environment into account.

Amino acid distributions

Amino acid distributions in PBs remain close to those previously observed (see supplementary data 3): for any position, we observe a simple increase in the number of over- and underrepresentations due to the size of the databank. The amino acid distributions in PBs *m* and *d* are close to those observed in α-helices (L, A, M, Q, E) and β-strands (I, V, F, Y), respectively. Amino acid preferences for these PBs extend over several positions in the range $[i - 2, i + 2]$. The preferred amino acids in the N- and C-caps of repetitive structures (*k*, *l*, *n*, *o*) are S, N, D, E, P, and Q. These preferences are consistent with other descriptions.[1,62] PBs associated with nonrepetitive structures conserve high amino acid specificity in specific positions, as previously described.[21,36] The positions within the range $[i - 2; i + 2]$ remain most significant. This observation suggests that the introduction of secondary structure prediction would improve the global PB prediction rate. Nonetheless, the correspondence of the PB *m* with the α-helix and of *d* with the β-strand, while clear from an amino acid viewpoint, is not as strong from a structural point of view. For example, PB *m* is associated with α-helices at a frequency of 92% and *d* with β-strands at a frequency of 72%.

Prediction Results

The prediction methods

Six approaches were compared: the first three methods used a score based on Bayes' rule while the latter three took a secondary structure prediction (from PSIPRED, see Methods for details) into account.

The first, simplest prediction approach (*classical*) used one amino acid occurrence matrix for each PB. The next two approaches were based on the SF concept, that is, several occurrence matrices may be associated with one PB, computed with either a simple learning method (*old SF approach*[21]) or a sophisticated learning method (*improved SF*, this article). The latter three approaches used mixed prediction, combining PSIPRED and PB prediction with a weighting parameter, β. Hence, secondary structure information plays a role in predicting the PB, through: (1) a direct combination of secondary structure prediction and PB prediction, or (2) applying the secondary structure prediction during SF optimization, or (3) applying secondary structure prediction after SF optimization. To ensure appropriate comparisons between the two SF strategies, we used the same number of SFs as in our previous study (26 SFs: 6 different SFs for PB *m*, 3 for PB *d*, 2 for PBs *b*, *c* and *f*, and one for the others).

Quality of the prediction was assessed with different indexes and was computed as an average of the prediction rates for each sequence protein in the validation test. Q_{16} is the reference measure (see Methods), comparable to Q_3 for the three states in secondary structure prediction. An additional index, Q_{14}, evaluates the prediction of the 14 nonrepetitive PBs, that is, all the PBs except *d* and *m*.

Table III sums up the best Q_{16} and Q_{14} prediction rates, and the best prediction efficiency for each PB by method, once all the parameters were optimized. For the first three methods, which did not use secondary structure information, the Q_{16} prediction rate ranged from 35.4 to 48.7%. As expected, because the amino acid specificities were preserved, enlarging the databank did not modify the results from the basic standard method: 35.4% is similar to the values we previously observed.[21] In contrast, results with the new SF method showed marked strong improvement over the previous SF procedure. Introducing secondary structure information in each procedure improved the Q_{16} prediction rates. The values ranged from 41.2 to 49.9%, depending on the procedure. This change improved the prediction rate for PBs *m* and *d* most strongly. For Q_{14}, however, which computes the prediction rate average by considering only nonrepetitive blocks, that is, excludes *m* and *d*, this improvement vanished and Q_{14} values remained close to 36.5%, except with the simplest method, the value of which dropped to 20%. Clearly, in the latter case, secondary structure information introduces large errors into the prediction of some of the nonrepetitive PBs and results in prediction values close to random. Using the SFs produces a Q_{14} value of 37.4%, compared with 35.4% with the classical procedure. This value is also better than that found with the previous SF procedure (36.5%[21]).

In summary, the best results are obtained with the new SF method and with it, secondary structure information slightly improves the results, but can weaken the prediction of some PBs. We discuss the different procedures in detail below.

Classical prediction with the secondary structures

This approach uses a confusion matrix between the three states of the secondary structures and the 16 states of the PBs. A first method used the confusion between the *true* secondary structures and the *true* PBs (data not shown). The Q_{16} value equalled only 33.6%, and seven PBs had a prediction rate below 10%, extremely poor. Alternative confusion matrices were established (*true* vs. *predicted* states, *predicted* vs. *true* states); surprisingly, the best results were obtained with the confusion matrix based on the PSIPRED *predicted* secondary structures and the *predicted* PBs (with *classical prediction*). The optimal parameter β, β^*, equals 0.55 and ensures a Q_{16} value of 41.2%. This method provided better predictions for 91% of the proteins than either the *classical prediction* method or the previous version of the SF approach. Unfortunately, the Q_{14} value decreased strongly (20% compared to 36%). This method appears to entail a strong bias that essentially favors the two repetitive PBs, *m* and *d*, at the expense of the other PBs. As mentioned above, the amino acid distribution is essentially identical for α-helices and PB *m* and for β-strands and PB *d*. Thus, secondary structure prediction results in overprediction of the *m* and *d* states. The procedure thus focused on the bijective relations associating α secondary structures with PB *m* and β structures with PB *d*. No such relation connects coils with the 14 remaining PBs, their relations are only unidirectional (from PBs to coil). Hence, although the global prediction rate appears adequate, prediction rates for most of the PBs dropped. Interestingly, the prediction rates for PB *m* and the α-helix PSIPRED prediction rate were rather close (respectively, 67 and 72%) and our approach is even better for PB *d* compared to the β-strand prediction (respectively, 56.7 and 43%). These values were computed for one sequence (no alignment).

Control parameters of the SF learning algorithm

One of the major advances presented here is the new procedure for defining the SFs. Different parameters control the process now, and they have been carefully optimized. In the first phase of the new SF algorithm (see Fig. 4), similar to the old SF procedure, the learning coefficient α_0 has been changed to provide optimal results (0.05 instead of 0.01). For the 2000 simulations, the Q_{16} rate ranged from 30.0 to 41.8%, which is slightly better than our previous rate (40.7%[21]). This improvement is also due to a well-balanced choice between the learning and validation sets. With $\alpha_0 = 0.01$, Q_{16} does not exceed 41.0%.

Adding the iterative learning phase to the new SF procedure increased the prediction rate to the range [44–46%] for a few cycles (less than 20). Succeeding cycles improved the prediction rate for some PBs, but at the expense of the global Q_{16} prediction rate, which fell to the range [40–42%]. Sometimes, the Q_{16} value exceeded 50%. Prediction was again, however, strongly unbalanced in favor of the repetitive PBs, clearly due to overtraining.

In accordance with the rules specified in the method section, we selected a series of SFs that corresponded to a Q_{16} value of 48.7% (cf. *improved SFs* in Table III). The final process increased the Q_{16} value by 8.0% over the previous SF definition,[21] and by 6.9% over the results of the first phase.

SFs and PSIPRED

Secondary structure prediction is added to the SFs in two different ways: the secondary structure predictions are used (1) during the SF learning process [Eq. (13)], or (2) after the SF optimization (see Sequence families in the Methods section).

During the SF training, the score is weighted with the secondary structure prediction ($\beta = 0.55$ remains constant). This procedure affected only the Q_{16} computation: the prediction rate rose from 41.2 to 44.7%. More interestingly, it improved for 95% of the proteins. The prediction is better distributed over the entire PB set than with the combined *classical*/PSIPRED approach. The Q_{14} value was also much better, but PBs associated with β-strands (PBs *b*, *c,* and *f*) and some coils (PBs *g* and *i*) were clearly underpredicted. The Q_{14} value was, however, lower than it was with the improved SF approach or even with the previous SF method.

When we combine the results of the SF approach with the PSIPRED results, an optimal Q_{16} prediction rate, equal to 49.9%, is obtained for $\beta = 0.20$ (with the previous β value—0.55—Q_{16} equals 46.9%). This Q_{16} value is slightly better (+1.2%) than the prediction rate from the

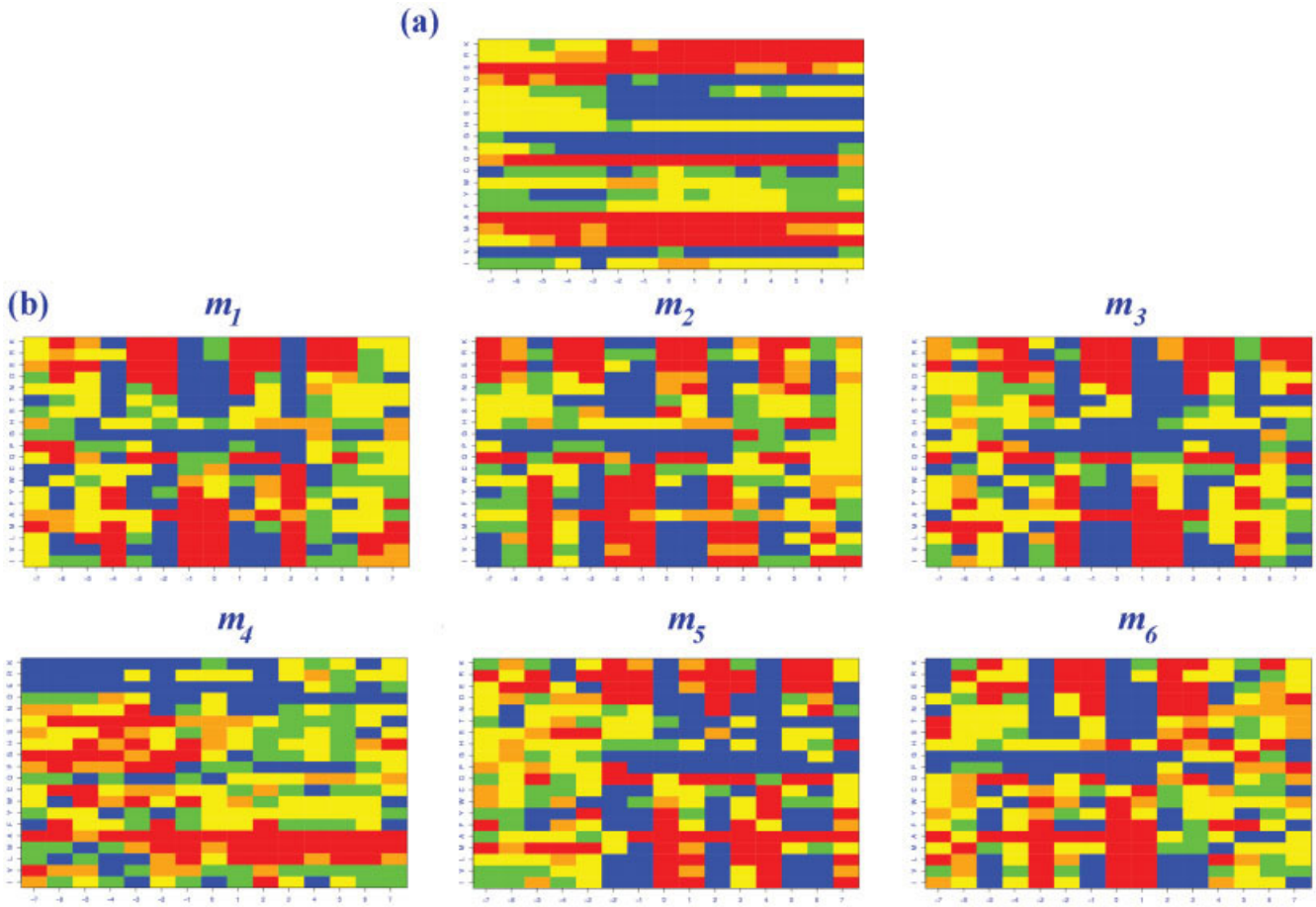


Fig. 5. *Z*-scores matrices of PB *m* and its sequence families. Amino acid distribution of (**a**) the Protein Block *m* and (**b**) PB *m* SFs (noted from m_1 to m_6) in terms of *Z*-scores, with (blue): *Z*-score $< (-4.4)$, (green): $(-4.4) \leq$ *Z*-score $< (-1.96)$, (yellow): $(-1.96) \leq$ *Z*-score < 1.96, (orange): $1.96 \leq$ *Z*-score ≤ 4.4 and (red): *Z*-score > 4.4. Positive *Z*-scores (respectively negative) correspond to overrepresented (respectively underrepresented) amino acids.

improved SF approach. This improvement too is essentially due to the improvement in the rates for PB *m* (>2%) and PB *d* (~7%). Other PBs increased slightly, including PBs *e*, *g*, *i*, *j,* and *o*. The Q_{14} value, however, remained roughly equal to the *old SF approach*[21] alone and lower than the Q_{14} rate with the improved SF approach.

As we showed above, using the new SFs and not considering secondary structures ensures the greatest improvement in prediction rates, including both Q_{14} and Q_{16}. In the paragraphs below, we will focus on this approach.

Sequence–Structure Relations through the New SF Procedure

The SF approach reinforced sequence–structure relations within the same local protein structure, that is, different sequence signatures for the same PB. The learning procedure we chose makes it possible to evaluate the amino acid contribution to the improved prediction. In this section, we describe some of these factors and discuss the rules and parameters that govern the efficiency of the SF procedure.

Sequence families associated with PB m

Optimal SF capture and discriminate sequence signatures better for each PB. Figure 5 provides an illustration, a detailed analysis of the amino acid distribution in the different SFs associated with PB *m,* with information for each position in sequence window *l*. Figure 5(a) shows the amino acid distribution in PB *m* before the SF process. Position 0 shows that Leu, Met, Ala, Gln, Glu, Arg, and Lys are clearly overrepresented. Only Ala is overrepresented all along the PB window (red horizontal line). For position 0, Pro, Gly, Ser, Thr, Asn, and Asp are underrepresented but only Gly is underrepresented along a large part of the window (blue horizontal line, position -6 to $+7$). In most cases, position specificity is weak (long horizontal lines). In contrast, application of the SF procedure [Fig. 5(b)] results in breaks in the long horizontal lines, which indicate clear position specificity. The amino acid that are under- and overrepresented are grouped interestingly according to their physicochemical properties. Moreover, strong specificity is observed at the dipeptide level, that is, when two consecutive positions involve the same set of amino acids. For example, for SF m_3, Ala, Gln, Asp, Glu, Arg, and Lys

TABLE IV. Transition Rates between the Sequence Families[a]

	b_1	b_2	c_1	c_2	d_1	d_2	d_3	f_1	f_2	m_1	m_2	m_3	m_4	m_5	m_6
a															
b_1								25.69	12.38						
b_2								17.82	10.65						
c_1		7.42	7.87	11.94		7.93									
c_2	9.86	10.48	8.22	10.74		7.84									
d_1	16.15	13.28	21.83	11.69	15.71	20.89	28.32								
d_2	16.30	33.35	22.91	14.31	15.01	18.76	21.49								
d_3	12.88		16.73	13.39	23.32	23.03	20.55								
e			5.21		7.17	6.96	8.96								
f_1	7.60	6.02	8.63	11.04	16.06	7.59	11.04								
f_2	7.16	5.13		14.31	13.82	5.06									
g															
h															
i															
j															
k								51.49	74.43						
l	8.99	8.51													
m_1														44.82	52.41
m_2										52.27				24.09	
m_3										30.42	33.62	7.74	10.37		
m_4													53.23	5.72	
m_5											5.51	24.94	7.88	12.71	15.36
m_6											26.92	47.67			9.52
n										5.57	8.80				
o															
p															

[a]For clarity, frequencies ≤5.0% are omitted.

are overrepresented in positions −1 and 0, and Ile, Val, Leu, Met, Phe, Tyr, Pro, and Gly are underrepresented. Leu and Met are overrepresented in the corresponding positions of PB m. The amino acid distribution in SF m_4 is different from that of the other SFs involving m and from the original PB m distribution. Only 36 over- and 43 underrepresentations of different amino acids are found in the window [−4; +4] compared with 50–53 over- and 59–71 underrepresentations for the other SFs of m. It is the least informative SF. In comparison, PB m has 56 over- and 55 underrepresentations.

Analysis of sequence family transitions

We previously pointed out that the most frequent series of PBs generate different highly overlapping patterns,[36] with strong specific transitions. These structural transitions are frequently associated with specific amino acid transitions. The SF clustering procedure captures specific amino acid transitions between PBs, and the resulting SFs reflect the different amino acid transition patterns. Table IV offers a detailed analysis of the preferential transitions between the SFs. For example, the two SFs (f_1 and f_2) associated with PB f show a preferential transition to the SF of PB k and to the SFs of PB b (b_1 and b_2). Nevertheless, behavior is slightly different according to the SF: SF f_2 forwards preferentially to SF k compared with SF f_1. In contrast, SF f_1 forwards preferentially to SF b_1 or b_2. Similar behavior is observed for the SFs of PB c. For the transition SF b to the SF d, SF b_2 forwards mainly to d_2 and never to d_3 but SF b_1 does not show such specialization.

More SFs were generated for the PBs d and m (3 and 6, respectively). Most of the transitions (70%) for SF d_3 occur towards SFs of PB d. No strong preference for any of these SFs is observed. Only 62% of the transitions from SF d_2 forward to SFs of PB d. The other transitions involve transitions quantitatively similar to those involving the SFs of PB c or f. Finally only 54% of the transitions from SF d_1 are to SFs of PB d. Typically, SF d_1 is associated with the amino acid signature of the PB d exits, SF f_1 and f_2. Of the six SFs of PB m, four (m_1, m_3, m_5, and m_6) are preferentially (91–86%) followed by SFs of PB m. They spread out mainly over two SFs (50% on average for the first and 27% for the second). The transition rate from SF m_4 to SFs of PB m is 84%. Its main transition is to itself, and the secondary transitions are evenly distributed over all of the other SFs of m. Finally, SF m_2 is the amino acid signature for PB m exits (22% to SFs not associated with PB m).

Analysis of the Improved Sequence Family-Based Prediction

Analysis of the predicted states

Using the SF procedure does not necessarily improve each PB prediction rate. For example, PB b is the most difficult PB to predict: even with two SFs, its prediction rate climbs only from 12.7 to 23.3%, a result less than that obtained with the old SF procedure (27%). It was not possible to exceed 25.0% in any of the simulations. Adding a third SF for PB b creates a strong imbalance and the Q_{16} value drops. Nevertheless, the new SF procedure improves the prediction for many of the PBs with one SF over

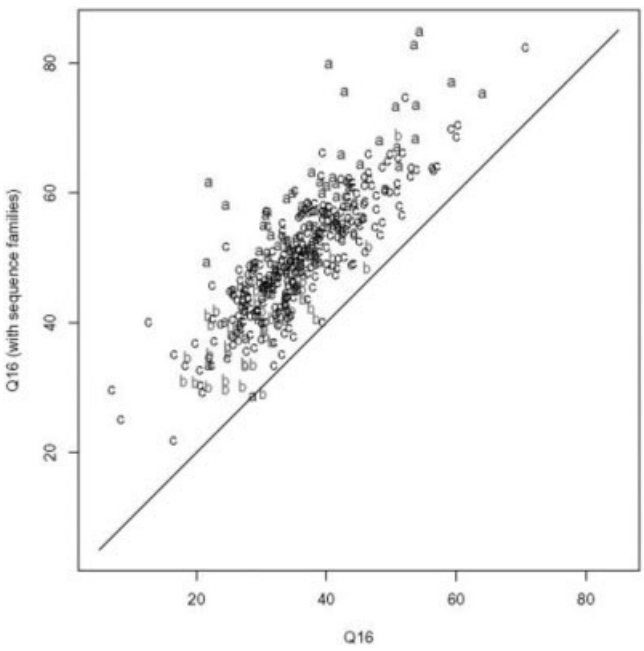


Fig. 6. Improvement of the prediction rate after introducing sequence families. The initial prediction rate Q_{16} is given in abscissa and the prediction rate obtained with the SFs in ordinate. The proteins are labeled (**a**) for all-α proteins, (**b**) for all-β proteins, and (**c**) for others according to the rules set forth by Michie et al.[56]

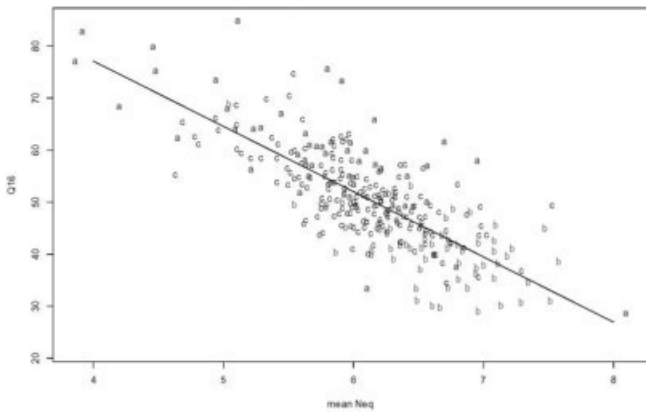


Fig. 7. Computation of expected prediction rate. $\overline{N}_{eq}$ in abscissa and the Q_{16} values in ordinate. The linear regression gives a relation $Q_{16}^{E} = -12.6\ \overline{N}_{eq} + 127.2$ with a squared correlation coefficient, R^2 value, of 0.88 ($p < 10^{-9}$).

previous levels: PBs *a* (+3.9%), *e* (+2.3%), *j* (+1.3%), *l* (+4.6%), *n* (+7.0%), and *o* (+2.1%). In a few cases, PBs *g* (−0.5%), *h* (−1.7%) and *i* (−0.2%), the procedure is slightly less efficient but the values remain close to the initial prediction rates.

A confusion matrix between the *true* PBs and the *predicted* PBs has been computed. It shows that prediction specificity remains high, that is, no PB is predicted preferentially at the expense of another PB (data not shown). Of the confusion between PBs, only 6% is at a rate larger than 10%, and 86% less than 5%. The most confused predictions concern PBs *i* and *j* (26%). This value is similar to that obtained without SFs (23%). The confusion in the prediction of these PBs may be due to their low frequency (1.9 and 0.8%, respectively) and their proximity in terms of structural characteristics and amino acid specificity.

Individual proteins

Figure 6 shows the Q_{16} prediction for the validation set with the combination of classical prediction and the improved SF approach. Prediction is clearly better for all but one protein. The all-α protein (noted a) prediction rates rise from 39.1 to 59.1%, the all-β protein (noted b) rates from 31.4 to 41.4%, and the others (noted c) from 36.2 to 50.1%. The only exception concerns the central part of the tailspike protein (~100 residues),[63] composed mainly of coils (61%) and β-strands (39%). Its prediction rate is poor (29% with the *classical prediction* and 28% with the *improved SF*-based prediction). Its PB composition (and secondary structure) is far from the classical average observed for soluble globular proteins, and a large conformational change occurs when the protein is involved in the functional tetramer structure.

We explored two additional points that allow us to propose methods for quantifying the relevance of the prediction and the accuracy of the predicted local structures.

Estimation of the prediction rate

Previous studies[21,36] showed that the N_{eq} index (based on the Shannon entropy of the score distribution) is a measure of the local confidence in the prediction (lower values = better confidence). We note that its value dropped from 7.8 to 5.2 PB-equivalents from *the classical prediction* to the *improved SF*-based prediction, a change that illustrates the improvement of the sequence–structure relation.

Analyzing the distribution of the mean N_{eq} per protein, we found a linear relation between the Q_{16} value and the mean N_{eq} (cf. Fig. 7). The linear regression gives the following parameters [cf. Eq. (20)]:

$$Q_{16}^{E} = -12.6\overline{N}_{eq} + 127.2 \qquad (21)$$

where Q_{16}^{E} denotes the expected prediction rate. The value of R^2, the squared correlation coefficient, is 0.88 ($p < 10^{-9}$). The larger the N_{eq}, the smaller the Q_{16}^{E}. Thus, from this equation, we can evaluate the expected prediction rate solely on the information from score distributions.

The Q_{16}^{E} value has a mean error of 5% and a standard deviation of 10%. The difference between the predicted and the observed Q_{16} values is less than 2% for 28% of the proteins; for 33%, the difference ranges from 2 to 5%, for 28%, from 5 to 10%, and, for 11%, more than 10%. Interestingly, the Q_{16}^{E} value is underestimated for the all-α proteins (average error = 6.3%), and overestimated for the all-β proteins (average error = 5.4%).

Local structural prediction

The prediction is expressed as a series of letters, that is, predicted PBs. We consider here the quality of predictions

for 3D local folds. To assess this, we superimposed all the predicted PBs, represented by the 3D prototype of the PB, over the real 3D fragments and computed the mean RMSD and RMSDA values.

The mean RMSD value equalled 0.60 Å. PBs *k* through *p* were in the range (0.48–0.66 Å) with a mean RMSD of 0.48 Å for PB *m*. PBs *a* through *f* were in the range (0.57–0.76 Å) and *g, h*, and *i* in the range (0.63–0.91 Å). These values include both correct and incorrect PB predictions. These values are rather small for a prediction rate close to 50%, which probably mean that the wrongly predicted sites are represented by PBs close to the true 3D local folds. The mean RMSDA value was 56°. We note that PBs *m, d,* and *l* were correctly approximated with RMSDA values of 39, 55, and 51°. All the others range from 61 to 71°.

Furthermore, we determined that the efficiency of the approximation was correlated with the quality of the local prediction. For this purpose, we used the prediction confidence indexes, that is, the N_{eq} values, and we divided the predicted fragments into population quartiles based on their local N_{eq} values. For the prediction of the first quartile, the RMSD was 0.42 Å and the RMSDA 33°, for the second, 0.61 Å and 57°, respectively, for the third 0.64 Å and 60°, and for the last, 0.70 Å and 70°. There is an obvious linear relation between the N_{eq} values and 3D approximation: low N_{eq} values are clearly associated with well approximated local structures.

The mda120 value is an alternative index frequently used to evaluate the quality of 3D prediction. It is computed as the number of dihedral angles in a predicted 3D local fold that deviate by less than 120° from the true corresponding structure. With the *classical prediction* method, it equalled 65.1%, while with SF-based prediction, it reached 79.2%. We note that the overall predictions for all PBs improved because of the sequence–structure specificity. The best PBs were PB *m* (mda120 = 89.1%), *d* (83.8%), *n* (79.5%), and *a* (78.2%). The lowest was PB *j,* which increased from 54.4% to only 56.4%.

DISCUSSION AND CONCLUSION

Comparison with Other Alphabets

As already noted, a structural alphabet can be defined correctly with a limited number of proteins.[20] Here, we have confirmed this observation with new larger datasets. Variations with results from our previous work are fairly slight.[21,36] The main difference lies in the features of PB *j*, the least frequent and most variable PB. The local sequence–structure relations obtained with the new datasets remain very stable.

Various structural alphabets based on different geometrical descriptors and classification methods have been proposed. We compared our structural alphabet with the "oligons" defined by Mitchelli and coworkers[10] and with the series of prototypes defined by Kolodony and coworkers.[9]

Micheletti's team uses an iterative approach and computes every RMSD between every fragments of their databank. Their approach (1) creates clusters of local folds based on the distribution of RMSD, (2) searches for the most populated cluster, (3) selects it as an oligon, (4) eliminates the fragments associated with it from the analysis, and then (5) returns backwards to step (1). Hence, it creates a hierarchy in the cluster definition: the first is more important than those that follow.

The comparison (with RMSD criteria, see supplementary data 4) of the 16 PBs with the 40 oligons of length 5 shows that all the PBs, except *j* (which is associated with oligon 35), can be associated with one of the first ten oligons. Some oligons do not correspond to only one PB. The relevance of their prototypes can be quantified by the average number of states per residue, called the complexity index and proposed by Levitt's group. This measure is related both to the library size and the length of the fragments. The higher this measure, the better the 3D structural approximation. The "oligon" approach first creates large populated clusters with α-helices and β-sheets, but cannot accurately distinguish N- and C-caps of repetitive structures, even with a complexity of 6.32. Oligon 1 is associated with PBs *m*, *n,* and *l*, oligon 2 with PBs *d* and *c,* and oligon 9 with PB *b* and *h*. In all, 13 PBs are associated with the first nine oligons.

The method presented by Kolodony's group is based on a modified *k-means* approach. The different libraries are designed for the best fit and reconstruction of the protein structures. A large set of different libraries (from 4 to 250 structural prototypes, *k*) is designed with four different prototype lengths, *r* (from four to seven residues). An important point is that fragments never overlap. The learning method consists in (1) selecting *k* protein fragments as *k* initial prototypes, (2) associating each protein fragment with its closest prototype, with the criterion of RMSD, (3) modifying the *k* prototypes according to their associated fragments, and (4) repeating the process until convergence. Fragments too distant from any prototype are considered outliers and eliminated. To compare with the PBs, we selected the closest states for comparison with the PBs: those with $k = 10$ and $k = 20$ for length $r = 5$, that is, noted respectively as K10 and K20.

Overall, K10 had a complexity of 3.16 and approximated the 3D structures locally with an average RMSD of 0.57 Å. For K20, the complexity was 4.47 and the average RMSD 0.47 Å. In our case, the values were 4.00 and 0.41 Å, respectively. Given the different sizes of the datasets used, our results are slightly better. The RMSD comparison shows that even though the methodology is quite different, the K10 and K20 libraries are fairly similar to the PBs.

The principal differences to be noted between our approach and Kolodny's are that our method uses: overlapping fragments to determine PBs, the transitions between them, all the fragments to define them, and finally, the information of dihedral angles.

To summarize, these comparisons show that PBs describe local structures correctly and detect small changes in local structure. The latter point is especially interesting because this limited number of prototypes (that is, 16) makes possible a good balance between description of the 3D structures and their prediction from sequences. Karchin et al.[38] assessed the features of our structural alphabet

quantitatively, noting its good predictive power in terms of bits per position and mutual information. They found it yielded the best values among nine structural alphabets for those criteria (see Tables II and III of Ref. 38).

Turns

A comparison of geometrical features in classic turns shows their association with particular PBs, such as *g,* and with some specific series, such as *ehia*. Nevertheless, there is no systematic equivalence between turns and specific PB combinations. Turns are defined by the delimitation of secondary structures, while PB series are defined by themselves, without any a priori decisions. Turn prediction thus depends on the quality of secondary structure prediction and is very sensitive to the low frequencies of some turns.[64] PB prediction may be helpful in progressing beyond these two limitations, and may thus more accurately predict the local protein structure associated with these regions.

Prediction

Most structural alphabets were not designed to predict 3D structures but rather to describe 3D structures accurately. For prediction, Bystroff and Baker have preferred to focus on optimizing the sequence–structure relation to define recurrent motifs called I-sites.[19] Consequently, the I-site library may well miss structural motifs that do not correlate well with a sequence motif. Nonetheless, this approach has had significant success in predicting structural motifs in proteins.[32] Direct comparison with our results is difficult because of the variable length of the representative fragments and the evaluation measure (mda120) used. Results with the alphabets of Camproux et al. and Hunter and Subramaniam can be compared more easily. Both use a sequence–structure relation deduced from the clustering procedure for predictions. Camproux's alphabet (12 Structural Building Blocks) is based on motifs 4 C_α length in length and uses a sophisticated HMM learning method.[20] It has been used mainly in a Bayesian approach for predicting the loop regions that connect α–α and β-β secondary structures. The prediction rate was close to 30% for fragments with lengths ranging from three to six residues.[25] Hunter and Subramaniam describe fragments (called "centroids") seven residues in length, obtained from a hypercosine clustering method.[23] The alphabet used for prediction involves 28 centroids, and the Bayesian prediction rate reaches 40%.[11] However, this value is strongly biased towards the most frequent centroids at the expense of the others. Eight of the 28 centroids are predicted at a rate above 20% but only four above 50%. In addition, eleven centroids are not predicted at all.

The prediction rate with our approach and the new SF method improved markedly (the Q_{16} value rose from 35.4 to 48.7%), increasing by 8.0% over previous results.[21] In a random simulation with PB frequencies, the Q_{16} value reached a maximum of 7.5%. Moreover, this improvement was not biased towards the most frequent PBs (*m* and *d*), although they improved strongly as well (from 39.7 to 68.3% and from 28.3 to 47.3%, respectively, for gains of 17.5 and 12.5% over our previous findings). The Q_{14} value also improved, from 35.4 to 37.4%, 0.9% over the previous Q_{14} value. In contrast, using secondary structure prediction directly in the PB prediction process did not significantly improve the Q_{16} value and even decreased the Q_{14} value slightly. Secondary structure prediction clearly introduces a bias that favors the prediction of the repetitive PBs at the expense of the PBs associated with less structured zones.

All the predictions in this article are based on a single sequence. Secondary structure predictions improved substantially when multiple sequence alignments were used (from 65 to 80% on average).[65,66] Such improvement may similarly be obtained for PB prediction based on sets of homologous sequences. Recently, Pei and Grishin combined evolutionary and structural information to predict local protein structures and emphasized the interest of predicting blocks.[67] The method they propose is very interesting and gives good results in terms of Q_3. Nevertheless, they use a large set of fragments for predictions rather than grouping these local structures according to their structural similarity.

Finally, an important point in this article involves the assessment of an index based on the Shannon entropy. We showed that this index is directly related to the quality of prediction and, more importantly, that local 3D approximation is correlated with it. The smaller this index, the less the uncertainty in the 3D local structures predicted. This index will be particularly useful in elaborating a new hierarchical strategy for prediction, focusing first on the regions with low N_{eq} values and progressively extending this prediction towards high N_{eq} regions.

Future work will focus on strategies for reconstructing protein structures from the PBs predicted. They may take into account the local quality of the prediction, that is, the N_{eq} index, and the PB overlaps. Reconstruction may be based on the use of predicted dihedral angles,[68] information that may be especially important in the loop regions.[69]

ACKNOWLEDGMENTS

C.E. and S.H. are Professors at the University Paris 7, Denis-Diderot, Paris. A.d.B. is a researcher at the French Institute for Health and Medical Research (INSERM).

REFERENCES

1. Liu WM, Chou KC. Singular points of protein β-sheets. Protein Sci 1998;7:2324–2330.
2. Bansal M, Kumar VL. HELANAL—a program to characterize helix geometry in proteins. J Biomol Struct Dyn 2000;17:811–819.
3. Pauling L, Corey RB. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc Natl Acad Sci USA 1951;37:235–240.
4. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci USA 1951;37:251–256.
5. 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.
6. de Brevern AG, Camproux A-C, Hazout S, Etchebest C, Tufféry P. Beyond the secondary structures: the alphabets. In: Sangadai SG, editor. Recent Advances in Protein Engineering. Vol. 1. Trivandrum, India: Reserach Signpost; 2001. p 319–331.

7. Karchin R. Evaluating local structure alphabets for protein structure prediction. PhD Comput Sci 2003.
8. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
9. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments models native protein structures accurately. J Mol Biol 2002;323:297–307.
10. Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 2000;40:662–674.
11. Hunter CG, Subramaniam S. Protein local structure prediction from sequence. Proteins 2003;50:572–579.
12. Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP. Clustering of protein structural fragments reveals modular building block approach of nature. J Mol Biol 21004;338:611–629.
13. Unger R, Harel D, Werland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5:355–373.
14. Prestelski SJ, Williams AL Jr, Liebman MN. Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results. Proteins 1992;14:430–439.
15. Unger R, Susman JL. The importance of short structural motifs in protein structure analysis. J Comput Aid Mol Des 1993;7:457–472.
16. 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.
17. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. J Mol Biol 1990;213: 327–336.
18. 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.
19. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence–structure motif. J Mol Biol 1998;281:565–577.
20. Camproux AC, Tufféry P, Chevrolat JP, Boisvieux J-F, Hazout S. Hidden Markov Model approach for identifying the modular framework of the protein backbone. Protein Eng 1999;12:1063–1073.
21. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271–287.
22. Camproux AC, Gautier R, Tufféry P. A Hidden Markov Model derived structural alphabet for proteins. J Mol Biol 2004;339:591–605.
23. Hunter CG, Subramaniam S. Protein fragment clustering and canonical local shapes. Proteins 2003;50:580–688.
24. Camproux A-C, Tufféry P, Buffat L, André C, Boisvieux J-F, Hazout S. Using short structural building blocks defined by a Hidden Markov Model for analyzing patterns between regular secondary structures. Theor Chem Acc 1999;101:33–40.
25. Camproux A-C, de Brevern AG, Hazout S, Tufféry P. Exploring the use of a structural alphabet for a structural prediction of protein loops. Theor Chem Acc 2001;106:28–35.
26. Fourier L, Benros C, de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 2004;5:58.
27. de Brevern AG, Hazout S. Compacting local protein folds by a "Hybrid Protein Model." Theor Chem Acc 2001;106:36–47.
28. de Brevern AG, Hazout S. Improvement of "Hydrid Protein Model" to define an optimal repertory of contiguous 3D protein structure fragments. Bioinformatics 2003;19:345–353.
29. Benros C, de Brevern AG, Hazout S. Hybrid Protein Model (HPM): a method for building a library of overlapping local structural prototypes. Sensitivity study and improvements of the training. IEEE Int Work NNSP 2003;1:53–70.
30. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D, Rossetta in CASP4: Progress in *ab initio* protein structure prediction. Proteins 2001;37:199–126.
31. Bonneau R, Strauss CEM, Rohl CA, Chivian D, Bradley P. Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.
32. Bystroff C, Shao Y. Fully automated ab initio protein structure prediction using I-Sites, HMMSTR and Rosetta. Bioinformatics 2002;18:S54–S61.
33. Tsai C-J, Maizel JV Jr, Nussinov R. Anatomy of protein structures: visualizing how a one-dimensional protein chain folds into a three-dimensional shape. Proc Natl Acad Sci USA 2000;97:12038–12043.
34. Tsai C-J, Nussinov R. The building blocks folding model and the kinetics of protein folding. Protein Eng 2001;14:723–733.
35. Tsai H-H, Tsai C-J, Ma B, Nussinov R. In silico protein design by combinatorial assembly of protein building blocks. Protein Sci 2004;13:2753–2765.
36. de Brevern AG, Valadié H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: a new approach to the sequence–structure relationship. Protein Sci 2002;11:2871–2886.
37. de Brevern AG, Benros C, Gautier R, Valadié H, Hazout S, Etchebest C. Local backbone strucgure prediction of proteins. In Silico Biol 2004;4:381–386.
38. Karchin R, Cline M, Mandel-Butfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 2003;51:504–514.
39. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of a secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
40. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
41. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
42. Noguchi T, Matsuda H, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). Nucleic Acids Res 2001;29:219–220.
43. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Databank. Protein Sci 1992;1:409–417.
44. Wang G, Dunbrack JL Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
45. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and strutures. J Mol Biol 1995;247:536–540.
46. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Res 2000;28:254–256.
47. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69.
48. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer-Verlag; 2001.
49. Rabiner LR. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc IEEE 1989;77:257–285.
50. Sippl MJ, Stegbuchner H. Superposition of three-dimensional objects: a fast and numerically stable algorithm for the calculation of the matrix of optimal rotation. Comp Chem 1991;15:73–78.
51. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
52. Labesse G, Colloc'h N, Pothier J, Mornon, J-P. PSEA: a new efficient assignment of secondary structure from Cα trace of proteins. Comput Appl Biosci 1997;13:291–295.
53. Rose GD, Gierasch LM, Smith JA. Turns in peptides and proteins. Adv Protein Chem 1985;37:1–109.
54. Chan AW, Hutchinson EG, Harris D, Thornton JM. Identification, classification, and analysis of β-bulges in proteins. Protein Sci 1993;2:1574–1590.
55. Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C, Saviano M. Discovering protein secondary structures: classification and description of isolated α-turns. Biopolymers 1996;38:705–721.
56. Michie AD, Orengo CA, Thornton JM. Analysis of domain structural class using an automated class assingment protocol. J Mol Biol 1996;262:168–185.
57. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
58. Bystroff C, Thorsson V, Baker D. HMMSTR: a Hidden Markov Model for local sequence–structure correlations in proteins. J Mol Biol 2000;301:173–190.

59. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
60. Humprhrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graph 1996;14:33–38.
61. Chou KC. Prediction of tight turns and their types in proteins. Anal Biochem 2000;286:1–16.
62. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7:21–38.
63. Steinbacher S, Baxa U, Miller S, Weintraub A, Seckler R, Hubet R. Crystal structure of phage P22 tailspike protein complexed with *Salmonella* sp. O-antigen receptors. Proc Natl Acad Sci USA 1996;93:10584–10588.
64. Kaur H, Raghava GPS. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci 2003;12:923–929.
65. Thompson MJ, Goldstin RA. Predicting protein secondary structure with probabilistic schemata of evolutionarily derived information. Protein Sci 1997;6:1963–1975.
66. Kloczkowski A, Ting KL, Jernigan RL, Garnier J. Comgining the GOR V, algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins 2002;49:154–166.
67. Pei J, Grishin NV. Combining evolutionary and structural information for local protein structure prediction. Proetins 2004;56:782–794.
68. Jurkowski W, Brylinski M, Konieczny L, Wiiniowski Z, Roterman I. Conformational subspace in simulation of early-stage protein folding. Proteins 2004;55:115–127.
69. Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. Bioinformatics 2004;20:1612–1621.