"Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence

A G de Brevern[1,*,a], C Etchebest[1,*], C Benros[1,2] and S Hazout[1,†]

[1] *INSERM, U726, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), Université Paris 7, case 7113, 2, place Jussieu, 75251 Paris Cedex 05, France*

[2] *CNRS, UMR 5086, Equipe de Bioinformatique et RMN structurales, IBCP UCBL, 7, passage du Vercors, 69 367 Lyon Cedex 07, France*

[a] *Corresponding author (Fax, 33-1-43-26-38-30; Email, debrevern@ebgm.jussieu.fr)*

The description of protein 3D structures can be performed through a library of 3D fragments, named a structural alphabet. Our structural alphabet is composed of 16 small protein fragments of 5 C_α in length, called protein blocks (PBs). It allows an efficient approximation of the 3D protein structures and a correct prediction of the local structure. The 72 most frequent series of 5 consecutive PBs, called structural words (SWs) are able to cover more than 90% of the 3D structures. PBs are highly conditioned by the presence of a limited number of transitions between them. In this study, we propose a new method called "pinning strategy" that used this specific feature to predict long protein fragments. Its goal is to define highly probable successions of PBs. It starts from the most probable SW and is then extended with overlapping SWs. Starting from an initial prediction rate of 34.4%, the use of the SWs instead of the PBs allows a gain of 4.5%. The pinning strategy simply applied to the SWs increases the prediction accuracy to 39.9%. In a second step, the sequence-structure relationship is optimized, the prediction accuracy reaches 43.6%.

[de Brevern A G, Etchebest C, Benros C and Hazout S 2006 "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence; *J. Biosci.* **32** 51–70]

1. Introduction

Until recently, the folded state of protein was classically described through the arrangement of secondary structures (2D): α-helix (Pauling and Corey 1951a), and β-strand (Pauling and Corey 1951b) and the non-periodic coil (non-α and non-β), (Eisenberg 2003). The simplicity of the description facilitated classification (Murzin *et al* 1995; Orengo *et al* 1997) and visualization (Sayle and Milner-White 1995; Humphrey *et al* 1996; Koradi *et al* 1996). Likewise, the sequence-structure relationship is easily tractable and generally strong enough for allowing 2D predictions with a high performance rate. For instance, prediction in three states using both neural networks and homology reaches now an accuracy rate close to 80% (Jones 1999; Petersen *et al* 2000; Pollastri *et al* 2002; Pollastri and McLysaght 2005). Lastly, comparison of predicted pattern of secondary structures with 2D pattern observed in known structures provides interesting tools for finding low-homology related proteins and extracting interesting functional features (Girod *et al* 1999; Geourjon *et al* 2001; Errami *et al* 2003). However, different limitations of such a description remain. For instance, depending on the assignment algorithms used, discrepancy exists about the extent of secondary structures and more precisely about the location of the Ncap or Ccap of the periodic regions (Colloc'h *et al* 1993; Cuff and Barton 1999; Fourrier *et al* 2004; Martin *et al* 2005). Moreover and more importantly, 50% of the protein structures are in the

Keywords. *ab initio*; bayesian prediction; local protein structures; structural alphabet

***** Both authors contributed equally to this work.

† Deceased.

Abbreviations used: PB, Protein block; PSOWs, preferential succession of overlapping structural words; SF, sequence family; SWs, structural words.

coil state, a state not defined *per se*. Some attempts were performed to overcome this simplified classification and to characterize more precisely the coiled state (Némethy and Printz 1972; Richardson *et al* 1978; Milner-White 1990; Sibanda and Thornton 1991; Ring *et al* 1992; Chan *et al* 1993; Rohl and Doig 1996; Wintjens *et al* 1996; Oliva *et al* 1997; Wojcik *et al* 1999; Espadaler *et al* 2004).

From few years, due to the increasing availability of 3D atomic data, a new view of 3D protein structures emerged and a more complete and accurate description of the 3D protein backbone was reached. In this context, different methods, casting off the need for defining repetitive structures, have been developed and proved their efficiencies (Unger *et al* 1989; Prestrelski *et al* 1992; Unger and Sussman 1993; Schuchhardt *et al* 1996; Fetrow *et al* 1997; Camproux *et al* 1999b, 2004; Hunter and Subramaniam 2003a; Tendulkar *et al* 2004; Sander *et al* 2006). The concept of structural alphabet, namely, a set of average protein fragments able to approximate locally the protein backbone with efficiency, thus appeared and became a very attractive and powerful description. Sequence-structure relationship can be extracted from the set of prototypes, allowing 3D-structure prediction from sequence (Bystroff and Baker 1998; Camproux *et al* 1999a, 2001; de Brevern *et al* 2000; Hunter and Subramaniam 2003b; Karchin 2003; Etchebest *et al* 2005). For a review about the structural alphabets, see de Brevern *et al* (2001). Recently, Pei and Grishin (2004) have emphasized the interest of blocks for predicting local protein structures. In the same way, the results of Tsai and coworkers (Tsai *et al* 2004) and Baker group (Chivian *et al* 2005) show clearly the interest of using local prototypes to design protein folds.

In a previous work (de Brevern *et al* 2000), we have defined a 16-states structural alphabet by using an unsupervised classifier close to the self-organizing maps (Kohonen 1982, 2001) and hidden Markov model (Rabiner 1989). Each state, called Protein Block (PB) is an average 3D-conformation of 5 C_α in length, which approximates locally the 3D protein backbone with an average root mean square deviation (*RMSd)* of 0.41Å (de Brevern 2005). They could be used with a good efficiency to compare protein structures that are encoded as sequences of PBs (Tyagi *et al* 2006). Karchin and co-workers have compared the PBs' features with 8 different structural alphabets. Their results have shown that PBs alphabet is the most informative one (Karchin *et al* 2003). In addition, a study has pointed out a specific dependence between the PBs and the physico-chemical properties of amino acids (de Brevern and Hazout 2000). A Bayesian probabilistic strategy gave a prediction rate equal to 34.4%; this rate was improved by an optimization of the sequence–structure relationship to 40.7% (de Brevern *et al* 2000). Moreover, PB series of more than 10 C_α length clustered by a fuzzy approach have shown a good structural approximation (Benros *et al* 2003; de Brevern and Hazout 2003). These PBs can be used in a structural homology search (de Brevern and Hazout 2001) and prediction (de Brevern *et al* 2005; Benros *et al* 2006). However, in these studies (de Brevern *et al* 2000, 2004), even if the PBs are overlapping, some geometrical incompatibility between the predicted PBs may occur because the prediction is performed independently for each PBs. Aiming to overcome such shortcomings, we concentrated on longer fragments still based on PBs description. Indeed, similarly to Fetrow *et al* (1997), we observed some over-represented successions of blocks. Focusing on 5-PBs series, we extracted, from an encoded 3D-structure databank (de Brevern *et al* 2002), the 72 most frequent series called structural words (or SWs). From the sequence–structure relationship observed in the SWs, we developed a strategy based on Bayes' rule, to predict the 3D-structure in terms of PBs from the sequence. Prediction rate reached 38%, a significant improvement compared to the previous study. Coupled with the prediction rate, we defined a confidence index of the prediction, derived from Shannon entropy (Shannon 1948). This index pointed out striking differences of the prediction rate along the sequence, i.e. some regions are easily predictable while others are not. Related to this observation, we propose, in the present work, a novel prediction method called "pinning strategy". This method rests on the overlapping features of SWs and aims (i) at first selecting the most predictable regions ('pinning') in terms of SWs, and then (ii) at extending the pinned regions on both ends until a predictability limit is reached. The principle therefore consists in assembling predicted SWs along a relevant pathway. We also test the potential interest of introducing homology for improving the prediction. Finally, we propose an index to estimate the relevance of the prediction. A large set of proteins is tested and improved prediction results are obtained in most cases. We present as an illustration the results obtained for the signal transduction protein of *Escherichia coli* and four homologous proteins.

2. Materials and methods

2.1 *Protein blocks and structural words*

To facilitate further reading of the manuscript, we briefly summarize the main points about the PBs and the SWs.

The PBs correspond to a set of 16 local prototypes of 5 residues in length (de Brevern *et al* 2000). They are based on (Φ, Ψ) dihedral angle description and are labelled from *a* to *p* (see figure 1 and figure 1 of Fourrier *et al* 2004). They were obtained by an unsupervised classifier similar to self-organizing maps (Kohonen 1982, 2001) and hidden Markov models (Rabiner 1989). The PBs set, denoted with letters from *a* to *p*, constitutes a structural alphabet (de Brevern

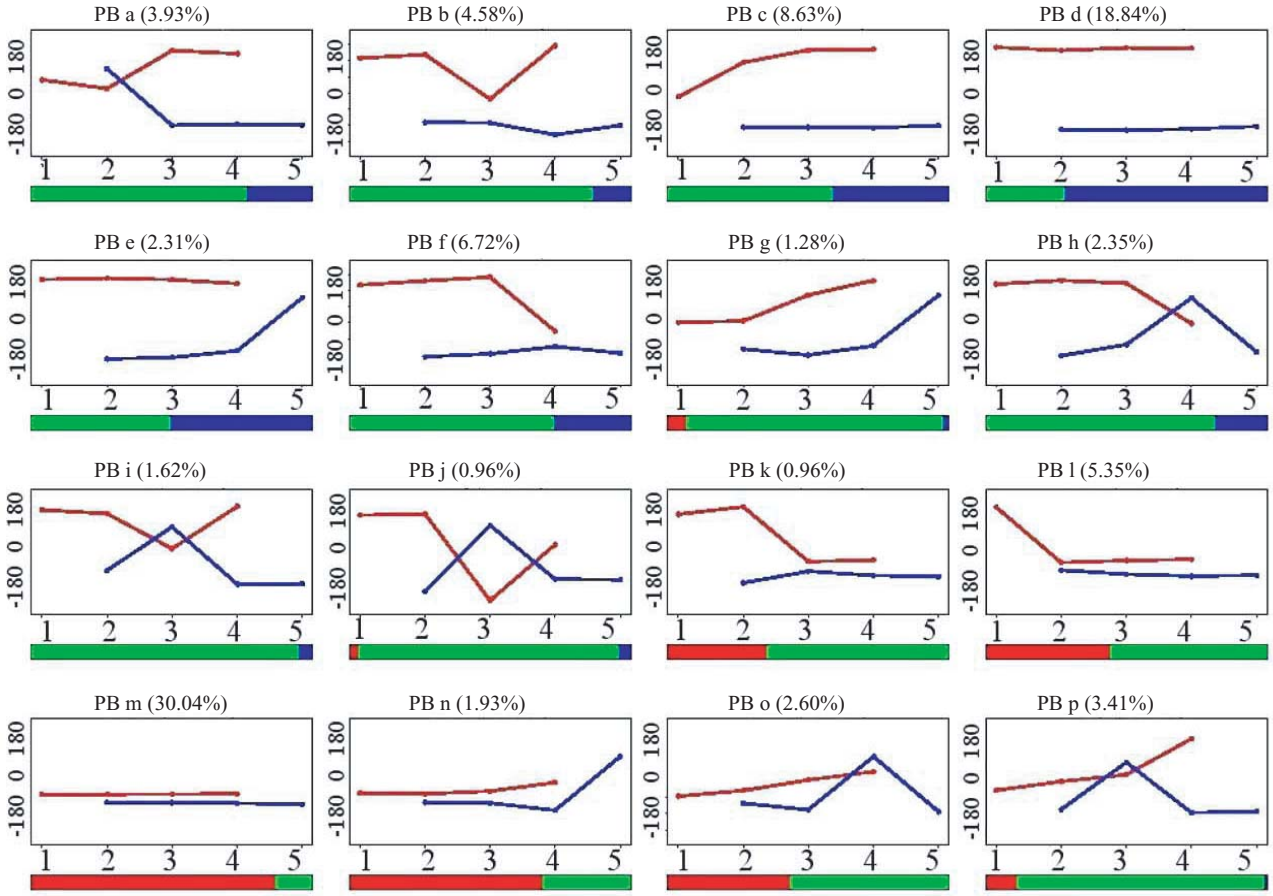


Figure 1. The backbone dihedral angle vectors (blue φ, red ψ) for the 5 C_α in the 16 PBs. Coloured bands (red for α-helix, blue for β-strand and green for coil) display the cumulated occurrence frequencies of the third residue in the secondary structures. These bands show the PBs location within secondary structures assigned according to a consensus approach proposed in Colloc'h *et al* (1993). It has to be noticed that the PBs description is more accurate than the conventional secondary structure description. A coarse correspondence between the two descriptions helps for comparison (see table 1 of de Brevern 2005) for more details).

et al 2001). This structural alphabet allows a correct approximation of local protein 3D structures with a root mean square deviation (*rmsd*) equals to 0.41 Å (de Brevern 2005). For more details, see *www.ebgm.jussieu.fr/~debrevern.*

The SWs (de Brevern *et al* 2002) are defined as the most frequent series of five PBs with an occurrence larger than 100 (cf. figure 2d to e, see also Appendix 1). They were obtained from the analysis of a non-redundant encoded databank composed of 1,403 proteins and 320,005 residues (see de Brevern *et al* 2002 and Etchebest *et al* 2005, for details and *http://www.ebgm.jussieu.fr/~debrevern/SWs*).

The analysis gives 72 most frequent SWs. They cover 92% of the residues of the non-redundant databank. For each SW, the 3D structural approximation has been analysed. As the transitions between successive PBs are highly specific, most of the SWs are overlapping. So we can build long continuous PB chains (see figure 8 of de Brevern *et al* 2002). The major features of the SWs may be summarized as: (i) their frequency varies between 17.3% and 0.18%, (ii) PB *m* (~ core of α-helix) and PB *d* (~ core of β-strand) are involved in many SWs (19 and 44 respectively), (iii) only one SW has no prefix and suffix SW, and 6 have only one prefix or one suffix SW, and (iv) the structural meaning of SWs is relevant with an average *rmsd* of 0.70 Å for 9 C_α length (de Brevern *et al* 2002).

The analysis of the sequences in the databank associated to the given SW (cf. figure 2f) permitted to deduce a sequence-structure relationship for each SW. It provided an amino acid occurrence matrix for each SW and for each position along the SW (cf. figure 2g). This matrix is the basis of the scoring schema, which is further used, in the pinning strategy (see below).

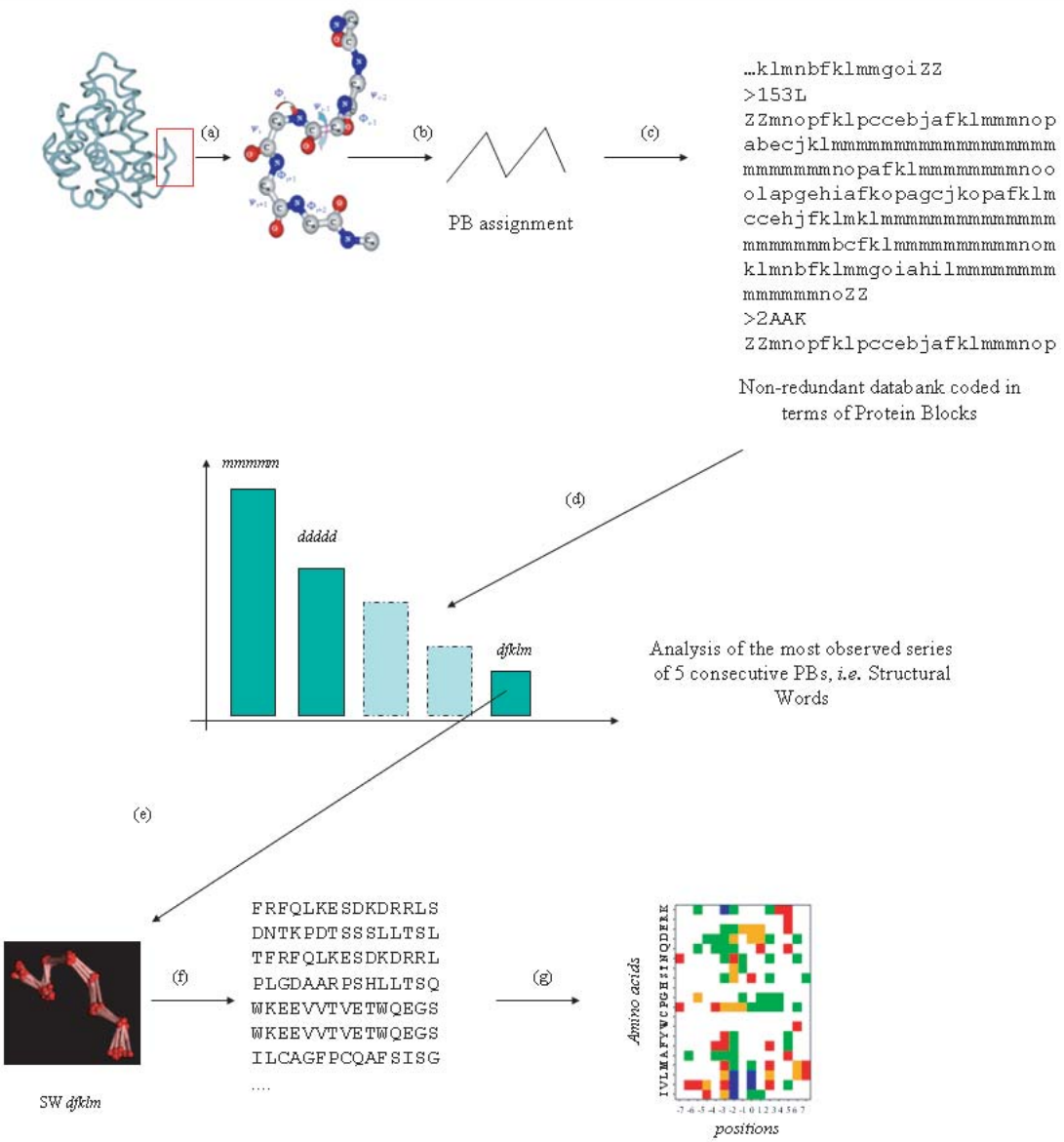


Figure 2. Coding the structural databank and computation of the amino acid matrices of the SWs. **(a)** Each protein residue is translated into dihedral angles (φ, ψ). **(b)** Fragment *f,* 5 residues in length and centered in position *i,* is selected as a succession of 8 dihedral angles (from ψ_{i-2} to φ_{i+2}). (**c**) The *rmsda* is computed between each PB and fragment *f*. The minimal *rmsda* value is selected, and allows assignment of the PB in position *i*. All the protein fragments are coded according to this approach (see de Brevern *et al* 2000). (**d**) From the distribution of the series of 5 consecutive PBs, the 72 most occurring SWs are selected and (**e**) analysed. (**f**) For each SW, the corresponding sequence fragments are selected and (**g**) used to compute an occurrence matrix.

2.2 *Principle of prediction by a "pinning strategy"*

Pinning strategy consists in searching for the "preferential succession of overlapping structural words" (PSOWs), namely, in chaining the predicted SWs in adequacy with the protein sequence studied (figure 3). The pinning strategy may be decomposed into three phases: (i) calcula-tion of an adequacy score matrix for all the fragments of a given length M composing the protein sequence, (ii) selection of the seeds, i.e. the SWs to be initially pinned, along the protein sequence, and, (iii) extension of the seeds into PSOWs.

First step – Assessing of the sequence-SW adequacy by a score matrix: Similarly to the ratio R_k used for the PBs prediction (see equation A1 in Appendix 1), we define a sequence-SW adequacy score S_p involving the amino acid properties of the structural word SW_p (p varying to 1 to N,

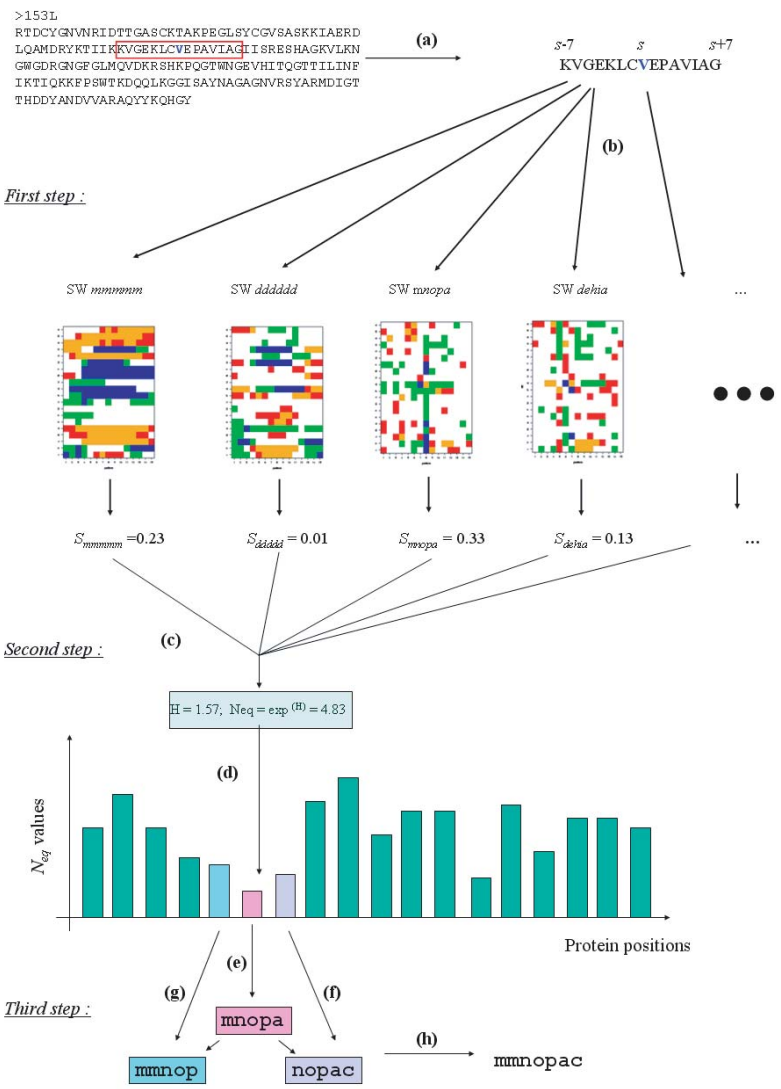


Figure 3. Pinning prediction strategy. (**a**) Assessment of the scores associated to the 72 SWs. To predict the local protein structure associated to a given position *s*, an amino acid fragment of length 2*l*+1 (*l* = 7, centered on the central residue *s*) is taken. (**b**) Then, each amino acid occurrence matrices associated to the SWs are used to compute a score based on Bayes' rule. (**c**) Location of the seed and definition of the structural pathways. From the prediction scores of the SWs, an entropy score can be computed and translated into an index called N_{eq} (number of equivalent). (**d**) The principle of the pinning strategy is to search for the lowest N_{eq} values. (**e**) In the example given, it corresponds to the position *i* associated with the SW *mnopa*. This initial position is called a seed. Then we search for the most probable SWs that overlap the SW at positions *s*–1 (i.e. prefix) and *s*+1 (i.e. suffix). So here, (**f**) SW *nopac* at position *s*+1 and (**g**) SW *mmnop* is found at position *s*–1. (**h**) This selection creates the beginning of a PSOW *mmnopac*. The process is iterated until no overlapping SW associated to a high score can be found.

with N =72 in our study) by:

$$S_p = \frac{P(X_s|SW_p)}{P(X_s)} = \frac{P(SW_p|X_s)}{P(SW_p)} . \quad (1)$$

The ratio S_p measures the extension of information provided by the amino acid chain X_s in the prediction of the structural word SW_p. The term $P(X_s \mid SW_p)$ is obtained from the occurrences of the amino acids in a given position of the sequence window X_s encompassing the structural word SW_p. We assume that the sequence window X_s (a_{s-l}, …, a_s,…, a_{s+l}) is composed of 2*l*+1 independent residues. Even if slightly crude, this assumption is necessary because an appropriate learning would require a dataset which is not yet available. Indeed, the database becomes rapidly sparse even at the dipeptide level.

Therefore, we have:

$$P(X_s \mid SW_p) = P(a_{s-l} \mid SW_p) \times \ldots \times P(a_s \mid SW_p) \times \ldots \times P(a_{s+l} \mid SW_p) \quad (2)$$

Each relative occurrence frequency $P(a_i \mid SW_p)/P(a_i)$ of a given amino acid *a*, located in the *i*th position of the

sequence window is computed as the ratio between the frequency of a_i observed in the SW_p and the frequency of a_i observed in the training databank.

As a result, a score matrix $F(s, p)$ is composed of the adequacy sequence-SW scores S_p computed for every structural word SW_p centered in the position s of the sequence.

Second step – Selection of the seeds with a score diversity index: The second step of the pinning strategy consists in locating along the studied protein the SWs (namely the 'seeds') from which we build the PSOWs. These positions s along the sequence are selected thank to a diversity index established from the score matrix. This index, we previously introduced (de Brevern *et al* 2000), is based on the information theory and estimates the predictability power of the sequence. It is defined as the exponential of the Shannon entropy $H(s)$ (Shannon 1948) of the probabilities $F^*(s, p)$ for a given position s with

$$F^*(s, p) = \frac{F(s, p)}{\sum_{q=1}^{N} F(s, q)} \quad (3)$$

and $H(s)$ defined as

$$H(s) = -\sum_{p=1}^{N} F^*(s, p) \ln F^*(s, p) \quad (4)$$

$$N_{eq}(s) = \exp [H(s)]. \quad (5)$$

This index denoted N_{eq}, for "equivalent number of structural words" varies between 1 and N, with N=number of structural words. For $N_{eq} = 1$, only one SW seems to be optimally adequate for the studied sequence window. For $N_{eq} = N$, the whole SWs are evenly distributed with respect to the adequacy scores; in this case, no SW prediction can be performed. Consequently, a low SW_p -value indicates a high information content of the sequence relative to the local protein structure, hence a high predictability level.

Therefore, the second step of the pinning strategy consists in defining a list where the positions, along the sequence, are ranked in ascending order according to the N_{eq} values. The first selected position s^* corresponds to argmin$[N_{eq}(s)]$. In this location, the top-scoring structural word (noted SW^*) is then considered, defining the first seed.

Its associated index and adequacy score are p^* = argmax $[F(s^*, p)]$ and $F_{max} = F(s^*, p^*)$ respectively. For example, in figure 3e, the seed is the SW *mnopa* corresponding to a C-cap α-helix (see Appendix 2 for another example).

Third step – Extension of the seeds: The location s^* is then extended into a series of overlapping SWs in maximum adequacy with the sequence and with the selected word SW^*. We define the terms of prefix words and suffix words. A structural word SW_o (respectively SW_q) is a prefix word (resp. a suffix word) of another word SW_p when the 4 last PBs of SW_o (resp. the 4 first PBs of SW_q) are identical to the 4 first PBs of SW_p (respectively the 4 last PBs). Prefix (resp. suffix) of SW_p^* is defined as (i) the SW with the maximum adequacy score for the selected position o (resp. q) and if (ii) this score is higher than a given user-defined threshold F_0. For instance in figure 3f, the optimal suffix word is *nopac*. The selection of the optimal prefix word of SW_p^* is carried out similarly. For position s^*-1, in figure 3g, the prefix SW *mmnop* is found. The extension is performed at the extremities of PSOWs while the scores of the suffix and prefix words are higher than a user-defined threshold F_0.

When no further extension is possible from a given seed, a new seed is selected from the list (step 2). If the chosen seed overlaps some suffix or prefix picked in the extension process, the next seed in the list is then considered. The process is iterated until a new seed with no overlapping is obtained.

The pinning processing is finally stopped when no new seed can be introduced or when the N_{eq} -value for any seed exceeds a given threshold N_{eq0}. Thus, the "pinning strategy" only depends on two threshold parameters: N_{eq0} for selecting the seeds and, F_0 for limiting the extension. With this prediction strategy, we reject *a fortiori* the regions of the protein sequence where the sequence-SW adequacy is too low. We have fixed N_{eq0} to a high value (e.g. 15) to provide an appropriate number of seeds.

To assess the quality of the strategy, we define two measures. The first one compares the number of predicted PBs with the true total number of PBs, i.e. the sequence length minus 4. This value will be called "the covering ratio" in the following. If all the PBs of the sequence were predicted, the covering value would be 100%. The second measure is the prediction rate itself Q_{16}, defined as the ratio of PBs, correctly predicted, over the total PBs predicted for the protein (those defining the "covering ratio").

2.3 *Improved prediction by using the sequence families*

In our previous works, we introduced a procedure able to improve strongly the prediction rate (de Brevern *et al* 2000; Etchebest *et al* 2005). This procedure lies on the observation that many different sequences may be associated to a given local fold ("*n* sequences for one fold"). Indeed, the learning process yielding to the definition of PBs (and consequently to SWs) is based on pure geometric considerations, the sequence-structure relationship being deduced *a posteriori.* The matrix $P(Xs \mid SWs)$, expressing the sequence-structure relationship along the window s, results in fact, from the superposition of n different sequences folding in the same

state. The sequence family (SF) concept thus aims at detecting and grouping the similar groups of sequences in the set of sequences folding in a given state, and accordingly defining new specific sequence-structure relationships for the given local fold.

In our previous approach, the procedure was based on the proximity of a sequence window to a profile, i.e. the occurrence frequencies of amino acids in the different positions of the window. The new approach is now based on the proximity of a sequence window to a reference sequence window specific to each sequence family. It is similar to the conventional clustering method, *k*-means (Hartigan and Wong 1979).

The process starts with *g* groups represented by a sequence chosen at random. Sequence windows are shared among the *g* groups using a distance based on a similarity score. This similarity score is computed using the BLOSUM62 matrix (Henikoff and Henikoff 1992). Then, inside each group, the sequence whose sum of distances with respect to the other fragments is minimal, is selected (the barycentre). The whole process is iterated until no modification of the *g* reference sequence windows is observed. Finally, the whole sequence windows are distributed into the groups, then for each SW, the *g* occurrence matrices are defined. The number *g* of sequence families depends on the SW frequency in the training databank. In practice, the most frequent SWs are split into 2 to 15 clusters.

In the building of the score matrix, only the maximal score $F(s, p)$ upon all the sequence families of the same SW_p is conserved. In addition, to assess the quality of the prediction, we perform a cross-validation consisting in splitting the set of proteins into five subsets, 4 for the training, and the last subset for the validation. So the learning subset ranges from 572 to 577 proteins and the validation subset between 140 and 145 proteins. For each subset, the SFs are re-defined. Depending on the cross validation step, the total number of occurrence matrices associated with the 72 SWs may vary between 124 and 129. Twenty-four SWs are associated with sequence families: SW *mmmmm* with 15 sequence families, SW *ddddd* with 8, 3 SWs with 4, 3 SWs with 3, and 15 SWs with 2 SFs.

2.4 *Definition of an index for assessing the prediction accuracy*

The amino acid distribution is the basis of any prediction method. Most strategies developed for predicting 3D structure from sequence are confronted to a dilemma: (i) the necessity to trap the main sequence determinants of the structure; that supposes a large set of data of being available; and (ii) to predict protein sequences with, in fact, many singularities; that means sequences with features that may be far from the average properties of the learning dataset. In many cases, the prediction rates obtained are inhomogeneous and depend on the sequence examined. Here we propose a relationship able to estimate the expected prediction rate using few parameters. The method is based on linear multiple regression using the following parameters;

(i) $\Delta f^R(i)_k$, the relative variation of frequencies equals to $\Delta f(i)_k/f(i)$ of amino acids with $\Delta f(i)_k = f(i)_k - f(i)$) where $f(i)_k$ and $f(i)$ denote respectively the frequency of the amino acid i in the protein k and in the database.

(ii) P_{Neq}^{5}, the relative proportion of N_{eq} positions with $N_{eq}<5.0$ computed as:

$$P_{Neq}^{5}=\Delta f_k(N_{eq}<5)/<f(N_{eq}<5)>,$$

with

$$\Delta f_k(N_{eq}<5)= f_k(N_{eq}<5) - <f(N_{eq}<5)> f_k(N_{eq}<5)$$

is the proportion of N_{eq} smaller than 5 in the protein k and $<f(N_{eq}<5)>$ mean proportion of N_{eq} smaller than 5, 10% in our study.

(iii) F_{50}, the relative proportion of positions associated to a score > 50.

$$F_{50} =\Delta f_k(F_{max})/<F_{max}>,$$

with F_{max}, the maximal score (see above),

$\Delta f_k(F_{max})=f_k(F_{max}) - <f(F_{max})>$ with $f_k(F_{max})$ proportion of F_{max} larger than 50 in the protein k and $<f(F_{max})>$ mean proportion of F_{max} larger than 50, 31% in our study.

The multiple linear regression performed gives a significant multiple correlation ($R^2 = 0.485$, P-value < 1.10^{-9}).

The final equation obtained is:

$$Q_{PSe} = 43.50 - 3.56\,I - 3.61\,V - 4.26\,L - 0.92\,M - 0.23\,A - 0.41\,F - 3.97\,Y - 1.71\,W - 1.45\,C - 4.78\,P - 10.07\,G - 0.27\,H - 5.02\,S - 4.85\,T - 1.06\,N - 1.45\,Q - 6.59\,D - 4.02\,E - 1.14\,R - 0.72\,K + 6.73\,P_{Neq}^{5} + 10.12\,F_{50}.$$

Thus, for a given sequence, we can compute an *expected* prediction rate using the amount of each amino acid in the sequence, and the values F_{50} and P_{Neq}^{5}, also related to the characteristics of the sequence.

3. Results

3.1 *Efficiency of the structure prediction by a pinning strategy*

The results obtained with the pinning strategy without and with the sequence family approaches are given in table 1. As a reference, we compute the prediction in terms of PBs and in terms of SWs applying the Bayes' rule, similarly to our previous works (de Brevern *et al* 2000, 2002; Etchebest *et al* 2005). The prediction rate Q_{16} of the Bayesian approach with the PBs is 34.4%. This value is unchanged compared to the initial work, while the size of the databank is doubled (from

Table 1. Prediction rates. The different prediction using the Bayesian prediction with the PBs or with the SWs, and, with the pinning strategy with or without SFs.

	Prediction with PBs		Prediction with SWs		
Prediction	Bayes' rule	Bayes' rule	Bayes' rule	Pinning strategy	Pinning strategy
Sequence families	No	Yes†	No	No	Yes
Q_{16} value (%)	34.4	40.7	38.9	39.9	43.6

† The SFs used are the ones defined in a previous work (de Brevern *et al* 2000).

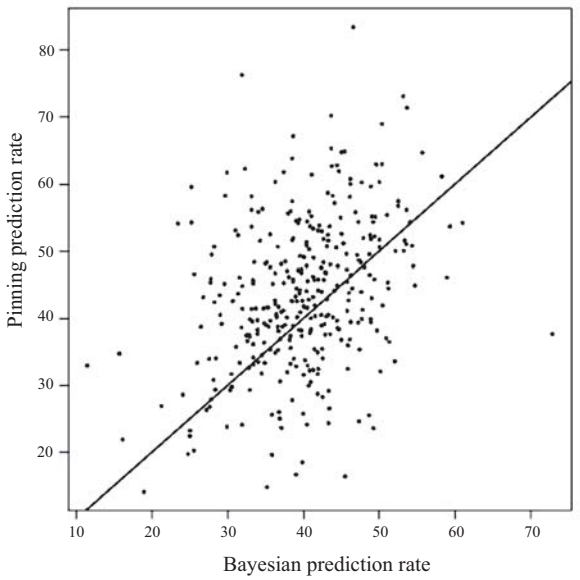


Figure 4. Improvement of the prediction rate using pinning strategy. The figure shows the comparison of the Bayesian prediction rate with the SWs (x-axis) and the pinning prediction with the SWs including the sequence families approach (y-axis). The values are in percent.

86,628 to 180,854 residues). When SWs are considered, Q_{16} increases to 38.9%. The pinning strategy with SWs yields an improvement of 1% for the Q_{16} value, compared to the SW Bayesian approach. These values are obtained with N_{eq} and F_0 optimized parameters (see §2). The N_{eq} parameter that controls the choice of the seed is fixed to 15, while, F_0 devoted to the extension is fixed to 4 (see §4). Finally, when sequence families are considered (see Appendix 3 for some details), the pinning strategy efficiency reaches 43.6% on average. A cross-validation procedure used for assessing the reliability of the sequence families shows a significant improvement, in any cases. Indeed, depending on the considered databank subset, the Q_{16} values range between 42.7% and 45.3% for a covering range of [81.4%–86.3%]. Thus, the pinning strategy with the SFs has Q_{16} values higher of 9.1% and 4.5% compared to the prediction of the PBs and SWs respectively. Interestingly, these values are associated to an average covering range of 85%, i.e. only 15% of the residues are not predicted. If we compare these values with the distribution of periodic structures *m* and *d* (30% and 19% respectively), we clearly show that the prediction is not biased towards repetitive structures.

In addition, the comparison of prediction rates between the Bayesian prediction with SWs and prediction rate of the pinning strategy with the SFs (figure 4) shows that 64.3% of the studied proteins get a positive gain. We observe that the Q_{16} improvement is not correlated to the initial prediction rate, and not related to the size of the protein or the protein structural class (all-α, all-β, $\alpha\beta$ others).

3.2 *Examples of a structure prediction by the pinning strategy*

Among the 717 proteins tested, the signal transduction protein of *Escherichia coli* (PDB label: 3chy, see Appendix 4) of 128 amino acids is representative of a large number of proteins. We present here the prediction using the pinning strategy and illustrate the role of different parameters. Moreover, we explore the putative contribution of homologous sequences on the structure prediction.

3.2a *Prediction:* Figure 5 shows the predictions of the protein structure from the sequence by the pinning strategy. Figure 5a gives the sequence, the classical secondary structure description, the coding in terms of PBs, the prediction of PBs with our previous approach (LocPred, de Brevern *et al* 2004) and the N_{eq} of pinning strategy. Figure 5b shows the results of pinning approach extending from the lowest N_{eq}, step-by-step with the seeds and the structural pathways. In parallel, are given the resulting covering ratio and Q_{16} value.

When the covering ratio increases until 87.1% (106 residues) (figure 5b), the prediction rate Q_{16} progressively decreases to reach a plateau close to 68.8%, namely 73 PBs among the 106 predicted PBs. The first seed is centered in position 27 with the SW *mnopa* [positions 25 to 29]. An extension of 29 PBs $kbc_2d_4fklm_{12}nopabd$ is obtained with 24 correctly predicted PBs [positions 3 to 31]. Extension is mainly located towards the prefix ends. Compared to the true assignment (figure 5a), the prediction is locally false since the triplet *fkl* [positions 15 to 17] is shifted to one position

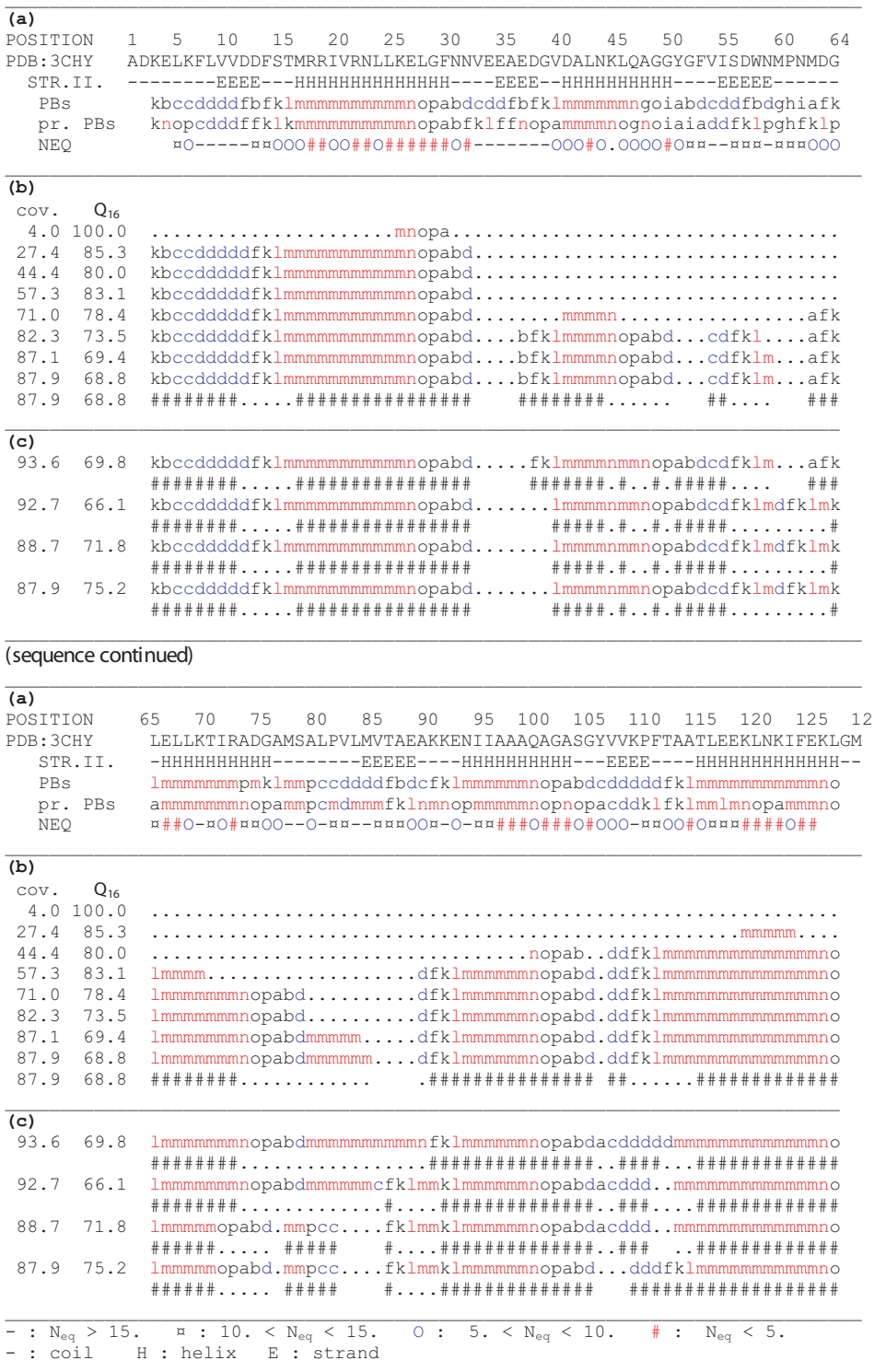
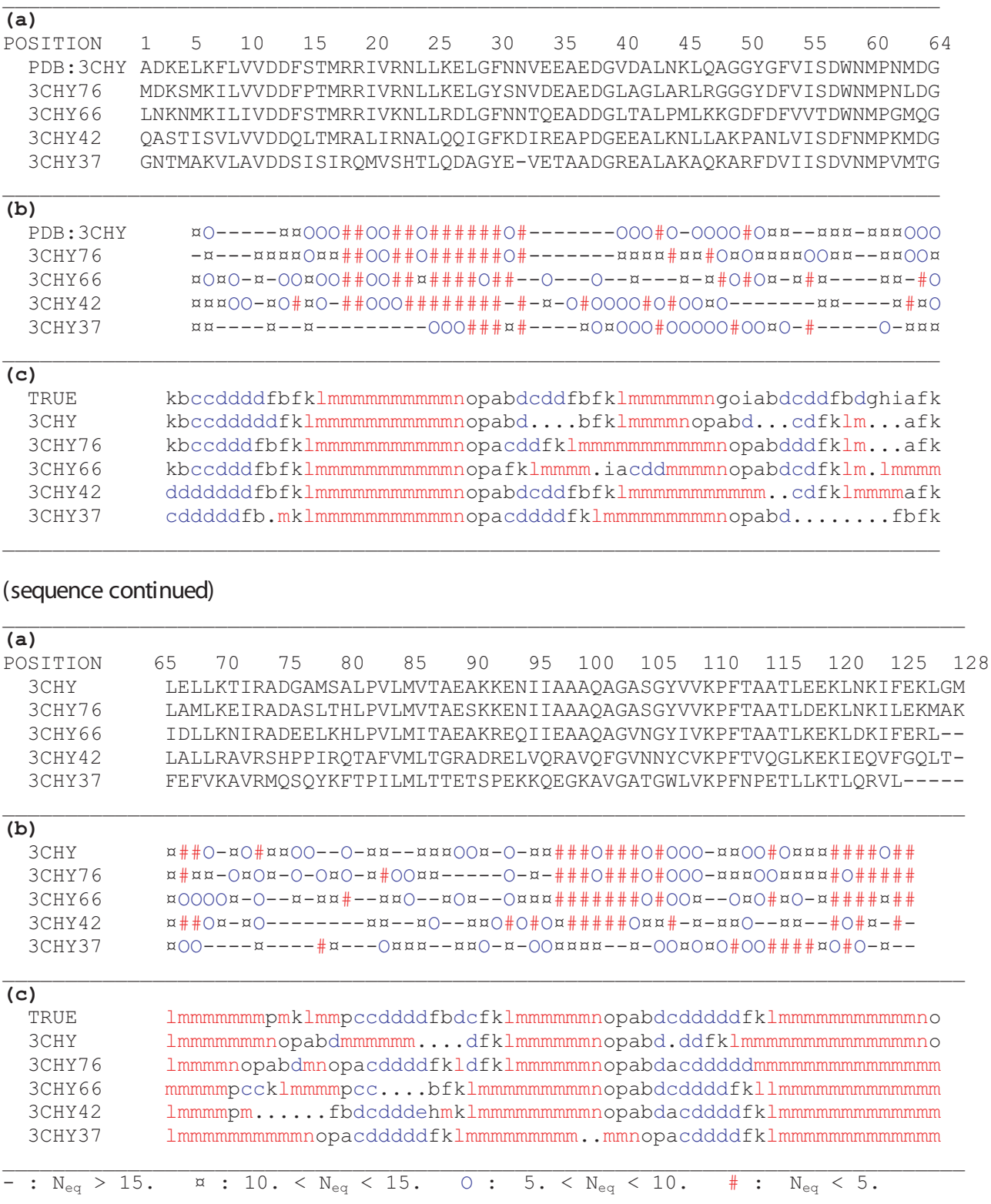


Figure 5. Building of the PSOWs the pinning strategy for the signal transduction protein of *Escherichia coli.* (**a**) The sequence of *3chy* is given with the consensus secondary structure assignment (Colloc'h *et al* 1993), the PBs assignment, the PBs prediction done with LocPred software (LocPred; de Brevern *et al* 2004; Alland *et al* 2005) and the N_{eq}-values associated to the prediction with SWs. (**b**) Selection of the different seeds and extension for $N_{eq0} = 15$ and $F_0 = 4$. On the left part of each line are given in percent the extent of the covering ratio (cov.) of the amino acid sequence and the prediction rate associated (Q_{16}). (**c**) Final results obtained for $F_0 = 8$, 12, 20 and 24. F_0 is a parameter controlling the extension of the seed (i.e. the initial structural word) into structural pathways. # corresponds to a correctly predicted PB.



(sequence continued)

Figure 6. Prediction for four homologous sequences of 3chy. The sequence identity equals to 76%: SwissProt code (Bairoch *et al* 2004) *p94342* (labelled 3chy[76]), 66%: *q9kd5* (3chy[66]), 42%: *q87718* (3chy[42]), and 37%: *q9kkk8* (3chy[37]). (**a**) multiple alignment of the five sequences (CLUSTALW (Thompson *et al* 1994)), (**b**) the variation of N_{eq}-value along each sequence and (**c**) the predicted PBs for N_{eq0}=15 and F_0=4.

backward. Indeed, a PB *l* is predicted at position 14 instead of position 15. This shift is already present when the original procedure (figure 5a, LocPred) based on Bayes' rule is considered. The second seed is centered in position 120 with the SW *mmmmm* [positions 118 to 122]. It leads to a 21-PBs chain $d_2fklm_{14}no$ with 15 correctly predicted PBs. In this last extension, the repetitive *m* region is too long compared to the true assignment. It is important to notice that this second seed is correctly predicted compared to the Bayes' prediction with PBs, which gives the PB series *mnopa*. Conversely, the extension to the left is wrongly predicted whereas the PB prediction predicts correctly the triplet *fkl* at positions 111 to 113.

Generally, we observe that the seeds selected in the first ranks give correct extended chains. PBs *m* are in most cases well positioned. Some regions are not predicted, e.g. positions 32 to 36 and the fourth β-strand (absent in the prediction due to high N_{eq} values) [positions 84 to 88], or wrongly defined, e.g. the N-cap of the fifth α-helix [position 114] (the α-helices are generally predicted longer).

In figure 5c, we report the impact of the extension threshold F_0 on the prediction results with F_0 values equal to 8, 12, 20 and 24. The prediction rate varies from 66.1% to 75.2% for a covering ratio extending from 87.9% to 93.6% (corresponding to 7 additional PBs). Overall, we observe a large agreement between the predictions. However, two regions are particularly variable. The first one extending from residues 32 to 35 (PBs *cddf*, an extended structure) corresponds to a region with high N_{eq}-values. Thus, in this region, different SWs are equally compatible with the sequence. This may be due to the presence of 3 glutamic acids in positions 33, 34 and 36, an occurrence probably sufficiently rare for explaining the weakness of the sequence-structure relationship. The second region located between residues 85 to 88 (PBs *dfbd*) corresponds to a zone where the results are strongly dependent on the parameters F_0 chosen.

This example is quite satisfying: the sequence is enough informative for predicting a large part of the 3D structure with a convenient reliability. Thus, the prediction rate is around 70% for a covering ratio close to 85%. This result corresponds to a strong improvement compared to the results obtained using the Bayes' rule that gives a prediction rate of 53% for the same protein.

3.2b *Analysis of homologous proteins of 3chy:* It is generally observed that homology strongly improves the success rate when used in the prediction paradigm. In the present case, we have tested the putative gain on the prediction when homologous sequences are considered. We present here the results obtained for the 3chy protein that is representative of a large number of proteins. In this case, the pinning strategy was applied to four homologous proteins. The homologous sequences were selected from SwissProt database (Bairoch *et al* 2004) (codes: *p94342*, *q9kd5*, *q87718*, *q9kkk8*) using BLAST (Altschul *et al* 1990), and aligned with CLUSTALW (Thompson *et al* 1994). The identity between *3chy* and *p94342*, *q9kd5*, *q87718*, *q9kkk8* equals respectively to 75.8%, 65.9%, 41.7% and 36.6%. The sequences will be noted $3chy^{76}$, $3chy^{66}$, $3chy^{42}$ and $3chy^{37}$ in the following.

In figure 6, are given the Neq values along the sequence for the *3chy* and its 4 homologous sequences as well as the corresponding predictions in terms of PBs obtained with the pinning strategy.

The following remarks may be pointed out: (i) The N_{eq}-variations are very similar between the studied protein and its homologous ones, apart the last one $3chy^{37}$. In the regions corresponding to PBs *m*, the sequence information level is high (small N_{eq} value). Interestingly, some regions are more informative (low N_{eq} values) in the homologous sequences than in the true 3*chy,* for instance, the N-ter region of $3chy^{66}$ and $3chy^{42}$, or the region located in the positions 36-53 for $3chy^{37}$. (ii) As expected, $3chy^{76}$ give close results to *3chy*. (iii) The covering ratio is similar in the four cases with a value close to 88%. (iv) A large similarity in the prediction is observed apart the region 70–89 where the predictions are ambiguous. A very interesting result concerns the PBs *d* in the region 80–89, missed when the sequence of *3chy* itself is examined, but may be recovered when considering three homologous proteins. We observe that among the five helical regions, three are correctly predicted [positions 16 to 25, 94 to 99 and 113 to 127] and are associated to low N_{eq} values. The two last ones show more confusing results: (i) positions 39 to 45 correspond to an extended structure in $3chy^{66}$ with high N_{eq} values (low predictability), and (ii) for positions 66 to 72, the length of the helices largely differ. Nonetheless, it is possible to find correct consensus prediction with a simple majority rule. For instance, a helical structure is found in positions 39 to 45 for four sequences among the five studied. We can make the same remarks for the extended structures *dddd* in positions 82 to 85 that are found for $3chy^{76}$, $3chy^{42}$ and $3chy^{37}$ but absent for *3chy*.

In conclusion, the analysis of the homologous sequences seems to bring additional information for the prediction of the structure of the studied protein. The use of different homologous sequences permits to point out sequence regions that appear more informative for predicting the 3D structure. This encouraging result must however be consolidated by a more detailed study.

3.3 *Prediction reliability*

The different prediction methods based on sequence information (*ab initio* methods) are frequently faced to a main difficulty: the prediction accuracy may depend on the protein examined. Here we propose a new approach aiming at estimating the relevance and the confidence of the

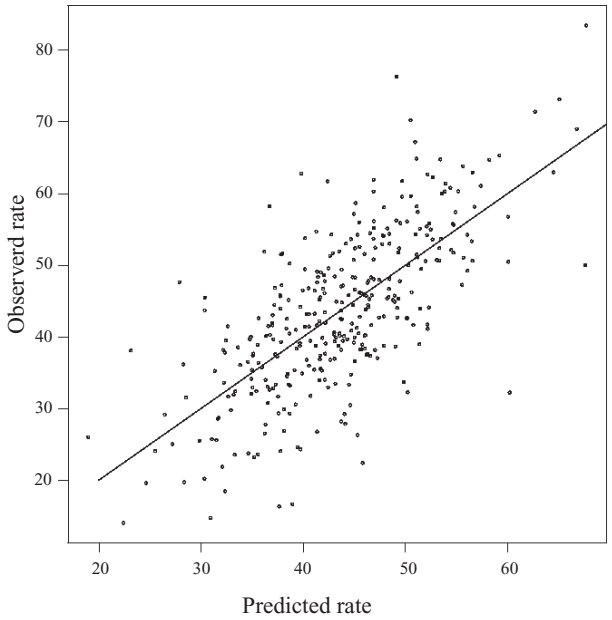


Figure 7. Estimation of the prediction rate. Comparison of the *expected* prediction rate Q_{PSe} computed with a multiple regression equation and the observed prediction rate Q_{16} (in percent).

prediction of the pinning approach. This method combines through a multiple linear regression *a priori* and *a posteriori* parameters: the amino acid composition of the protein (noted $\Delta f^{R}(i)_{k}$), the proportion of positions where the N_{eq} is less than a threshold fixed at 5.0 (noted P_{Neq}^{5}) and the proportion of positions associated to a score larger than a threshold fixed at 50 (noted F_{50}). A significant multiple correlation, $R^2 = 0.485$, (P-value < 1.10^{-9}) is obtained. The equation of the regression is given in the caption of figure 7. From this equation, an *expected* prediction rate was deduced, for each sequence in the databank. The value was compared to the *observed* prediction rate for the same sequence. For the whole set of proteins tested, the results show a strong linear correlation between the two measures. Consequently, we get *a priori* a confidence index of prediction reliability when the true 3D structure of the studied protein is unknown. We also performed a sensitivity analysis of this confidence index to the parameters used and we showed that (see Appendix 5) most of information relies on the two first parameters (P_{Neq}^{5} and F_{50}).

4. Discussion

In the present section, we discuss the impact of different parameters in the prediction efficiency of protein structure by the pinning strategy, such as the definition of a SW library, the number of sequence families, and, the values of the threshold parameters.

4.1 *Definition of a SW library*

First, the length of SWs (five PBs) was chosen to ensure both a structural meaning and a tractable number of prototypes to perform a prediction method. A smaller length leads to a smaller combination of PBs and a more accurate structural description. For a similar extent of the covering zone (around 92%) with series of three PBs, 64 motifs are only needed. Nevertheless, in terms of the pinning strategy the use of smaller words implies an increase of the number of possible transitions from a given word, i.e., the number of possible overlapping suffix and prefix words. Moreover, the amino acid distribution may be less specific and could lead to a lower rate of prediction. Conversely, a longer size of the structural words provides a reduction of the occurrence frequencies due to a larger combination of PBs, and an increasing structural variability. This lengthening provides *a fortiori,* a reduction of the mean numbers of suffix and prefix words per SW. Accordingly, the pinning strategy becomes less efficient particularly in the structural pathway building. Thus, the length chosen (i.e., 5 PBs corresponding to 9 residues) represents an appropriate compromise allowing a correct structural variability (*rmsd* ranged between 0.34Å and at most 1.11Å) and an accurate amino acid distribution specificity per SW. This choice is enough relevant for yielding a prediction rate close to 44% for long series of PBs. In addition, the covering ratio is highly dependent on this number ranging from 55% for 4 SWs, 80% for 20 SWs and 92% for 72 SWs. From this number, the value of the covering ratio reaches a plateau of 93% for 140 SWs. Therefore, a library of 72 SWs allows an almost maximal covering ratio and ensures a number of transitions between SWs large enough to build the PSOWs.

4.2 *Number of sequence families*

As shown previously, the use of sequence families permits to improve the correct prediction rate thanks to an optimal splitting of the sequence window set associated with a given SW. The resulting subsets (denoted by SF-sets) are accordingly more homogeneous in terms of amino acid composition. It must be noted that the number of SF-sets plays a major role in the structure prediction improvement. The number of SF-sets was chosen according to the following constraints: (i) the number of sequences window in each SF-set must be similar whatever the own frequency of SWs; (ii) this number must be large enough for building an occurrence matrix (>100 sequence windows per subset); (iii) the amino acid specificities per position must be more strongly marked; (iv) a significant improvement of the prediction rate must be obtained. The numbers of SF-sets per SW chosen here obey all the enacted rules, ensuring a

better balance between SF-sets for a given SW. In addition, the amino acid specificity is more marked, what explains the prediction rate gain.

4.3 *Choice of the threshold parameters in the pinning strategy*

The N_{eq0} parameter mainly controls the choice of the seed but only partially its extension into a structural pathway. For small N_{eq0} (for instance N_{eq0}<5), the prediction rate is more accurate as noted by the high weighting value of P_{Neq}^{5} in the multiple regression of the accuracy rate (see §3.5). Conversely, the extent of the prediction region is smaller. Unlikely, for large N_{eq0}, the covering ratio increases to the detriment of the prediction rate. The parameter F_0 is in contrast only devoted to the extension and mainly governs the covering ratio. This value may change drastically according to the F_0-values (see figure 5); but the prediction rate is slightly modified. Finally, the threshold parameters we chose, (N_{eq0} = 15 and F_0 = 4) results from an optimal compromise between a maximal prediction rate (close to 44%) and a large covering ratio (around 85%).

5. Conclusion

In the present paper, we propose at the same time a new approach for predicting the local 3D structure of proteins but also an index aiming at estimating the quality of the prediction.

One of the main choices carried out in the pinning strategy lies in the selection of a seed based on a minimal N_{eq} and coupled to a maximum score value. We have assessed the fact that the N_{eq} is an efficient index to start the construction of the pathways, i.e. always associated to a correct SW. The pathway extends by choosing new words with a maximum score value and compatible with the pathways previously defined. The use of SWs, focusing on the most observed local folds, improves the sequence-structure adequacy. The pinning strategy takes advantage of the SW overlapping**,** so facilitating the building of the different continuous pathways. The need for compatibility implies that the final prediction is not necessarily the optimal solution for whole SWs. Nonetheless, the strategy provides an efficient way for 3D reconstruction. Indeed, the predicted SWs correspond to 3D protein fragments in the structural databank. Finally, even if the improvement of the final prediction rate is limited, the procedure provides a good rate of prediction, close to 44% for an average covering ratio of 85%, with local fragments structurally quite compatible. Furthermore, we show that homology could provide significant improvement for the prediction rate.

As a final point, we address an important aspect of the prediction, never previously considered, namely a quantitative way for assessing the prediction accuracy. The index is only based on the amino acid composition of the sequence and the corresponding distribution of the scores.

Additional improvements are possible. Indeed, as noticed by one of the reviewer, the choice of prefix (or suffix) SWs with last 4 PBs strictly identical to the ends of the SW selected seed, is quite restrictive. SWs (or PBs) actually can be structurally similar. Accordingly, we could compute a structural confusion matrix between PBs (or SWs) and thus select among fuzzy prefixes (suffixes) with a high adequacy score.

In conclusion, the field of pinning strategy is very large: it could be applied as a preliminary step to detect far homologous proteins in protein homology modelling. It could constitute an alternative way to threading approaches or at least, speed the detection of the appropriate candidates for threading. It surely could be of great interest in *de novo* modelling. Indeed, the use of dihedral angles – basis of the PB definitions – is quite interesting to explore folding process (Jurkowski *et al* 2004); this information is of particular interest outside the repetitive structures (Kuang *et al* 2004).

In the future, we would like to improve the pinning strategy by a fast and efficient procedure of optimal pathways defined based on, for instance, a dynamic programming algorithm (conventionally used in sequence alignment). The procedure has already been applied in a similar context (Benros *et al* 2004). However, many difficulties remain that indeed could make this strategy hard to use, even if preliminary results were promising. Alternatively, a greedy algorithm could be tested and would allow us to find sub-optimal structural pathways. Other important improvements are in progress and concern mostly the development of an appropriate algorithm for ensuring more accurate successions of SWs. It should in particular take advantage of the amino acid content of homologous sequences of the studied protein. It could be also interesting to analyse the PSOWs in regards to existing description of protein structures such as super secondary structures (Efimov 1997), Protein Units (Sowdhamini and Blundell 1995; Gelly *et al* 2006) or domains (Alexandrov and Shindyalov 2003).

Acknowledgements

This paper is dedicated to the memory of Pr. Serge Hazout. This work benefited from grants from the Ministère de la Recherche and from “Action Bioinformatique inter EPST” number 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.”).

Appendix 1. Structural words and prediction

A1. *"Structural words": over-represented PB series*

From the encoded structure set, we have extracted the over-represented series of 5 PBs corresponding to local backbones composed of 9 C_α from the whole PB chains (cf. figure 2d to 2e). Each series is defined as a SW. Only 72 SWs with a frequency more than 0.18% have been selected and constitute a library of reference fragments. 92% of the encoded protein structures are recovered by these SWs. The SW number is very small compared to the 16^5 expected combinations. It points out the high dependence between the PBs. The structural stability of the SWs is assessed by the *rmsd*. It ranges between 0.34 Å and 1.1 Å (de Brevern *et al* 2002).

A2. *Prediction by a Bayesian probabilistic approach*

We recall the principle of prediction of the encoded structure protein from the sequence used in the previous works (de Brevern *et al* 2000, 2002, 2004; Fourrier *et al* 2004; Etchebest *et al* 2005). In practice, we extract at each position s of the studied structure a sequence window Xs of length w ($w = 2\,l + 1$, with $l = 7$, the sequence window is centered on position s). The Bayes' theorem relates the conditional probability $P(PB_k \mid X_s)$ of observing the block PB_k given an amino acid chain X_s (a_{-l}, …,a_s,…, a_{+l}) to the conditional probability $P(X_s \mid PB_k)$ of observing the chain X_s encompassing a particular block PB_k by the following equation:

$$P(PB_k \mid X_s) = \frac{P(X_s \mid PB_k) \times P(PB_k)}{P(X_s)}. \qquad \text{(A1)}$$

$P(X_s)$ and $P(PB_k)$ denote the prior probability of observing a given amino acid chain X_s in a given site s of the studied protein, and the prior probability of observing the block PB_k in the databank respectively. To define the optimal protein block, noted *PB** centered in a given amino acid fragment X_s, a score of adequacy between sequence and *PB*, R_k is computed:

$$R_k = \frac{P(X_s \mid PB_k)}{P(X_s)} = \frac{P(PB_k) \mid X_s}{P(PB_k)}. \qquad \text{(A2)}$$

The ratio R_k is computed for each PB among the 16. The optimal *PB** is defined as the top-scoring PB. To assess the predictions, we compute the accuracy Q_{16}, i.e. the proportion of PBs correctly predicted. This value is equivalent to the Q_3 value for the secondary structures (only 3 states).

Appendix 2. Principle of the pinning strategy

Figure A2 shows an example of pinning prediction. The sequence VIYLNVDNETLSKLV, centred on position s (amino acid N) corresponds to the minimal N_{eq} value of this sequence, i.e. position s*. For this position the maximal $F(s^*, p^*)$ value corresponds to SW *klmmm*, i.e. p*. So, the prefix s*-1 is found as *fklmm*, the SW compatible with p* associated to the maximal F value. In the same way, the prefix s*-1 is as *lmmm*. It leads to a PSOW *fklmmmn*.

Appendix 3. Example of sequence families

The amino acid occurrence matrices associated to SWs are the basis of our prediction methods. In a previous work, we have developed the concept of "*n* sequences for 1-fold", with *n* the number of sequence clusters associated with a given local protein structure, defined as one PB. For each PB, the corresponding set of sequences is split into *n* groups. These groups are optimized with a learning approach related to Self-Organizing Maps (Kohonen 2001). These new amino acid occurrence matrices are called Sequence Families (see §2). This process allows defining more relevant relationship between local protein structure and their associated sequences, ensuring in fact a better prediction. This idea is extended to the SWs, i.e. the most frequent SWs will not be associated to one amino acid occurrence matrix but to *n* amino acid occurrence matrices.

As illustration, we focus on SW *dfklm*, a transition from an extended structure to a helical structure (cf. figure 2f, *rmsd* = 0.71 Å). Figure S3 gives the over-represented amino acids in the sequence window for SW *dfklm* and its two sequence families namely *SF1* and *SF2*. The over-represented amino acids are selected from the occurrence matrix computed from sequence windows of 15 amino acids encompassing this SW. Each SW is characterized by a sequence signature, i.e. a set of amino acids overrepresented in some window positions (i.e. Z-score > +4). The specificity of a position is quantified by the relative entropy index (Kullback-Leibler asymmetric divergence measure, *KLd,* Kullback and Leibler 1951).

The sequence signature of SW *dfklm* is:

$$[(G)_{-5}\,(V)_{-4}\,(PVLI)_{-2}\,(PDST)_{-1}\,(PE)_0\,(DE)_{+1}\,(DEQ)_{+2}(L)_{+3}(LM)_{+6}].$$

The indices are relative to the central position labelled 0, and in parenthesis are indicated the amino acids overrepresented in a given position. Some sites are highly specific (Z-score of an amino acid > 4.0), like positions (-2), (+3), and (+6) exhibiting systematically hydrophobic residues or the positions (0) to (+2) with a charged residue. The informative positions (*KLd* > 1.0) mainly range between (–2) and (+3).

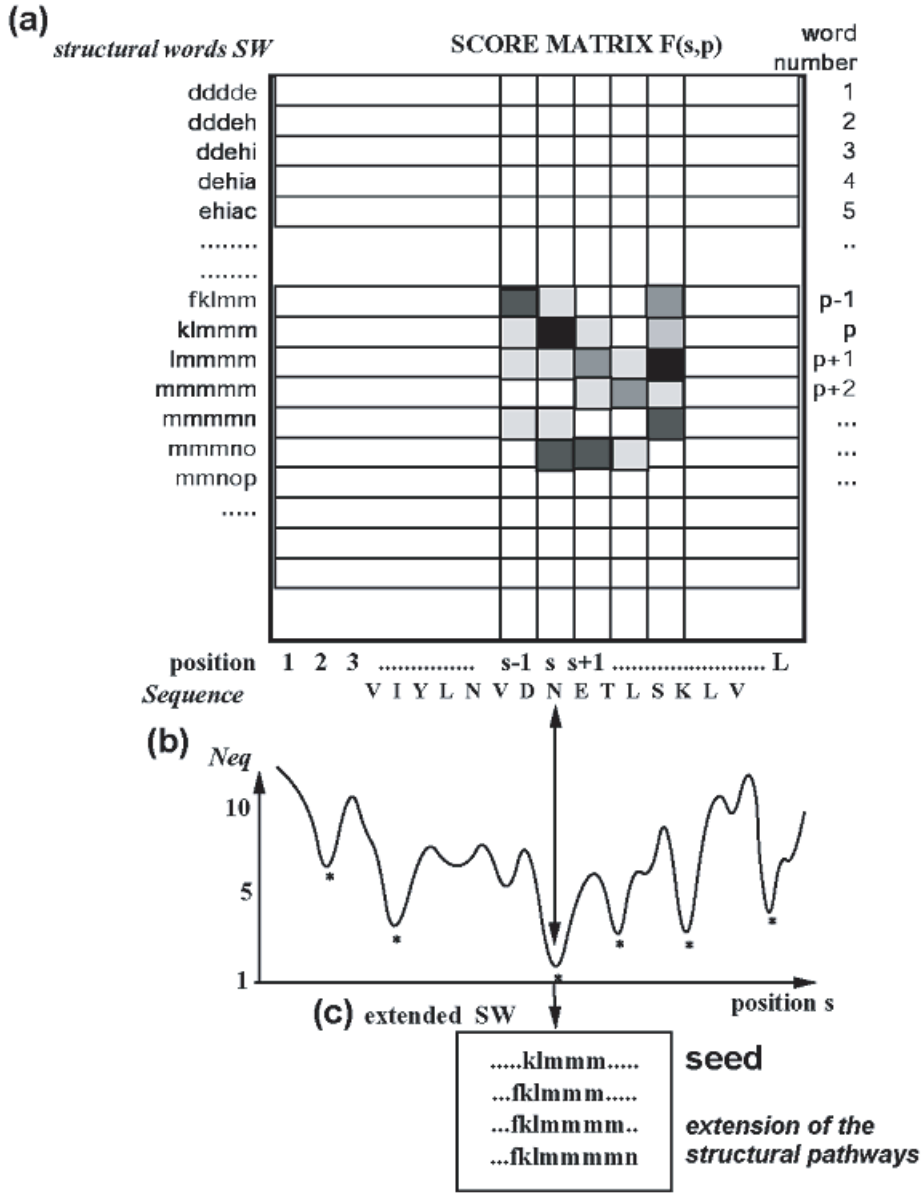


Figure A2. In the upper part is reported the adequacy score matrix in a given position *s* (second column) for the 72 SWs (horizontal lines) for a given sequence window VIYLNVDNETLSKLV. The grey levels correspond to increasing scores. In the bottom part, is indicated the variation of the N_{eq} values in each position along the sequence. The different minima used for the pinning strategy are indicated with a star. The first seed in that example corresponds to the position centred on *s*.

The highest *KLd* value is in position (–1), i.e. the amino acids of this position are particularly important for this SW. In the same way, the sequence signature shows that a proline or a glutamic acid might occupy 3 consecutive positions in the signature of SW *dfklm* (positions -2 to 0, and 0 to +2 respectively).

When the amino acid distribution of SW *dfklm* is examined through the sequence family clustering procedure, the corresponding signatures *SF1* and *SF2* are now respectively

SF1 : $[(P)_{-6}\ (G)_{-5}\ (VI)_{-4}\ (LIYF)_{-2}\ (PDST)_{-1}(PE)_{0}\ (E)_{+1}\ (DE)_{+2}\ (V)_{+3}\ (K)_{+4}\ (E)_{+5}\ (L)_{+6}]$

and *SF2* :$[(G)_{-5}(P)_{-2}(ST)_{-1}(E)_{+1}(DQ)_{+2}(L)_{+3}(L)_{+4}(A)_{+5}\ (LM)_{+7}]$.

Compared to the amino acid signature of SW *dfklm*, we point out the following observations: (i) Overrepresented prolines (Z-score > 4) are now in positions (–1) and (0) for

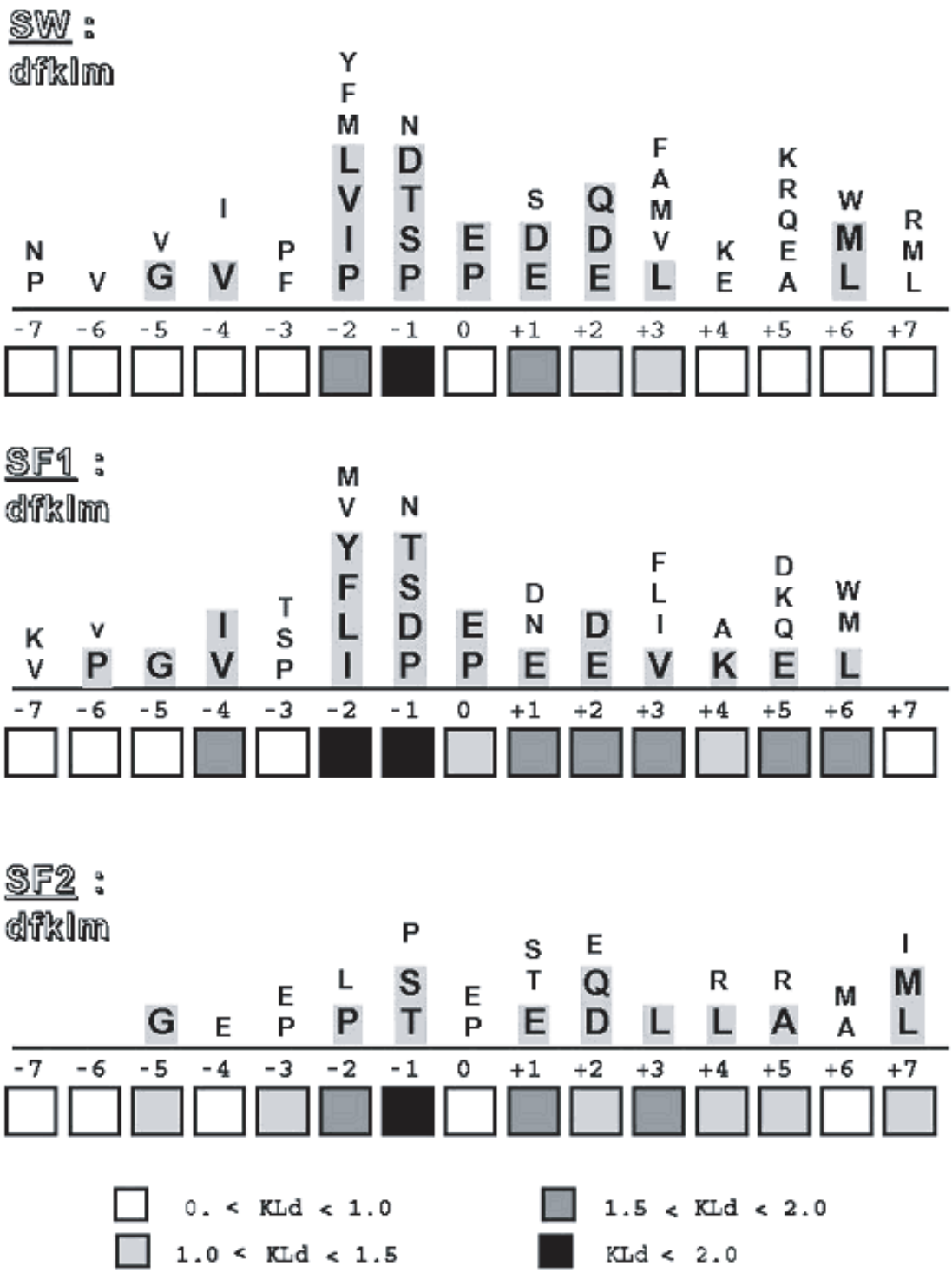


Figure A3. Example of the sequence family concept**.** SW *dfklm* and its two sequence families, SF1 and SF2. Above each position are indicated the over-represented amino acids (Z-scores > 2 and in the grey box Z-scores > 4). Z-scores allow appreciating the relevance of amino acids (see Methods (de Brevern et al. 2000)). Each SW is characterized by a sequence signature, i.e., a set of amino acids over-represented in some window positions. The specificity of a position is quantified by the relative entropy index (Kullback-Leibler asymmetric divergence measure, *KLd,* Kullback and Leibler 1951), measuring the divergence between the observed amino acid frequencies in a given position and that observed in the databank. It allows analysing the information relative to each position. *KLd* is given below each position in four levels.

SF1, whilst it is in position (–2) for SF2. (ii) Overrepresented glutamic acid is also informative; their locations for *SF1* are similar to those of SW before clustering and found also in position (+1) for *SF2*. (iii) Position (–2) becomes entirely hydrophobic in *SF1* (no proline observed), position (–1) remains highly informative in both. Notable differences exist between the two signatures as for instance the locations of hydrophobic sites in position (–4) for *SF1* whilst in position (+7) for *SF2*. These various points show the features specific to the SFs, i.e. informative occurrences can be dispatched into different SFs like the prolines of positions –2 to 0; or not like the glycine of positions –5. In most of the cases, the positions become more informative, i.e. associated to highest *KLd* values. In the same way, some amino acid, not significant in the initial amino acid occurrence matrix, play a significant role for one particular SF, e.g. leucine in positions (+4) of *SF2*.

This example shows clearly the concrete interest of defining sequence families: a same local fold is associated to different types of sequence signatures. Consequently a correct definition of the sequence families should improve the structure prediction from the sequence.

Appendix 4. Visualization of N_{eq} distribution along the protein sequence of 3chy

Figure A4 shows the fold of this protein (Flavodoxin-like) composed of 3 layers $\alpha/\beta/\alpha$ with a parallel β-sheet of 5 strands. The structure is coloured according to the level of the N_{eq}-index (low values mean high information content). We point out that the α-helices globally show lower values than the β-strands. This lower level of sequence information in the β-strand region has already been noticed (de Brevern *et al* 2000) and explains the difference observed in the prediction rate of PB *m* and PB *d* 50.6% and 34.6% respectively. It has to be noted however that high sequence information levels are not systematically associated to α-helices or β-strands but may be found also in non periodic regions described by PBs *g* to *j*. For instance, two α-helices (located in the right part of the figure) present a central zone of weak predictability, i.e. associated with higher N_{eq}-values. In contrast, the loop including residues 59 to 65 (coded *ghiafkl*) and the loop including residues 110 to 113 (coded *dfkl*) exhibit significant predictability. Clearly, the sequence information may be relevant whatever the PB involved. Thus the

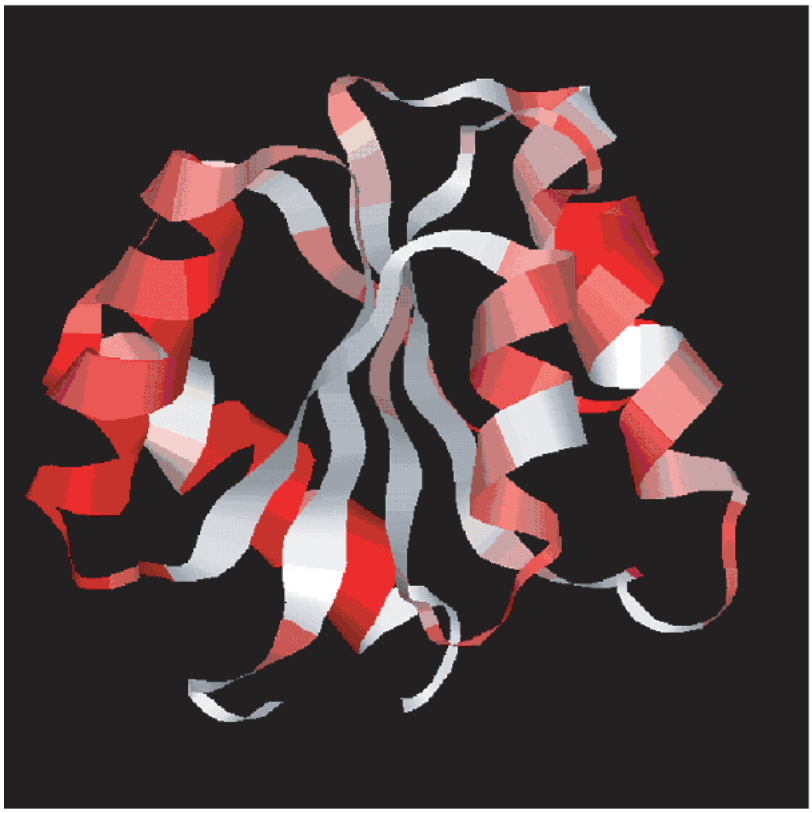


Figure A4. Distribution of the N_{eq}-value along the 3D structure of the signal transduction protein of Escherichia coli (PDB code: 3chy). The colours vary linearly from red for $N_{eq} \leq 5.0$ to white for $N_{eq} \geq 15.0$. Thus, the red parts correspond to low values (less than 5, high sequence information level) and the white parts to high value (more than 15, low sequence information level).

prediction is not limited to the sole periodic repetitive PBs such as *d* or *m* but also to the regions connecting repetitive structures.

Appendix 5. Prediction reliability

A position *s* with a $N_{eq}(s)$ value less than 5 (P_{Neq}^{5}) provides a small number of candidate SWs for the prediction and a maximal score for a SW more than 50 (F_{50}) ensures *a priori* a correct prediction. In the equation, the value 43.5 is an average prediction value obtained with the set of average values (composition, P_{Neq}^{5} and F_{50}). This average value is counterbalanced by the other factors, i.e. the amino acids. All the coefficients involving the amino acids in the equation above are negative while the coefficients involving P_{Neq}^{5} and F_{50} are positive and larger. It means that larger is the proportion of sites with N_{eq} less than 5, higher is its prediction rate. Similarly, larger is the proportion of sites with a score greater than 50, higher is the prediction rate.

By analysing the dependency between each input and the output (i.e. the observed prediction rate), we obtain significant correlations for the most of the used variables for a type-I error of 5%. 10 among the 20 amino acids contribute strongly to the prediction rate. Moreover, the analysis of the significant partial correlation (i.e. a correlation between a given input and the output, the other inputs are considered fixed) shows that no significant partial correlation is observed except for the two last parameters, $P5_{Neq}^{5}$ and F_{50}. This regression exhibits that the main factors implicated in the evaluation on the prediction rate are P_{Neq}^{5} and F_{50}. Thus, we have carried out two other multiple regressions, the first one only based on the relative variations of amino acid frequencies, the second one only based on N_5 and F_{50}. We find significant multiple correlations $R^2 = 0.35$ for both models. The combination of theses information, the amino acid composition of the protein sequence and the adequacy of the sequence with the SW library quantified by the parameters (P_{Neq}^{5}, F_{50}), ensures an improvement of the Q_{16} estimation.

References

Alexandrov N and Shindyalov I 2003 PDP: protein domain parser; *Bioinformatics* **19** 429–430

Alland C, Moreews F, Boens D, Carpentier M, Chiusa S, Lonquety M, Renault N, Wong Y, Cantalloube H, Chomilier J *et al* 2005 RPBS: a web resource for structural bioinformatics; *Nucleic Acids Res.* **33** W44–W49

Altschul S.F, Gish W, Miller W, Myers E W and Lipman D J 1990 Basic local alignment search tool; *J. Mol. Biol.* **215** 403–410

Bairoch A, Boeckmann B, Ferro S and Gasteiger E 2004 Swiss-Prot: juggling between evolution and stability; *Brief Bioinform* **5** 39–55

Benros C, de Brevern A G, Etchebest C and Hazout S 2006 Assessing a novel approach for predicting local 3D protein structures from sequence; *Proteins* **62** 865–880

Benros, C, de Brevern A G and Hazout S 2003 Hybrid Protein Model (HPM): A Method For Building A Library Of Overlapping Local Structural Prototypes. Sensitivity Study And Improvements Of The Training; in *IEEE Workshop on Neural Networks for Signal Processing* (Toulouse, France) pp 53–72

Benros C, de Brevern A G and Hazout S 2004 Predicting Local Structural Candidates from Sequence by the "Hybrid Protein Model" Approach; in *12th Intelligent Systems for Molecular Biology (ISMB) / 3rd the European Conference on Computational Biology (ECCB)*, Glasgow

Bystroff C and Baker D 1998 Prediction of local structure in proteins using a library of sequence-structure motifs; *J. Mol. Biol.* **281** 565–577

Camproux A C, Brevern A G, Hazout S and Tufféry P 2001 Exploring the use of a structural alphabet for structural prediction of protein loops; *Theor. Chem. Acc.* **106** 28–35

Camproux A C, Gautier R and Tuffery P 2004 A hidden markov model derived structural alphabet for proteins; *J. Mol. Biol.* **339** 591–605

Camproux A C, Tuffery P, Buffat L, Andre C, Boisvieux J F and Hazout S 1999a Using short structural building blocks defined by a Hidden Markov Model for analysing patterns between regular secondary structures; *Theor. Chem. Acc.* **101** 33–40

Camproux A C, Tuffery P, Chevrolat J P, Boisvieux J F and Hazout S 1999b Hidden Markov model approach for identifying the modular framework of the protein backbone; *Protein Eng.* **12** 1063–1073

Chan A W, Hutchinson E G, Harris D and Thornton J M 1993 Identification, classification, and analysis of beta-bulges in proteins; *Protein Sci.* **2** 1574–1590

Chivian D, Kim D E, Malmstrom L, Schonbrun J, Rohl C A and Baker D 2005 Prediction of CASP-6 structures using automated Robetta protocols; *Proteins (Suppl. 7)* **61** 157–166

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

Cuff J A and Barton G J 1999 Evaluation and improvement of multiple sequence methods for protein secondary structure prediction; *Proteins* **34** 508–519

de Brevern A G 2005 New assessment of Protein Blocks; *In Silico Biol.* **5** 283–289

de Brevern A G, Benros C, Gautier R, Valadie H, Hazout S and Etchebest C 2004 Local backbone structure prediction of proteins; *In Silico Biol.* **4** 381–386

de Brevern A G, Camproux A-C, Hazout S, Etchebest C and Tuffery P 2001 Protein structural alphabets: beyond the secondary structure description; in *Recent research developments in protein engineering* (ed.) S Sangadai (Trivandrum: Research Signpost) pp 319–331

de Brevern A G, Etchebest C and Hazout S 2000 Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks; *Proteins* **41** 271–287

de Brevern A G and Hazout S 2000 Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties; *IEEE – Comput. Soc.* **S1** 49–54

de Brevern A G and Hazout S 2001 Compacting local protein folds with a "hybrid protein model"; *Theor. Chem. Acc.* **106** 36–47

de Brevern A G and Hazout S 2003 'Hybrid protein model' for optimally defining 3D protein structure fragments; *Bioinformatics* **19** 345–353

de Brevern A G, Valadie H, Hazout S and Etchebest C 2002 Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship; *Protein Sci.* **11** 2871–2886

de Brevern A G, Wong H, Tournamille C, Colin Y, Le Van Kim C and Etchebest C 2005 A structural model of a seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC); *Biochim. Biophys. Acta* **1724** 288–306

Efimov A V 1997 Structural trees for protein superfamilies; *Proteins* **28** 241–260

Eisenberg D 2003 The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins; *Proc. Natl. Acad. Sci. USA* **100** 11207–11210

Errami, M, Geourjon C and Deleage G 2003 Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures; *Bioinformatics* **19** 506–512

Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles F X, Sternberg M J and Oliva B 2004 ArchDB: automated protein loop classification as a tool for structural genomics; *Nucleic Acids Res.* **32** D185–188

Etchebest C, Benros C, Hazout S and de Brevern A G 2005 A structural alphabet for local protein structures: Improved prediction methods; *Proteins* **59** 810–827

Fetrow J S, Palumbo M J and Berg G 1997 Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme; *Proteins* **27** 249–271

Fourrier L, Benros C and de Brevern A G 2004 Use of a structural alphabet for analysis of short loops connecting repetitive structures; *BMC Bioinformatics* **5** 58

Gelly J C, de Brevern A G and Hazout S 2006 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments; *Bioinformatics* **22** 129–133

Geourjon C, Combet C, Blanchet C and Deleage G 2001 Identification of related proteins with weak sequence identity using secondary structure information; *Protein Sci.* **10** 788–797

Girod A, Ried M, Wobus C, Lahm H, Leike K, Kleinschmidt J, Deleage G and Hallek M 1999 Genetic capsid modifications allow efficient re-targeting of adeno-associated virus type 2; *Nat. Med.* **5** 1438

Hartigan, J A and Wong M A 1979 k-means; *Appl. Stat.* **28** 100–115

Henikoff S and Henikoff J G 1992 Amino acid substitution matrices from protein blocks; *Proc. Natl. Acad. Sci. USA* **89** 10915–10919

Humphrey W, Dalke A and Schulten K 1996 VMD: visual molecular dynamics; *J. Mol. Graph.* **14** 33–38, 27–38

Hunter C G and Subramaniam S 2003a Protein fragment clustering and canonical local shapes; *Proteins* **50** 580–588

Hunter C G and Subramaniam S 2003b Protein local structure prediction from sequence; *Proteins* **50** 572–579

Jones D T 1999 Protein secondary structure prediction based on position-specific scoring matrices; *J. Mol. Biol.* **292** 195–202

Jurkowski W, Brylinski M, Konieczny L, Wiiniowski Z and Roterman I 2004 Conformational subspace in simulation of early-stage protein folding; *Proteins* **55** 115–127

Karchin R 2003 *Evaluating local structure alphabets for protein structure prediction*, Ph. D. thesis, University of California, Santa Cruz, USA

Karchin R, Cline M, Mandel-Gutfreund Y and Karplus K 2003 Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry; *Proteins* **51** 504–514

Kohonen T 1982 Self-organized formation of topologically correct feature maps; *Biol. Cybern.* **43** 59–69

Kohonen T 2001 *Self-organizing maps* 3rd edition (Springer) pp 501

Koradi R, Billeter M and Wuthrich K 1996 MOLMOL: a program for display and analysis of macromolecular structures; *J. Mol. Graph.* **14** 29–32

Kuang R, Leslie C S and Yang A S 2004 Protein backbone angle prediction with machine learning approaches; *Bioinformatics* **20** 1612–1621

Kullback S and Leibler R A 1951 On information and sufficiency: *Ann. Math. Stat.* **22** 79–86

Martin J, Letellier G, Marin A, Taly J-F, de Brevern A G and Gibrat J-F 2005 Protein secondary structure assignment revisited: a detailed analysis of different assignment methods; *BMC Struct. Biol.* **5** 17

Milner-White E J 1990 Situations of gamma-turns in proteins. Their relation to alpha-helices, beta-sheets and ligand binding sites; *J. Mol. Biol.* **216** 386–397

Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 SCOP: a structural classification of proteins database for the investigation of sequences and structures; *J. Mol. Biol.* **247** 536–540

Némethy G and Printz M P 1972 The gamma turn, a possible folded conformation of the polypeptide chain. Comparison with the beta turn; *Macromolecules* **5** 755–758

Oliva B, Bates P A, Querol E, Aviles F X and Sternberg M J 1997 An automated classification of the structure of protein loops; *J. Mol. Biol.* **266** 814–830

Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B and Thornton J M 1997 CATH–a hierarchic classification of protein domain structures; *Structure* **5** 1093–1108

Pauling L and Corey R B 1951a Atomic coordinates and structure factors for two helical configurations of polypeptide chains; *Proc. Natl. Acad. Sci. USA* **37** 235–240

Pauling L and Corey R B 1951b The pleated sheet, a new layer configuration of polypeptide chains; *Proc. Natl. Acad. Sci. USA* **37** 251–256

Pei J and Grishin N V 2004 Combining evolutionary and structural information for local protein structure prediction; *Proteins* **56** 782–794

Petersen T N, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert G P and Lund O 2000 Prediction of protein secondary structure at 80% accuracy; *Proteins* **41** 17–20

Pollastri G and McLysaght A 2005 Porter: a new, accurate server for protein secondary structure prediction; *Bioinformatics* **21** 1719–1720

Pollastri G, Przybylski D, Rost B and Baldi P 2002 Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles; *Proteins* **47** 228–235

Prestrelski S J, Williams A L Jr and Liebman M N 1992 Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results; *Proteins* **14** 430–439

Rabiner L R 1989 A tutorial on hidden Markov models and selected application in speech recognition; *Proc. IEEE* **77** 257–286

Richardson J S, Getzoff E D and Richardson D C 1978 The beta bulge: a common small unit of nonrepetitive protein structure; *Proc. Natl. Acad. Sci. USA* **75** 2574–2578

Ring C S, Kneller D G, Langridge R and Cohen F E 1992 Taxonomy and conformational analysis of loops in proteins; *J. Mol. Biol.* **224** 685–699

Rohl C A and Doig A J 1996 Models for the 3(10)-helix/coil, pi-helix/coil, and alpha-helix/3(10)-helix/coil transitions in isolated peptides; *Protein Sci.* **5** 1687–1696

Sander O, Sommer I and Lengauer T 2006 Local protein structure prediction using discriminative models; *BMC Bioinformatics* **7** 14

Sayle R A and Milner-White E J 1995 RASMOL: biomolecular graphics for all; *Trends Biochem. Sci.* **20** 374

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

Shannon C 1948 A mathematical theory of communication; *Bell Syst. Tech. J.* **27** 379–423

Sibanda B L and Thornton J M 1991 Conformation of beta hairpins in protein structures: classification and diversity in homologous structures; *Methods Enzymol.* **202** 59–82

Sowdhamini R and Blundell T L 1995 An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins; *Protein Sci.* **4** 506–520

Tendulkar A V, Joshi A A, Sohoni M A and Wangikar P P 2004 Clustering of protein structural fragments reveals modular building block approach of nature; *J. Mol. Biol.* **338** 611–629

Thompson J D, Higgins D G and Gibson T J 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice; *Nucleic Acids Res.* **22** 4673–4680

Tsai H H, Tsai C J, Ma B and Nussinov R 2004 In silico protein design by combinatorial assembly of protein building blocks; *Protein Sci.* **13** 2753–2765

Tyagi M, Sharma P, Swamy C, Cadet F, Srinivasan N, De Brevern A G and Offmann B 2006 Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet; *Nucleic Acids Res***.** (in press)

Unger R, Harel D, Wherland S and Sussman J L 1989 A 3D building blocks approach to analyzing and predicting structure of proteins; *Proteins* **5** 355–373

Unger R and Sussman J L 1993 The importance of short structural motifs in protein structure analysis; *J. Comput. Aided Mol. Des.* **7** 457–472

Wintjens R T, Rooman M J and Wodak S J 1996 Automatic classification and analysis of alpha alpha-turn motifs in proteins; *J. Mol. Biol.* **255** 235–253

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

ePublication: 6 September 2006