Assessing a novel approach for predicting local 3D protein structures from sequence

Cristina Benros*, Alexandre G. de Brevern, Catherine Etchebest and Serge Hazout

Equipe de Bioinformatique Génomique et Moléculaire (EBGM),
INSERM U726, Université Denis DIDEROT - Paris 7, case 7113,
2, place Jussieu, 75251 Paris, France

* Corresponding author:
mailing address: Benros C., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U726, Université Denis DIDEROT - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
E-mail: benros@ebgm.jussieu.fr
Tel: (33) 1 44 27 99 24
Fax: (33) 1 43 26 38 30

Running title: Local Protein Structure Prediction

key words: library of fragments, sequence-structure relationship, local structure prediction, structural candidates, *ab initio*.

ABSTRACT

We developed a novel approach for predicting local protein structure from sequence. It relies on the Hybrid Protein Model (HPM), an unsupervised clustering method we previously developed. This model learns 3D protein fragments encoded into a structural alphabet of 16 Protein Blocks (PBs). Here, we focused on 11-residue fragments encoded as series of 7 PBs and used HPM to cluster them according to their local similarities. We thus built a library of 120 overlapping prototypes (mean fragments from each cluster), with good 3D local approximation, *i.e.*, a mean accuracy of 1.61 Å Cα *rmsd*.

Our prediction method is intended to optimize the exploitation of the sequence-structure relations deduced from this library of long protein fragments. This was achieved by setting up a system of 120 experts, each defined by logistic regression to optimize the discrimination from sequence of a given prototype relative to the others. For a target sequence window, the experts computed probabilities of sequence-structure compatibility for the prototypes and ranked them, proposing the top scorers as structural candidates. Predictions were defined as successful when a prototype less than 2.5 Å from the true local structure was found among those proposed. Our strategy yielded a prediction rate of 51.2% for an average of 4.2 candidates per sequence window. We also proposed a confidence index to estimate prediction quality.

Our approach predicts from sequence alone and will thus provide valuable information for proteins without structural homologues. Candidates will also contribute to global structure prediction by fragment assembly.

INTRODUCTION

Because of the importance of structural information for the functional characterization of proteins, the prediction of the three-dimensional (3D) structure of proteins from their amino acid sequences is a major scientific challenge. Currently, homology modeling yields the best results.[1,2] However, this approach requires a template protein with a known 3D structure and clear sequence similarity to the protein to be modeled. The task is much more difficult when the target protein does not have obvious homologues in the Protein Data Bank (PDB).[3] Fold recognition methods, which attempt to detect a structural template when sequence similarity is not immediately recognizable, can be an alternative,[4] albeit only partial, because the structural database, while large, is not complete.

Today, large-scale genome sequencing projects are producing numerous protein sequences for which there are no homologues with a known structure. Under these circumstances, any approaches that can provide information about 3D structure from the sequence alone (*ab initio* methods) are of great interest. As the last editions of the Critical Assessment of Methods for Protein Structure Prediction (CASP) show, successful predictions have been made for targets in the new fold category.[5-9] These successes remain limited, however, and some proteins still cannot be predicted.

Methods based on 3D fragment assembly have yielded the greatest progress in this field.[9,10,11] Addressing the problem of local protein structure prediction from sequence may therefore constitute a first step towards global structure prediction.

Considerable work has focused on analyzing the local conformations of available protein structures and trying to predict them from their sequences. The secondary structure provides a three-state description: repetitive α-helices and β-strands, and coils, which are defined as non-helical and non-extended. This description of protein structures is nonetheless crude, despite

the numerous attempts to depict the connecting regions more precisely.[12-16]

Many research groups have designed fragment libraries or structural alphabets to try to describe the local structural features of known protein structures more accurately.[17-29] These libraries or alphabets correspond to finite sets of protein structural fragments. They can be differentiated according to a variety of characteristics, including number of clusters (ranging from 4 to several hundred), fragment length (fixed or variable, from 4 to 9 residues), geometric descriptors (*e.g.*, Cα coordinates, φ, ψ dihedral angles) and clustering methods (*e.g.*, hierarchical clustering, neural networks). These differences depend primarily on the particular goal. The library size, for example, governs the quality of the protein structure description: very generally, a relatively large number of clusters is required for precise protein structure reconstruction,[25,26,27] whereas a small library facilitates the identification of relevant sequence-structure correlations when the goal is local structure prediction from sequence.[22,24,30]

A structural alphabet composed of 16 average protein fragments, 5 residues in length, called Protein Blocks (PBs), was developed in a previous work.[24] These PBs have been used both to describe 3D protein backbones and to predict local structures.[31-34] They have proved to be highly informative and useful for prediction purposes.[35]

The reliability of this structural alphabet for long fragments[31] enabled us to develop an unsupervised clustering method, which we call a Hybrid Protein Model (HPM). This method, which can capture long-range features of a succession of PBs, compresses a structural protein databank into a limited set of clusters.[36,37,38]

The HPM training principle is similar to that of Kohonen's Self-Organizing Maps (SOM).[39,40] Its originality is that it can learn long protein fragments previously encoded into series of PBs. HPM compacts a databank of such fragments into one "Hybrid Protein" (HP), by stacking them on the basis of the similarity of their PB series. Through this process, it

builds a library of clusters that group structurally similar fragments; each cluster is represented by a mean local structure prototype. Unlike standard clustering methods, HPM generates a library of overlapping prototypes. Its principal advantage is that it takes into account the dependence between successive local structures along the proteins by maintaining their continuity.

Two main characteristics affect the features of the final library built by HPM: the length of the protein fragments and the number of clusters in the library. A first HPM of 100 clusters, grouping a series of 10-PB fragments, was used for fine description of protein 3D structures and efficiently identified local structural similarities between two cytochromes P450.[36] Subsequent examination of a new learning approach led to an HPM of 233 clusters that grouped a series of 13-PB fragments.[37] Recent work has focused on improving detection of similarities between long fragments.[38]

The principal concept of HPM can be compared with that of HMMSTR.[41] As HPM extends the PB structural alphabet, HMMSTR extends the I-sites library.[22] They both enable description of the continuities of different sequence-structure motifs observed in proteins and they both represent overlapping motifs in a compact form. Nevertheless, they differ substantially. HPM is linear whereas HMMSTR has a branched topology. HMMSTR is trained simultaneously on sequence and structure databases, while HPM is trained only on structural data. Significant amino acid properties are then deduced from the clusters.

Our aim in this paper is to use HPM descriptions and features to predict local 3D protein structures from their sequences. For this purpose, we built a new HPM, modifying the number of clusters and the fragment length required. We also took advantage of the increased structural information provided by the PDB.[3] Next, we used the library of local structure prototypes constructed by the HPM to develop a prediction strategy, aiming at optimizing exploitation of the sequence-structure relations in this library. This was achieved by setting up

a system of experts, each defined by logistic regression and best able to discriminate from sequence a given local structure prototype relative to the others. The experts then computed probabilities for each prototype for a target sequence window, and the top scorers become structural candidates. The results were analyzed with different evaluation schemes, and a confidence index was proposed to assess prediction quality. The structural prototype candidates can contribute to *ab initio* tertiary structure prediction as structural constraints or as fragments in combinatorial assembly.

MATERIALS AND METHODS

The structural alphabet

The structural alphabet used in this work corresponds to a set of 16 short prototypes, each 5 residues in length, called Protein Blocks (PBs;[24] see supplementary data 1). Each PB is defined by 8 dihedral angles (ϕ and ψ). Very generally, PB *m* represents central α-helices and PB *d* central β-strands. PBs *a* through *c* primarily describe β-strand N-caps, and PBs *e* and *f*, β-strand C-caps. PBs *g* through *j* are specific to coils. Finally, PBs *k* and *l* and PBs *n* through *p* represent α-helix N- and C-caps, respectively. This structural alphabet approximates local 3D protein structures with a mean accuracy of 0.41 Å Cα *rmsd* (root mean square distance between α-carbon atoms).[34]

Structure databanks

Protein databanks. A non-redundant set of 675 protein structures was selected from the PDB-REPRDB database[42] according to the following criteria: X-ray structures with 2 Å or better resolutions, no more than 30% pairwise sequence identity, a Cα *rmsd* value larger than 10 Å between all representative chains. An updated databank, composed of 1,041 proteins,

was built from the same criteria.

Each protein structure was encoded into the structural alphabet according to the following coding principle: every overlapping fragment of 5 consecutive residues of a given protein was assigned to the PB most similar according to its dihedral angles.[24] The similarity criterion used is the *rmsda* (root mean square deviation on angular values).[20] Hence, each 3D protein structure was encoded by a string of characters, or a series of consecutive overlapping PBs.

Fragment databanks. Each protein structure, encoded in terms of PBs, was cut into overlapping fragments of L consecutive PBs. The positions for a given fragment are labeled from $-w$ to $+w$, and the index 0 is assigned to the central PB. Hence, the fragment length is defined as $L = 2w+1$, with w equal to 3: the fragment length in our study is 7 PBs. Since each PB is a 5-residue fragment and the successive PBs are overlapping, each fragment of L PBs has a corresponding amino acid fragment of length $L+4$, similarly labeled from $-(w+2)$ to $+(w+2)$, that is, 11-residue fragments. The first databank was composed of 139,503 fragments and the updated one of 251,497 fragments.

The first databank was used to train the Hybrid Protein Model. The training was carried out with three quarters of this databank (105,340 fragments), and the remaining quarter (34,163 fragments) was used to evaluate the Hybrid Protein stability. The second databank was divided into three subsets and used for prediction. Half the databank (the learning set) was used to compute the amino acid occurrence matrices (521 proteins; 125,074 fragments). One quarter (the parameterization set) was used to establish the expert scoring functions (261 proteins; 62,194 fragments). Finally, the remaining quarter was used as a validation set to assess the prediction method (259 proteins; 64,229 fragments).

Construction of the structural prototype library

Hybrid Protein Model. The structural prototype library is built by an unsupervised classifier named the Hybrid Protein Model (HPM), introduced in previous work.[36,37,38] Its main principles are summarized in supplementary data 2. Briefly, the Hybrid Protein (HP) corresponds to a self-organizing neural network. It is characterized by a ring of *N* neurons, each representing a cluster of structurally similar 3D fragments. Its training strategy consists of learning the protein structure fragments encoded into series of *L* PBs, by stacking them according to their similarities. Similar but not identical series of PBs may be grouped in the same cluster. The result is equivalent to a multiple structural alignment of the local protein structure fragments. Stacking allows the building of a "PB profile", that is, a series of PB probability distributions. Each region of this PB profile is *L* successive PB distributions in length and characterizes a cluster of structurally similar fragments, represented by a mean local structure prototype. The successive regions are overlapping and have (*L*-1) probability distributions in common, thereby defining overlapping local structure prototypes.

Definition of the structural prototypes. At the end of HPM training, each neuron or cluster, noted *s* (*s* varying from 1 to *N*), was associated with a set of 3D protein fragments (*L*+4) = 11 Cα long. For each cluster, we computed the pairwise Cα *rmsd* values between all the fragments and we selected a representative structural prototype, $\mathbf{P}_s$, which corresponds to the fragment with the smallest sum of Cα *rmsd* values of all the fragments in the cluster. We next ensured that each fragment was assigned to the closest prototype, that is, that with the smallest Cα *rmsd* value.

Characterization of the relation between amino acid sequence and structural prototype

Relative self-information matrices. To analyze the sequence-structure relations, we

considered enlarged sequence windows [-*m*; +*m*] of length $M = 2m + 1$, with $m = (w + 7)$. These correspond to a final window of $M = 21$ residues, to take account of the amino acid information content in the neighbourhood. We computed for each cluster an amino acid occurrence matrix of dimensions M x 20. This matrix was then translated into a "relative self-information matrix". The "relative self-information content" $i_s(y,AA_k)$ of a given amino acid AA_k ($k = 1, \ldots, 20$) in position y of the relative self-information matrix (y varying from $-m$ to $+m$) for a given local structure prototype $\mathbf{P}_s$ ($s = 1, \ldots, N$) is equal to:

$$i_S(y, AA_k) = \ln\left[\frac{p_S(y, AA_k)}{p_R(AA_k)}\right] \quad (1)$$

$p_s(y,AA_k)$ corresponds to the probability of observing amino acid AA_k in position y of the sequence windows of the fragments associated to the local structure prototype $\mathbf{P}_s$. $p_R(AA_k)$ is the reference frequency of the amino acid AA_k in the databank. The amino acid relative self-information content is positive (respectively, negative) when the amino acid is overrepresented (respectively underrepresented).

Analysis of the amino acid information content. To analyze the amino acid information content of each position y of the sequence windows associated with a given local structure prototype $\mathbf{P}_s$, we computed the relative entropy, defined by the asymmetric Kullback-Leibler discrepancy, $KLd_s(y)$.[43] It corresponds to the mathematical expectation of the relative self-information content of all the amino acids in position y:

$$KLd_S(y) = \sum_{k=1}^{k=20} p_S(y, AA_k) \ln\left[\frac{p_S(y, AA_k)}{p_R(AA_k)}\right] \quad (2)$$

This index measures the contrast between the amino acid frequencies observed in position y for structural prototype $\mathbf{P}_s$ and their reference frequencies. The significance of the $KLd_s(y)$ value is assessed by a χ^2 test, since the quantity $(2.NF_s.KLd_s(y))$, with NF_s the number of

fragments (or sequence windows) associated with the local structure prototype $\mathbf{P}_s$, follows a χ^2 distribution with 19 degrees of freedom.

Strategy for predicting local protein structure from sequence

The prediction strategy we propose is based on an expert system. One expert was defined for the optimal discrimination from amino acid sequences of a given local structure prototype relative to the others; there were N experts, one for each of the N local structure prototypes of the library. An overview of the prediction strategy is shown in Figure 1. For a target sequence window W of unknown local 3D structure, each expert gives its diagnosis, that is, the probability that the sequence window fits the structural prototype it characterizes. A jury then selects from among these diagnoses the structural prototype candidates for a given decision rule. The jury can also assess the candidate list by a confidence index.

Characterization of the expert system. We determined each expert by logistic regression (R software[44]). The logistic function computes the probability that a given sequence window W belongs to the 3D fragment cluster represented by the local structure prototype under study, $\mathbf{P}_s$. We used the parameterization set, from which two learning subsets of 3D fragments were defined for this logistic regression. The first subset, called the "positive set", is composed of fragments assigned to the prototype $\mathbf{P}_s$ cluster; the second or the "negative" set, is composed of an equivalent number of 3D fragments taken randomly from the other clusters. For each fragment of these two sets, we derived its "self-information input vector" from its corresponding amino acid sequence window W. The relative self-information content of the M amino acids of the sequence window W, defined from the relative self-information matrix of $\mathbf{P}_s$, constitutes the elements of this input vector. Hence, the y^{th} element of this input vector is the quantity $i_s(y,AA_k)$, AA_k denoting the amino acid in position y in the

sequence window *W*.

The logistic function or score $p(W / \mathbf{P}_s)$ quantifies the probability of the compatibility or fit between the sequence window *W* and the local structure prototype $\mathbf{P}_s$. It is expressed as:

$$p(W / \mathbf{P}_S) = \frac{1}{1 + \exp\left[-\phi_S(W)\right]} \tag{3}$$

$$\text{with} \qquad \phi_s(W) = w_0 + \sum_{y=-m}^{y=+m} w_y \; i_s(y, AA_k) \tag{4}$$

This logistic regression is well suited for optimally discriminating between the negative and positive fragment sets. The weights w_y allow it to assess the contribution to each expert's discriminatory power of the amino acids located in the different positions *y* of $\mathbf{P}_s$. The significance of each weight is assessed by a *t* test.[44]

To assess the relevance of each logistic regression, we determined the minimal error risk, denoted R_{min}. It corresponds to the minimal average fraction of false positives (that is, sequence windows of the negative set classified as positive) and false negatives (sequence windows of the positive set classified negative) obtained for a probability threshold p_{Rmin}.

Jury and decision rule. For a target sequence window *W*, each expert assesses the sequence-prototype fit by computing the probability $p(W / \mathbf{P}_s)$. A jury selects the best structural prototypes as candidates for the local 3D structure from among the *N* probabilities, ranked in descending order. The decision rule for selecting prototype candidates has two criteria: (i) $p(W / \mathbf{P}_s) > p_0$, with p_0 a user-set threshold for sequence-structure fit, and (ii) a fixed maximum number of candidates.

Evaluation of the candidates. The prediction strategy was assessed by applying it to the validation set proteins. Two evaluation schemes were tested. In the first a prediction for a

target sequence window was defined as successful when the prototype assigned from the structure was found among the candidates proposed from the sequence. The second used a geometric evaluation: a prediction was defined as successful based on its Cα *rmsd* from the true local structure. Different degrees of approximations were considered, that is, 1.5 Å, 2 Å, and 2.5 Å. For reference purposes, the distribution of the Cα *rmsd* of pairs of unrelated 11-residue fragments selected randomly from the databank has a mean of 4.5 Å (standard deviation, *sd* = 1.1 Å, see supplementary data 3). To assess the quality of our prediction strategy, we also compared the rates we obtained to rates computed from lists with the same number of local structure prototypes randomly drawn from the library.

Confidence index for assessing the selected prototype candidates. A confidence index *CI* was defined to assess the list of prototype candidates. Its aim was to quantify the probability of finding a successful candidate among those proposed. Here, successful is defined as having a Cα *rmsd* less than 2.5 Å from the true local structure. We began by dividing the fragments of the training set into two subsets: the first contained fragments for which at least one successful candidate was found among the proposed local structure prototypes, and the other grouped fragments with no successful candidate. A new logistic regression used an input vector composed of one *a priori* and three *a posteriori* data items. (i) *A priori information related to the target amino acid sequence window: amino acid classes.* We subdivided the 20 types of amino acids into four classes. The first two correspond to G and P, the third to the hydrophobic amino acids (I, V, L, M, A, F, Y, W, C), and the fourth one to polar amino acids (H, S, T, N, Q, D, E, R, K). The target sequence window was then encoded according to these classes. (ii) *Information about the selected prototypes*. We included a Boolean vector indicating for each of the *N* local structure prototypes whether it was selected as a candidate by the jury (1 if selected, elsewhere 0). (iii) *Information about the*

proportions of observed probabilities of sequence-prototype compatibility. The next elements of the input vector correspond to the proportions of probabilities within the ranges [0.9; 1.0], [0.8; 0.9], [0.7; 0.8], [0.6; 0.7] and [0.5; 0.6] among the N prototypes computed. (iv) *Information about the levels of the observed probabilities*. Finally, we added the differences between some probability values, according to their rank: $(1 - p_1)$, $(p_2 - p_1)$, $(p_5 - p_2)$, $(p_{10} - p_5)$ and $(p_{60} - p_{10})$, with p_k denoting the kth probability.

From the input vectors derived for all sequence windows of the training set, we estimated the optimal weights of the logistic function (see equation 3) to discriminate between the lists with and without a successful local structure prototype candidate. Next, we divided the distribution of the probabilities p given by the logistic function into 6 levels of confidence ranging from $CI = 1$ (low confidence) to $CI = 6$ (high confidence), based on the variation of the error risk. We assessed the relevance of the confidence index by computing the prediction rates according to the 6 levels of confidence for the validation set.

RESULTS

We built a library of representative protein structural prototypes to use for predicting protein local structures. We first examine the structural features of this library and then analyze the relation between amino acid sequences and local structure prototypes. We then assess the prediction strategy. Our approach is based on the use of an expert system able to discriminate local structure prototypes from amino acid sequences. An interesting point of this methodology is that it proposes not one but several structural prototype candidates. Analysis focuses on the expert system and the assessment of the proposed structural prototype candidates.

Construction of a library of 120 structural prototypes

The best library of structural prototypes, *i.e.*, the best Hybrid Protein (HP), was obtained after testing different numbers of clusters. We began testing with large numbers of clusters and selected the final HP as the best balance between many clusters and enough fragments per cluster to permit prediction. We checked the low structural variability of each cluster (see below). Figure 2 shows the final HP and its main features. It corresponds to a library of 120 clusters of 11-residue protein fragments (7 PBs). Each cluster shares similar local structures and is associated with a representative local structure prototype. The successive prototypes overlap.

A structural profile characterizes the prototype library. The Hybrid Protein matrix of PB distributions (see Figure 2a) is equivalent to a "structural profile" obtained by stacking the fragments encoded as PBs. The profile reveals interesting features. First, high structure specificity is observed along the HP since only a few PBs characterize each site. There are long series of identical PBs that correspond to repetitive structures. They mainly involve two specific PBs, *d* and *m*. In previous studies,[24,34] we have shown the strong correspondence between PB *m* and the central part of the α-helix, and to a lesser extent between PB *d* and the central part of the β-strand. Thus, the stretches of PB *m* correspond to helical structures (see for example prototype #24 in Figure 2d, visualized with VMD software[45]), while those of PB *d* globally characterize extended structures (see prototype #10 in Figure 2d). Three series of PB *m* are distinguishable, located in HP regions [#21 - #29], [#39 - #45] and [#67 - #71]. Six series of PB *d* are observed. Certain are highly specific to PB *d*, see, *e.g.* [#8 - #13], while others are associated with PBs related to extended structures, such as PBs *b* and *c*, see, *e.g.*, [#76 - #83]. The lengths of these series vary, depending on their location. They also differ by their local environment, that is, by the nature of the neighboring prototypes.

Figure 2b shows the number of protein fragments associated with the different clusters; it ranges from 308 to 4028. The first helical region [#21 - #29] is the most heavily populated, and the most fragments are associated with cluster #24. The fragments are more evenly distributed for the extended structures.

Substantial diversity of the local structures is observed for the non-repetitive HP regions that connect the repetitive regions. Interestingly, structure specificity is high, with clear PB signatures. For instance, the HP region [#27 - #33] corresponds to the exit of a helical structure defined by a well determined PB series, that is, *mmmnopa* (*e.g.*, see prototype #30 in Figure 2d). These kinds of particular signatures are highly recurrent in protein structures, as seen by their frequencies in Figure 2b. Other HP connecting regions are characterized by more PBs per site, *e.g.* the region [#85 - #91], which connects two extended structures.

Low geometric variability within each cluster. We want to stress that the clusters are defined by grouping the protein fragments according to their similarity in terms of a PB series. Thus, similar but not identical series of PBs may be grouped in the same cluster. To assess the quality of the 3D approximation, we computed ***a posteriori*** for each cluster its average Cα *rmsd* value (see Figure 2c), which we obtained by superimposing all the 11-Cα fragments of a cluster with their representative structural prototype. The local structure prototype of a cluster corresponds to the fragment closest in terms of Cα *rmsd* values to all the other fragments in the cluster. This prototype may thus be described both by its PB series and by its Cα coordinates. The average Cα *rmsd* value of 1.61 Å (standard deviation, *sd* = 0.77 Å) indicates that the geometric variability of the prototypes is low. This local structure approximation is quite satisfactory for fragments 11 residues long, since the random value for this length residue is 4.5 Å (*sd* = 1.1 Å, see supplementary data 3). The Cα *rmsd* values

computed for each cluster vary within the range [0.28 Å - 2.44 Å]. Cluster #26, located in the main repetitive helical region, has the lowest Cα *rmsd* value, and cluster #113, in a region connecting two extended structures, has the highest (see supplementary data 4 for details).

Prototype comparisons. HPM uses the structural dependence between successive local structures along the proteins to build this library of overlapping prototypes. The originality of this method therefore is to ensure the continuity between successive local structure prototypes. This property of continuity necessarily leads to a certain structural redundancy, that is, some Hybrid Protein regions may be structurally close. These similar prototypes, however, are differentiated by their transitions -- the prototypes that precede and follow them.

We investigated the extent of this structural redundancy by computing the Cα *rmsd* values obtained by the optimal pairwise superimposition of all 120 local structure prototypes (see Figure 3). This analysis showed that the structural redundancy was actually fairly weak. The Cα *rmsd* value for most of the prototype pairs was greater than 2.5 Å (displayed in blue). The mean Cα *rmsd* value was equal to 3.9 Å (*sd* = 1.0 Å). Of the total of 7140 prototype pairs, only 0.4% (31 pairs), 0.3% (21 pairs) and 1.8% (124 pairs), respectively, corresponded to prototypes less than 1 Å (in red), from 1 Å to 1.5 Å (in orange), and from 1.5 Å to 2 Å (in yellow). As expected, these pairs mainly involved prototypes located in repetitive HP regions. For instance, all the pairs within the helical region [#23 - #27] are less than 1 Å apart (see the red square in the bottom left corner). This figure also shows structurally similar prototypes belonging to distant HP helical regions: prototypes #22, #23 and #24 are less than 1 Å from prototypes #41 and #42. Similarities are also observed between extended regions, such as regions [#8 - #12] and [#96 - #100], and between non-repetitive regions, such as the helical capping regions [#20 - #22] and [#66 - #68].

Relations between amino acid sequence and structural prototype

After we built the library of structural prototypes, we sought to assess the specificity of the amino acid sequences associated with each cluster. We therefore computed an amino acid-related self-information matrix for each cluster (see Methods section). We showed in an earlier study[24] that prediction can be improved by enlarging the sequence window to five residues on each side. Hence, we considered enlarged sequence windows, 21 residues in length (noted from -10 to +10 and centred at 0) to take account of the amino acid content in the neighbourhood. The elements of the self-information matrices, *i.e.*, self-information content, provide information for each amino acid at each position. Positive (respectively negative) self-information content corresponds to overrepresented (respectively underrepresented) amino acids. The sequence information content of the different amino acid self-information matrices is measured by the asymmetric Kullback-Leibler discrepancy (*KLd*).[43] This index highlights the most informative positions.

All the clusters of the library show significant amino acid sequence specificity. Using the *KLd* index, we determined that the number of significantly informative ($\alpha < 0.001$; see Methods section) positions per self-information matrix ranged from 4 to 20 of 21, with a mean of 12.6. These positions are located mainly in the central region (from -5 to +5). The average number of informative positions among the 11 central positions equals 9.4. The most informative matrix is associated with cluster #27, which characterizes the exit from a helical structure. The least informative matrices are located in HP region [#50 - #53], where there are a relatively large number of PBs.

Figure 4 illustrates the representative local structure prototypes of clusters #67 and #15 (Figure 4a), their relative self-information matrices (Figure 4b) and their *KLd* distributions (Figure 4c). Cluster #67, corresponding to the N-cap of a helical structure, is highly

determined in both structure (mean Cα *rmsd* = 1.06 Å) and sequence (the *KLd* index highlights 14 informative positions). We observe high over- and underrepresentations (displayed respectively in red and blue, see caption of Figure 4 for details) of different amino acids, mainly in the central positions. These over- and underrepresentations are numerous at each position (vertical lines) and along the positions (horizontal lines). For instance, (P, G, S, T, D) are overrepresented in position -3. Several other overrepresentations, such as (Q, D, E, K) and (A, L, M), are also observed in other positions. There are also many underrepresentations, principally in position -3 (vertical blue line) where 11 of the 20 amino acids, mainly hydrophobic, are underrepresented. This is the most informative position of the matrix, with a *KLd* value equal to 0.43 (see Figure 4c). These amino acid preferences are consistent with those normally observed in helical structures and in helix-capping motifs.[46] Cluster #15, which corresponds to the exit from an extended structure, is slightly less structurally determined (average Cα *rmsd* = 1.96 Å) and less sequence-dependent (8 informative positions defined by the *KLd* index) compared with cluster #67. Among the amino acid preferences we observe overrepresentations of I and V in positions -4 to -2 and of P and D in positions [-1, 0]. These preferences are also consistent with other descriptions.[47]

Now that we have shown strong sequence-structure correlations in our library, our goal is to exploit this information for prediction purposes. Our previous prediction methods were based solely on the use of amino acid occurrence matrices and Bayes' rule.[24,33,34] To go further in exploiting the relations between sequence and structure, we developed an improved prediction strategy that relies on an expert system. The experts' principal task is the discrimination of a particular local structure prototype relative to the others on the basis of the amino acid sequence.

Analysis of the expert system

For each cluster of the library, we determined an expert able to distinguish between amino acid sequence windows that belong to the cluster, *i.e.* that adopt its local structure, and those that do not belong to the cluster (that is, the positive and negative subsets; see Methods section). The expert of each cluster was defined by a logistic function. It computes the probability of sequence-structure compatibility between a given target sequence window and its local structure prototype. Each logistic function was characterized by 21 optimized weights, one per position of the self-information matrix (see equations 3 and 4). Figure 4d, for example, shows for clusters #67 and #15 graphic indicators of the significance level of the weights assigned to the different positions of their self-information matrices (see caption for details). Positions with three stars contribute highly to the expert's discriminating power. In most cases, the positions with statistically significant weights are those for which the *KLd* values indicate strong amino acid specificity, *e.g.*, positions [-4; -3] and [-1; 1] for cluster #67 and, positions [-2, 0, 1] for cluster #15. A few positions not considered informative by the *KLd* index are nonetheless significantly related to the expert's discriminatory power and are thus assigned high weights, *e.g.*, position -5 compared to -2 and 3 of cluster #67. We also note that the enlargement of the sequence window captured additional information, as we can see from the stars assigned to positions located outside the central region, *e.g.*, positions -7, 6, 7 and 10 of cluster #15.

Assessment of the experts' discriminating power. To assess the discriminating power of each expert, we analyzed the distributions of the probabilities computed for the positive and negative subsets. As expected, the shapes of these distributions show most probabilities are close to 0 for the fragments that do not belong to the cluster (negative subset), and most are close to 1 for the fragments that do (positive subset; see supplementary data 5 which

depicts the distributions for cluster #67). The quality of discrimination for each cluster is assessed by analyzing the minimal average fraction of false-positive and false-negative fragments (defined as the minimal error risk, R_{min}, see Methods section). Overall, we found an average R_{min} equal to 25% (sd = 3.8%), corresponding to a p_{Rmin} about 0.5. This means that an expert correctly discriminated the right cluster for 75% of the sequence windows, on average. R_{min} values for the different clusters range from 16% (cluster #7, good structure determination and strong sequence specificity) to 35.3% (cluster #52, only 8 informative positions but structurally well determined).

Thus, we set up a system of experts and assessed the power of each to discriminate a given local structure prototype of the library from sequence, relative to the others. The paragraphs below focus on the assessment of our new prediction method.

Local structure prediction from sequence

A classic strategy would propose only the local structure prototype for which the associated expert returned the highest probability, that is, each sequence window would be associated with one predicted prototype. Due to the size of our library (120 prototypes), however, several prototypes have high probabilities. We thus chose a strategy that selects a -- limited -- series of candidates, *i.e.*, each sequence window is associated with one or a few predicted local structure prototypes. In practice, our strategy relies on the jury's decision rule, which ranks the prototypes associated with a high probability of sequence-structure compatibility (larger than a given threshold p_0) into a list of structural prototype candidates for a target sequence window.

The prediction was tested on a validation set of 259 proteins for each position along the target protein sequences. The lists of local structure prototype candidates were evaluated

under two schemes. The first, based on prototypes, retrieved the prototype assigned from structure among those predicted from sequence. This scheme is very stringent for a library the size of ours -- 120 clusters. A second evaluation scheme, based on Cα *rmsd* to the true local structure, was thus considered: it looked for at least one prototype candidate in the list structurally close to the true structure. It provided a more realistic and appropriate evaluation of the prediction's validity.

Optimal probability threshold for sequence-structure compatibility. For each expert, we set a probability threshold, noted p_0 (see above): above this threshold, the sequence window was considered associated with the prototype, otherwise not. To minimize the proportion of false-positive sequence windows, a threshold value p_0 substantially greater than p_{Rmin} was chosen. This choice is made to the detriment of the true positives. After a systematic search, we set p_0 to 0.8. This value was applied to all the experts, as it ensures similar proportions of false positives, with a mean of 5% (see example of cluster #67 in supplementary data 5).

For a given target sequence window, several experts may return a compatibility probability more than the threshold, thereby producing a list of local structure prototype candidates. The jury ranks them according to their probability values. The target sequence windows for which no candidate was proposed accounted for only 1.6% of the validation set. For these sequence windows, there were no prototypes with a probability above the threshold (p_0 = 0.8). When we exclude these windows, the number of prototype candidates per target sequence window averaged 6.3 (*sd* = 3.8). Less than 2.9% of the sequence windows had more than 15 prototype candidates among 120.

Evaluating the proposed lists of prototype candidates. The first evaluation scheme,

based on the assigned prototype, yielded a prediction rate of 35.0%; that is, experts were able to recognize on average 35.0% of the fragments belonging to their cluster solely from sequence information. This rate is clearly satisfactory given the size of the library. Moreover, it is significantly greater than the 5.1% rate obtained by randomly drawing the same number of prototype candidates from the library for each target sequence window. In addition, the top-scoring prototype candidate corresponded to the assigned structure with a frequency of 11.5%. This rate is also clearly better than the random rate (0.8%) for such a large library.

Geometric evaluation. The second evaluation scheme is based on a threshold Cα *rmsd* distance from the true local structure. This geometric evaluation is more appropriate since it takes into account the structural similarity of certain local structures to more than one prototype. Hence, it takes advantage of the structural proximity of some prototypes (see paragraph *Prototype comparisons*). The assigned prototype provides the best local approximation in terms of Cα *rmsd*, but other prototypes may also provide good 3D approximations, albeit not the best.

Three Cα *rmsd* thresholds - 1.5 Å, 2 Å and 2.5 Å – were used to quantify the prediction rates. They have not been used to define the clusters but are related to the geometrical characteristics deduced ***a posteriori*** from each cluster (see Figure 2c). The threshold of 1.5 Å is very stringent for fragments as long as 11 Cα, while 2.5 Å is close to the mean Cα *rmsd* value (2.44 Å) for the most variable cluster of the library (see paragraph *Low geometric variability within each cluster*). Even this threshold is stringent in comparison with the random value of 4.5 Å (*sd* = 1.1 Å), and the probability (or *p*-value) for a random match with Cα *rmsd* < 2.5 Å is 10^{-2} (see supplementary data 3).

Table I shows the prediction rates and the improvements over random rates when setting a maximum number of allowed candidates (noted *MNAC*). Our strategy produces a

variable number of candidates per sequence window. Accordingly, the mean number of candidates is always less than the *MNAC* value, as some sequence windows have fewer candidates proposed. For instance, for *MNAC* = 7, the mean number of candidates is 5.1. The prediction rate when considering only the top-scoring candidate (*MNAC* = 1) equals 29.5% for a Cα *rmsd* threshold of 2.5 Å. This value is clearly satisfactory given that the predictions are based on sequence information alone and with a large library of prototypes. As expected, the prediction rate decreases with stricter thresholds: 20.6% for 2 Å and 13.6% for 1.5 Å. It is nonetheless significant in such a context. When we considered the candidate lists, with a mean of 6.3 candidates, the prediction rate reached 55.4% for a threshold of 2.5 Å (38.3% for 2 Å and 23.7% for 1.5 Å).

To assess our prediction strategy, we compared these rates with prediction rates from lists of the same numbers of prototypes randomly drawn from the library. The prediction rate increased with the number of proposed candidates, but so did the random prediction rate. The difference between these rates is thus a good indicator of prediction quality. Our method showed clear improvements over random selection for all *MNAC* values and all thresholds. More importantly, the margin between the prediction strategy and random choice was highest when there were only a few prototype candidates per sequence window. Thus, for a *MNAC* = 5 (corresponding to an average value of 4.2 candidates), the prediction rate was 51.2% for a threshold of 2.5 Å (significantly better than the random rate, with a gain of 29.3%), 35.1% for 2 Å (improvement over random of 24.3%) and 22.2% for 1.5 Å (improvement of 17.0%). In view of these satisfactory results, the maximum number of candidates per sequence window was set at 5. We discuss the resulting lists in detail below.

Analysis of the prediction with four categories of prototypes. To provide a simplified evaluation of the prediction results, we defined four categories of prototypes, relative to the

secondary structures. They were obtained by hierarchical clustering of the prototypes from the matrix of Cα *rmsd* shown in Figure 3 (see supplementary data 6 for details). These rough categories are defined as helical, extended (core), edges (inputs and outputs of extended structures) and connecting structures. These categories include 16, 13, 40 and 51 prototypes, respectively. The proportion of fragments from the validation set in each category is respectively 26.9%, 10.6%, 23.8% and 38.7%. Proportions are similar for the training set.

Table II summarizes the contribution of the different categories of prototypes in the prediction results. The prediction rate was high for the helical prototypes -- 70.1% for a Cα *rmsd* threshold of 2.5 Å. The rates were also quite satisfactory for the strictest accuracy levels, *e.g.*, 53.1% for a 1.5 Å-Cα *rmsd* threshold. Improvements over random selections range from 48.3% to 53.0%. This category contains slightly fewer candidates, an average of 3.9. The prediction rate was also high for the prototypes corresponding to the core of relatively large extended structures -- 52.6% for a Cα *rmsd* threshold of 2.5 Å, 30.7% better than the random case. The third category (named edges) actually unites shorter extended structures and edges. Predictions rates were also satisfactory at a 2.5 Å threshold for this category (43.3%) and for the last, which groups prototypes corresponding to connecting structures (42.4%). The least stringent Cα *rmsd* threshold must be considered here since these categories are the most variable and their flexibility makes them hard to predict. Improvement over random prediction remains significant at approximately 20% at 2.5 Å.

Analyzing the prediction results for individual prototypes (see supplementary data 7), we found that each of the 120 prototypes contributed to the global predictions rates. At the Cα *rmsd* threshold of 2.5 Å, the individual prediction rates ranged from 18.6% for cluster #6 to 76.9% for cluster #26. The prediction rate exceeds 40% for 84 prototypes. For a stricter accuracy level, *e.g.* 2.0 Å, high prediction rates were obtained for prototypes belonging to the different categories even the most variable ones, *e.g.* 60.4% for prototype #67. In addition, we

observed that the prediction rate of a local structure prototype was related to its cluster mean Cα *rmsd* value (see supplementary data 8).

Six confidence levels assess the lists of prototype candidates. We defined six confidence levels for these predictions, from *CI* = 1 (low confidence) to *CI* = 6 (high confidence). They were obtained by a logistic regression intended to provide optimal discrimination between the lists that do and do not contain a successful prototype candidate at the Cα *rmsd* threshold of 2.5 Å. The input vectors included various items of information, about the amino acid target sequence window, the selected prototype candidates and their probabilities (see Methods section). These were selected as those that best discriminated between lists with and without successful candidates. An optimal weight was estimated for each, and its contribution to the discriminatory power of the logistic function assessed by a *t* test.

Table III summarizes the prediction results for the validation set, according to the six confidence levels. Results were similar for the training set. The logistic regression was conducted for a threshold of 2.5 Å but the results were also analyzed for 1.5 Å and 2 Å. The results show that the confidence levels defined are good indicators of the probability of finding a successful candidate among those proposed. 27.2%, 11.4% and 8.6% of the target sequence windows were assigned to *CI* = 4, *CI* = 5 and *CI* = 6, respectively, the three highest levels of confidence; prediction rates for those levels reached 60.8%, 75.5% and 85.8%, respectively for a threshold of 2.5 Å. The confidence index is also relevant for thresholds of 1.5 Å and 2 Å. We also analyzed the confidence levels for predictions for the four prototype categories defined above (see supplementary data 9). The level of confidence was highest for the helical structures (63.2% were assigned to the three highest confidence levels, with 22.9% in *CI* = 6) but was also high for the other categories (assignments to the three highest

confidence levels accounted for 54.3% of the core of extended structures, 41.2% of the edges and 37.7% of the connecting structures).

Examples of prototype candidates. Figure 5 shows examples of the prototype candidates proposed for *Escherichia coli* signal transduction protein CheY (PDB code 3CHY).[48] It is an alpha and beta protein, 128 residues in length. The prediction rate for this protein reached 63.9% for the Cα *rmsd* threshold of 2.5 Å (50.0% for 2 Å and 36.1% for 1.5 Å), and 67.6% of the protein fragments were assigned to the three highest prediction confidence levels, with 18.5% in *CI* = 6. Four examples of predictions are displayed in Figure 5, with the true local structure (in red), the assigned prototype (in green), and the selected prototype candidates (in blue).

The example labeled (a) corresponds to a helical structure with its exit. The assigned prototype, #28, provides an accurate 3D local approximation (Cα *rmsd* = 0.50 Å). The prediction results show that the latter is the top-scoring candidate, with a high probability of sequence-structure fit (p = 0.99). Two additional candidates are proposed: #69 and #41, which have lower compatibility probabilities (p = 0.86 and p = 0.81, respectively) and provide less accurate approximations of the true local structure, at distances of 2.49 Å and 1.91 Å. These local structures are nonetheless at least partially satisfactory, especially the second candidate, for the orientation of its exit. The confidence level assigned to this prediction is the highest, *CI* = 6.

Example (b) corresponds to the input of an extended structure. Five candidates with extended structures are proposed, and the top scorer is the assigned prototype, #59. It approximates the true local structure, with a Cα *rmsd* value of 1.02 Å. The second candidate also provides an acceptable local approximation (2.20 Å), but the three last candidates have Cα *rmsd* values greater than 2.5 Å, the least stringent threshold. The confidence level for this

prediction is 5.

The two last examples, (c) and (d), correspond to loops. In the global structure, they connect an α-helix to a β-strand, and a β-strand to an α-helix, respectively. The confidence levels for these predictions are low -- 2 and 3, respectively. Only one candidate is proposed for example (c), and its sequence-structure compatibility probability is relatively low (p = 0.86). It does, however, provide an acceptable approximation (2.12 Å). This result is especially satisfactory since the assigned prototype gives an approximation of 2.09 Å. This also emphasizes the fact that prototypes proposed as candidates but different from the assigned one can provide good 3D approximation. Finally, an unsuccessful prediction result is shown in example (d). Here, the assigned prototype is 3.05 Å-Cα *rmsd* from the true local structure. Thus, the prediction is necessarily unsuccessful despite a high probability of sequence-structure compatibility for the candidate.

DISCUSSION AND CONCLUSION

This paper is a first attempt to exploit the features of the Hybrid Protein Model (HPM), an unsupervised clustering method introduced in previous work[36,37,38] for predicting local 3D protein structures from amino acid sequences. Here, HPM is used to construct a library of 120 clusters grouping series of 7-Protein Block (PB) fragments, 11 residues in length. The prediction method we propose takes advantage of the sequence-structure relation deduced from HPM for long fragments, but also aims at optimizing the discrimination between the clusters on the basis of sequence.

A first point that requires discussion concerns the choice of the fragment length and of the number of clusters (or local structure prototypes) in the library. The parameters chosen

ensure a good balance between accurate description of 3D local structures and tractable prediction objectives. The fragment length of 7 PBs is shorter than we used in previous studies because we have found it to be clearly more appropriate and realistic for prediction than 10 or 13 PBs.[36,37] It corresponds to a length of 11 residues. A major advantage of using long fragments is that they capture long-range correlations, but as fragment length increases, the completeness of the fragment databases becomes a limitation. Nonetheless, the number of protein structures in the PDB has increased enormously in the last few years, and the observation is that the number of new-found folds is decreasing.[6] In addition, in a recent work, Du *et al.* (2003)[49] argued that there currently exists sufficient coverage to model even a novel fold using fragments from the PDB.

The principal characteristic of HPM is that it learns protein 3D fragments previously encoded into series of PBs. Two important points must be considered here: (i) the 16 PBs have proved reliable for long fragments, and (ii) the transitions between successive PBs are highly specific and lead to a limited number of PB combinations. A study focusing on the extension of the local backbone description identified the 72 most frequent series of five PBs, called Structural Words (SWs).[31] These provide good structural approximations for fragments that are already fairly long, *i.e.*, nine residues long and cover 92% of the amino acids. HPM is a probabilistic approach that goes further than the SW analysis. It enables us to capture the longer-range features of successions of PBs and condenses these long local structures according to the features they have in common. It also takes account of the structural dependencies observed in proteins. Unlike the SWs, HPM groups together similar but not always identical PB series and makes it possible to define a PB profile.

Using 120 prototypes ensures a good quality of 3D local approximation for 11-residue fragments, with a Cα *rmsd* of 1.61 Å between the fragments and their closest prototype. The main geometric features of 11-residue protein fragments have thus been captured. As the

number of local structure prototypes in the library is relatively large, these last may contribute to define and describe the boundaries of the secondary structures more accurately. It must be noticed that all the protein fragments are classified here; we have not removed outliers, contrary to other studies.[*e.g.*,26] In addition, The 120 clusters are sufficiently populated to identify relevant sequence-structure dependencies.

Many studies have investigated the classification of protein fragments, but only few for the purpose of local structure prediction.[*e.g.*,22,27,30] A direct comparison with our library is difficult because of major differences in characteristics, such as fragment length and number of clusters. Hunter and Subramaniam (2003)[30], for example, used a 28-cluster basis set of seven-residue prototypes, obtained from a hypercosine clustering method; it models the backbone structure with a mean accuracy of 1.23 Å Cα *rmsd*.

The prediction approach proposed here focuses on the ability to discriminate local structure prototypes from sequence. Our previous prediction methods used amino acid occurrence matrices and Bayes' rule,[24,33,34] as other works have.[30] Here, we developed an improved prediction strategy relying on a system of experts to optimize the exploitation of sequence-structure relations.

The prediction rate was 51.2%, a rate 29.3 percentage points higher than random prediction, for an average of 4.2 candidates per sequence window and a Cα *rmsd* threshold of 2.5 Å. The threshold of 2.5 Å is strict in comparison with the random value of 4.5 Å for 11-residue fragments. Beyond this threshold, the difference between the actual and random prediction falls, to 25.3% at 3 Å. This point further supports the validity of the threshold of 2.5 Å. Moreover, it must be noted that Yang and Wang (2003)[27] used a threshold of 2.4 Å to evaluate the percentage of nine-residue segments correctly predicted by their methodology.

The use of a system of experts, each characterized by a logistic function, has thus

proved effective. In addition, the expert system ensures that each prototype contributes to the global prediction rate, with satisfactory individual prediction rates. A major advantage of logistic regression is that the expert's discriminatory power relies in part on an assessment of the contribution of the amino acid occurrence matrix positions.

Our prediction strategy attempts to identify for each target sequence window a limited series of structural prototype candidates. The lists may contain prototypes that are close to one another structurally, and in such cases some could be reduced. For instance, we tested the effect of removing prototype candidates located less than 1.5 Å from higher-scoring ones. We then analyzed these lists, limiting the maximum number of allowed candidates to 5. Interestingly, the mean number of candidates decreased from 4.2 to 3.9, while the prediction rate rose slightly, from 51.2% to 51.5%, and the difference between HPM and random prediction from 29.3% to 31.3%. This reduction is dependent on the Cα *rmsd* threshold used, and its effect must be further analyzed, especially on the compatibility between successive prototypes.

Until now, few local structure prediction methods have been published. For easier and more relevant comparison purpose, we selected the study carried out by Yang and Wang (2003)[27] that includes itself comparisons with other works. They described a local structure prediction method which goals to predict the backbone conformation of nine-residue sequence segments using four states: A, B, G and E (see the Ramachandran plot in Figure 1 of Yang and Wang, 2003)[27]. We thus encoded the local structure candidates proposed with our strategy in terms of these four conformational states (A, B, G, E), and we computed a consensus prediction from the multiple alignment of these candidates. The originality of our prediction strategy is to propose a limited series of local structure prototype candidates for a query sequence window, associated to a confidence index. When considering only the top scoring candidate (*MNAC* = 1) or when considering the best candidate (the closest to the true local

structure) among an average of 4.2 candidates per sequence window (*MNAC* = 5) to compute the consensus, the backbone torsion angle prediction accuracy ranged from 63.6% to 75.8%. This result is comparable to the prediction accuracy obtained by Yang and Wang (2003)[27] which equals 74.7% when their method is combined with PSI-PRED (otherwise this value decreases of approximately 7%). This result is also comparable to HMMSTR[41] prediction accuracy (74.0%). With more sophisticated methods based on the use of support vector machines (SVMs) combined with PSI-BLAST[50] and PSI-PRED[51], Kuang *et al.* (2004)[52] reached a prediction accuracy of 77.3% (more details are given in supplementary data 10). Our method is based on logistic functions that are equivalent to perceptrons without a hidden layer. Improvements of our prediction results may be obtained with more sophisticated artificial neural networks or with SVMs. In addition, we note that our approach is specifically different from the others in that it makes prediction from the amino acid sequence alone (no sequence alignments). The evaluation of the prediction results must therefore be viewed in this context.

Yang and Wang (2003)[27] also used a Cα *rmsd* prediction accuracy measure similar to that of Bystroff and Baker (1998)[22]. It corresponds to the proportion of test residues for which at least one of the overlapping nine-residue segments is predicted correctly, that is, less than 1.4 Å from the true local structure. We computed the same criteria by considering only the nine central residues of our eleven-residue fragments. The proportion of correctly predicted residues we obtained ranged from 53.5% when considering only the top-scoring candidate (*MNAC* = 1) to 72.2% when considering the best candidate among those proposed (*MNAC* = 5). The rate of 62.1% of correctly predicted residues obtained by Yang and Wang (2003)[27] is comprised in the range defined by our values. Then, we did the same calculations with our eleven-residue fragments and Cα *rmsd* thresholds. The percentage of correctly predicted residues when considering only the top-scoring candidate ranged from 44.5% (Cα *rmsd* < 1.5

Å) to 62.7% (Cα *rmsd* < 2.0 Å). When considering the best candidate among those proposed, this percentage respectively ranged from 59.7% (Cα *rmsd* < 1.5 Å) to 80.1% (Cα *rmsd* < 2.0 Å) (more details are given in supplementary data 11). These comparisons further indicate the prediction capacities of our method.

Our approach could provide valuable information for proteins without structural homologues. Even in homology modeling, such approach would be useful, particularly in the regions of the target protein where the conformation may differ from the template, as in loop regions.[53] Our prediction results may nonetheless be improved by use of homology-derived sequence information, when available. Using multiple sequence alignments has, for example, permitted great improvements in secondary structure predictions.[54] Recently, Pei and Grishin (2004)[55] proposed a nearest-neighbor method for local structure prediction; it combines evolutionary and structural information. They showed that the combination of this information improved the accuracy of the applications of their methodology to secondary structure prediction (*i.e.,* 3 states).

Another major point of this paper is the definition of a confidence index. The results obtained showed that it is a good indicator of the probability of finding an acceptable candidate among those proposed.

Future work will focus on global structure prediction by fragment assembly. This kind of approach lies on the idea that proteins are constructed from a small catalog of recognizable parts that fit together in a limited number of ways.[56] We will use the properties of continuity and overlap of successive prototypes. In addition, the long length of the prototypes is well suited to this objective.[26] Possible structural pathways could be defined from the lists of prototype candidates, assessed by the confidence index. Because all the transitions between successive prototypes are not possible, combinatorial assembly will be limited.

Acknowledgments

The authors thank P. Fuchs, A.C. Camproux and D. Flatters for fruitful discussion, and F. Guyon and P. Tufféry for the superimposition software. The authors also thank the reviewers for their very useful comments. This work was supported by 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.”). CB has a grant from the Ministère de la Recherche. AdB is a researcher at the French Institute for Health and Medical Research (INSERM). CE and SH are Professors at the University Paris 7 - Denis-Diderot, Paris.

Captions

Figure 1: *Overview of the local prediction strategy.* (a) Each cluster *s* of the library, *s* varying from 1 to *N*, is represented by an average local structure prototype $\mathbf{P}_s$. (b) For each cluster, an amino acid relative self-information matrix of dimensions (20x21), *i.e.*, the 20 amino acids and a sequence window 21-residues long (labeled from -10 to +10), is computed. (c) For a target sequence window *W*, 21-residues long and of unknown structure, (d) a vector of relative self-information contents is derived for each of the *N* clusters of the library. (e) The resulting vectors are the inputs of the *N* experts, and each expert is characterized by a logistic function. (f) The output of each expert corresponds to the probability that the target sequence window is compatible with the local structure prototype of the cluster it characterizes, *i.e.*, $p(W / \mathbf{P}_s)$. They are analyzed by a jury that, based on a decision rule, does (when black) or

does not (when white) propose the structural prototype of the cluster as a candidate (g). Two prototype candidates are proposed in the example.

Figure 2: *Library of 120 structural prototypes built by the Hybrid Protein Model (HPM)*. (a) The library built by HPM has a length of N =120 clusters and is characterized by a PB profile, that is, a series of PB probability distributions. The gray level indicates the PB frequencies, which vary from 0 (white) to 1 (black). A cluster is defined by L = 7 PB distributions. Successive clusters are overlapping and have 6 PB distributions in common. Examples of some clusters are displayed: #10, #24, #30, #64 and #105. HPM is a closed linear neural network. Hence, cluster #120 is contiguous to cluster #1. (b) Distribution of the protein fragments. (c) Mean Cα *rmsd* value of each cluster (Å). (d) Example of structural prototypes of the library, displayed using the VMD software.[45] The N-cap is on the left and the C-cap is on the right. Each prototype is 11 Cα long and is characterized by a series of 7 PBs. For each prototype, the mean Cα *rmsd* value of its cluster and its standard deviation are also stated. The prototypes displayed correspond to an extended structure (#10), a helical structure (#24) and three capping structures (#30, #64 and #105). Prototype #30 overlaps prototype #24 over 5 Cα.

Figure 3: *Cα rmsd values computed after optimal pairwise superimposition of the 120 prototypes*. Colors are assigned as follows: (red): Cα *rmsd* < 1 Å; (orange): 1 Å ≤ Cα *rmsd* < 1.5 Å; (yellow): 1.5 Å ≤ Cα *rmsd* < 2 Å; (green): 2 Å ≤ Cα *rmsd* < 2.5 Å; and (blue): Cα *rmsd* > 2.5 Å.

Figure 4: *Analysis of the sequence – local structure relation*. The sequence-structure relation is illustrated for two clusters: #67 and #15. (a) The representative prototype is

displayed, and its corresponding series of 7 PBs is stated, with the mean Cα *rmsd* value of the cluster and the standard deviation. (b) The relative self-information matrix of dimensions (20x21), noted from -10 to +10 and centred in 0, is displayed. The color assignment is based on the distribution of the self-information content values of all the 120 matrices: top 5% of values (red), next 15% (orange), lowest 5% of values (blue), next 15% (green), otherwise (yellow). The amino acids from bottom to top are: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K. (c) *KLd* distribution. The significance threshold equals 0.02 for cluster #67, and 0.04 for cluster #15 ($\alpha < 0.001$; see Methods section). (d) The expert logistic scoring function assigns a weight to each position of the relative self-information matrix. A *t* test assesses their significance. Graphic indicators of the level of significance are assigned according to the *p*-value, as follows: '***': $0 \leq p < 0.001$, '**': $0.001 \leq p < 0.01$, '*': $0.01 \leq p < 0.05$ and '.': $0.05 \leq p < 0.1$.

Figure 5: *Examples of prototype candidates proposed for Escherichia coli signal transduction protein CheY (PDB code 3CHY)*. Four examples are displayed. The true local structure with its location in the protein is shown in red. The assigned prototype with the corresponding Cα *rmsd* of local approximation is green. The selected prototype candidates are displayed in blue and ranked according to the probability *p* of their sequence-structure compatibility; and the Cα *rmsd* of local approximation it provides is reported for each.

References

1. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Francisco M, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000 ;**29**:291-325.

2. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;**294**:93-96.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucl Acids Res* 2000;**28**:235-242.
4. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. *Proteins* 2001;**S5**:55-67.
5. Lesk AM, Lo Conte L, Hubbard TJ. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins* 2001; **S5**:98-118.
6. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins* 2003;**53**:436-456.
7. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. TOUCHSTONE: A unified approach to protein structure prediction. *Proteins* 2003;**53**:469-479.
8. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. *Proteins* 2003;**53**:480-485.
9. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* 2003;**53**:457-468.
10. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 1997;**268**:209-225.
11. Bradley Holmes J., Tsai J. Some fundamental aspects of building protein structures from fragment libraries. *Protein Science* 2004;**13**:1636-1650.

12. Donate LE, Rufino SD, Canard LHJ, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modelling and prediction. *Protein Sci* 1996;**5**:2600-2616.
13. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. *J Mol Biol* 1997;**266**:814-830.
14. Wojcik J, Mornon JP, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol* 1999;**289**:1469-1490.
15. Michalsky E, Goede A, Preissner R. Loops In Proteins (LIP) – a comprehensive loop database for homology modelling. *Protein Eng* 2003;**16**:979-985.
16. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJE, Oliva B. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res* 2004;**32**:185-188.
17. Unger R, Harel D, Wherland S, Sussman L. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 1989;**5**:355-373.
18. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol* 1990;**213**:327-336.
19. Prestrelski SJ, Williams Jr. AL, 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.
20. 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.
21. 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.

22. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motif. *J Mol Biol* 1998;**281**:565-77.
23. Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng* 1999;**12**:1063-1073.
24. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* 2000;**41**:271-287.
25. 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.
26. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol* 2002;**323**:297-307.
27. Yang AS, Wang L. Local structure prediction with local structure-based sequence profiles. *Bioinformatics* 2003;**19**:1267-74.
28. Hunter CG, Subramaniam S. Protein fragment clustering and canonical local shapes. *Proteins* 2003; **50**:580-588.
29. Camproux AC, Gautier R, Tufféry P. A Hidden Markov Model derived structural alphabet for proteins. *J Mol Biol* 2004;**339**:591-605.
30. Hunter CG, Subramaniam S. Protein local structure prediction from sequence. *Proteins* 2003;**50**:572-579.
31. 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.
32. de Brevern AG, Benros C, Gautier R, Valadié H, Hazout S, Etchebest C. Local

backbone structure prediction of proteins. *In Silico Biol* 2004;**4**:31.

33. Fourrier L, Benros C, de Brevern AG. Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics* 2004;**5**:58.
34. Etchebest C, Benros C, Hazout S, de Brevern AG. A structural alphabet for local protein structures: improved prediction methods. *Proteins* 2005;**59**:810-827.
35. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 2003;**51**:504-514.
36. de Brevern AG, Hazout S. Compacting local protein folds with a Hybrid Protein Model. *Theor Chem Acc* 2001;**106**(1/2):36-47.
37. de Brevern AG, Hazout S. Hybrid Protein Model for optimally defining 3D protein structure fragments. *Bioinformatics* 2003;**19**:345-353.
38. 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.
39. Kohonen T. Self-organizing formation of topologically correct feature maps. *Biol Cybernet* 1982;**43**:59-69.
40. Kohonen T. Self-Organizing Maps, 3rd edition. 2001;Springer-Verlag, Berlin, Germany.
41. 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.
42. Noguchi T, Matsuda H, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the protein data bank (PDB). *Nucl Acids Res* 2001;**29**:219-220.
43. Kullback S, Leibler RA. On information and sufficiency. *Ann Math Stat* 1951;**22**:79-86.

44. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comp Graph Stat* 1996;**5**:299-314.
45. Humphrey W, Dalke A, Schulten K. VMD - Visual Molecular Dynamics. *J Mol Graph* 1996;**14**:33-38.
46. Aurora R, Rose GD. Helix capping. *Protein Sci* 1998;**7**:21-38.
47. Liu WM, Chou KC. Singular points of protein β-sheets. *Prot Sci* 1998;**7**:2324-2330.
48. Volz K., Matsumura P. Crystal structure of Escherichia coli Che Y refined at 1.7-A resolution. *J Biol Chem* 1991;**266**(23):15511-9.
49. Du P, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng* 2003;**16**:407-414.
50. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389-3402.
51. Jones DT. Protein secondary structure prediction based on position-specific scoring matrix. *J Mol Biol* 1999;**292**:195-202.
52. Kuang R, Leslie CS, Yang AS. Protein backbone angle prediction with machine learning approaches. *Bioinformatics* 2004;**20**:1612-1621.
53. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. *Proteins* 2003;**53**:352-368.
54. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 2002;**47**:228-235.
55. Pei J, Grishin NV. Combining evolutionary and structural information for local protein structure prediction. *Proteins* 2004;**56**:782-794.
56. Fitzkee NC, Fleming PJ, Gong H, Panasik N Jr, Street TO, Rose GD. Are proteins

made from a limited parts list? *Trends in Biochem Sci* 2005;**30**:73-80.