BIOINFORMATICS

Vol. 19 no. 3 2003, pages 345–353
DOI: 10.1093/bioinformatics/btf859

'Hybrid Protein Model' for optimally defining 3D protein structure fragments

A. G. de Brevern and S. Hazout*

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

Received on February 1, 2002; revised on June 6, 2002; accepted on September 9, 2002

ABSTRACT

Motivation: Our aim is to develop a process that automatically defines a repertory of contiguous 3D protein structure fragments and can be used in homology modeling. We present here improvements to the method we introduced previously: the 'hybrid protein model' (de Brevern and Hazout, *Theor. Chem. Acc.*, **106**, 36–47, 2001) The hybrid protein learns a non-redundant databank encoded in a structural alphabet composed of 16 Protein Blocks (PBs; de Brevern *et al.*, *Proteins*, **41**, 271–287, 2000). Every local fold is learned by looking for the most similar pattern present in the hybrid protein and modifying it slightly. Finally each position corresponds to a cluster of similar 3D local folds.

Results: In this paper, we describe improvements to our method for building an optimal hybrid protein: (i) 'baby training,' which is defined as the introduction of large structure fragments and the progressive reduction in the size of training fragments; and (ii) the deletion of the redundant parts of the hybrid protein. This repertory of contiguous 3D protein structure fragments should be a useful tool for molecular modeling

Contact: debrevern@urbb.jussieu.fr

INTRODUCTION

Although it has been suggested that protein structures may adopt only a limited number of folds (Govindarajan *et al.*, 1999), the determination of their 3D structure from their sequence remains a difficult task (Baker and Sali, 2001). The conventional methods, such as homology modeling (Fiser and Sali, 2002) and threading (Kelley *et al.*, 2000) take advantage of the substantial growth of the Protein DataBank (Berman *et al.*, 2000). At the same time, the '*ab initio* modeling' strategy is still restricted to small proteins (Bonneau and Baker, 2001). At the last CASP4 workshop, *ab initio* modeling with some structural constraints showed very encouraging results in some complicated protein structure predictions (Bonneau *et al.*, 2001).

*To whom correspondence should be addressed.

Protein prediction is based first and foremost on an accurate knowledge of the available protein structures. These structures may be studied at different levels: (i) local, on the basis of its classical secondary structures (3-state coding) or of a 'structural alphabet' (n-state coding, $n > 3$); (ii) the protein domains; and (iii) the clusters of protein folds.

At the simplest level, the 3D structure description is often limited to sequences of secondary structures (α-helix, β-sheet and coil). Defining this simple 3-state alphabet is not an easy task. Accordingly, different algorithms have been developed for this; they use various criteria, such as dihedral angles (ϕ, ψ) distribution, energetic values, Cα distances, consensus or crystallographic approaches (Colloc'h *et al.*, 1993; Frishman and Argos, 1995; Labesse *et al.*, 1997; King and Johnson, 1999). The α-helix and β-sheet repetitive structures represent less than 50% of all protein folds. Attempts to classify coils have not yielded completely satisfactory results, due to their large conformational variability (Ring *et al.*, 1992)

To overcome these limitations, libraries of small prototypes have been built to describe protein structures in their entirety. They are based on different types of data (backbone description in terms of Cα, dihedral angles, or other angles), fragment lengths (from 4 to 9) or numbers (from 4 to 100) (for a review, see de Brevern *et al.*, 2001). We have chosen to use an alphabet composed of 16 Protein Blocks (PBs), each 5 Cα in length; it approximates protein 3D-structures with adequate accuracy and has been used in a Bayesian prediction of protein structures from their sequences (de Brevern *et al.*, 2000).

At the intermediate level, a protein is described as a set of protein domains, with the definition of domains dependent on the criteria used. Jones *et al.* (1998) have shown that the three classic algorithms—PUU (Holm and Sander, 1994), DOMAK (Siddiqui and Barton, 1995) and DETECTIVE (Swindells, 1995)—assign only 72% of proteins to the same cluster. Wernisch *et al.* (1999) and Taylor (1999) have developed new more precise definitions of domains. This type of research is essential to the understanding of protein folding and interactions (Jones *et al.*, 2000).

A higher level of protein description classifies them into different families, i.e. the proteins are described by their secondary structures. The most popular classifications are SCOP, which provides a detailed description of the structural and evolutionary relations between all proteins with a known structure (Murzin *et al.*, 1995), FSSP, based on exhaustive all-against-all 3D structure comparisons (Holm and Sander, 1996), and CATH, a hierarchical domain classification of protein structures (Orengo *et al.*, 1997). These classifications are used to find distant structural homologues (Bray *et al.*, 2000) or for structural genomics purpose (Pearl *et al.*, 2000). Systematic comparison of protein structure classifications of these three databases shows that they classify similarly only two thirds of all proteins (Hadley and Jones, 1999).

Describing and classifying protein structures are thus not easy tasks. The 'hybrid protein model' (HPM) attempts to tackle some of these issues. Because proteins have common local structures of various lengths, we tried to stack the structures locally. HPM is an unsupervised classifier, similar to Kohonen's self-organizing maps (SOM; Kohonen, 1982). These unsupervised methods are widely used in proteomics. For example, SOM can assess the composition of protein secondary structures from circular dichroism experiments (Unneberg *et al.*, 2001) or from the clustering of protein sequences into families (Ferrán *et al.*, 1994; Andrade *et al.*, 1997). Developing those approaches makes it possible to create an associative database of protein sequences, accessible via the internet (Hanke *et al.*, 1999). They can also be used to search for protein cleavage sites (Schneider *et al.*, 1998), map enzyme sites (Stahl *et al.*, 2000), or determine secreted proteins (Schneider, 1999).

In our approach, HPM builds a concatenation of local structures that share common parts (de Brevern and Hazout, 2001). After training a non-redundant protein databank, every local structure of every protein is located in a given position of the hybrid protein. Its principal interest is that it defines contiguous fold clusters. The improvements that we introduce to the HPM approach here affect two criteria: (i) continuity between consecutive hybrid positions; and (ii) redundancy within the hybrid protein. The strategies to accomplish this involve: (a) 'baby training,' by which large structural fragments are introduced, with their size progressively reduced during the training; and (b) the deletion of the redundant parts of the hybrid protein.

Hence, the hybrid protein helps us understand both its structures and its amino acid sequences. The quality of this optimal hybrid protein has been assessed by evaluating the 3D local folds at each site of the hybrid protein. Some of these are detailed here, and the informativity of the sequence is analyzed. In the last section, we will point out the various potential uses of this repertory of protein structure fragments for structure prediction.

MATERIALS AND METHODS

Databank of 3D protein structures encoded into protein Blocks

In a previous paper (de Brevern *et al.*, 2000), we established a structural alphabet for coding 3D protein structures; it is a set of 16 local prototypes, called Protein Blocks (PBs), that can approximate the protein backbone locally. The average *root mean square deviation* (*rmsd*) of the PBs is 0.58 Å. The 16 PBs are labeled by letters from *a* to *p*. PB *m* is the prototype of the central α-helix and *d* the prototype of the central β-sheet. PBs *a* to *c* primarily represent β-sheet N-caps and *e* and *f*, C-caps; *g* to *j* are specific to coils, *k* and *l* to the N-caps, and *n* to *p* to the C-caps of α-helices. This categorization provides a rough approximation of the PB locations in the protein folds.

Figure 1 shows the coding of a lysozyme (code PDB: 153l): every fragment of 5 consecutive residues is assigned to the corresponding Protein Block according to its series of dihedral angles. Thus a protein M amino acids long is translated into a string vector of $(M - 4)$ PBs. The interest of using a structural alphabet lies in its conversion of a 3D object (i.e. the protein backbone) into a string of characters (i.e. the associated PB series). This alphabet has also been used to predict local protein structure by a Bayesian approach (de Brevern *et al.*, 2000).

The databank used in our study is composed of 717 non-redundant proteins taken from the Protein DataBank (Berman *et al.*, 2000). Specifically, we selected from the PAPIA databank (Noguchi *et al.*, 2001) the chains with a resolution of 2 Å or less and an R-factor less than 0.2. Each structure selected had an *rmsd* value at least 10 Å larger than all the other structures selected and a sequence identity of 30% or less. The entire 3D protein structure of each protein selected was encoded into PBs. Hence, the databank is composed of 177 986 PBs.

'Hybrid protein model' (HPM)

In another previous paper (de Brevern and Hazout, 2001), we developed a novel training approach called the 'hybrid protein model' (HPM). Its goal is to compact the protein structure encoded in PBs into clusters of contiguous 3D structure fragments. Hence, the hybrid protein is a chimerical protein composed of N sites and for which every position i is defined by a probability distribution $f_i(b_n)$, with b_n denoting one of the 16 PBs ($n = 1, 2, \ldots, 16$). Figure 2 summarizes the principle of the training. Every 3D protein structure is cut into overlapping series of L PBs. A fragment of 13 PBs is taken randomly from the non-redundant databank, here *mmmmmmnopafkl* (cf. Figure 2a). The fragment is presented to the hybrid protein and a score is computed to find the best fit between the fragment and the hybrid protein region. The position associated with the highest score is noted by a

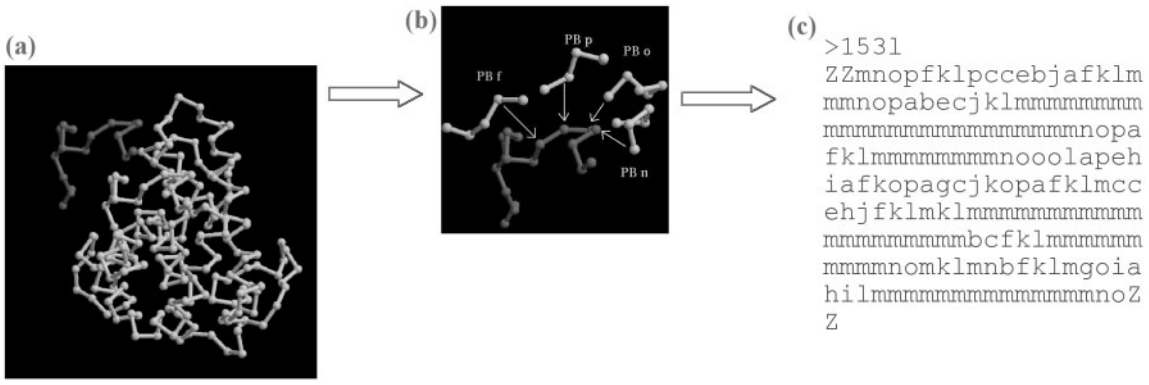

Fig. 1. Coding of a lysozyme (code PDB: 153l) into protein blocks (PBs). (a) 3D structure representation using XmMol (Tuffery, 1995) and Raster 3D (Merritt and Bacon, 1997). (b) Local coding of 153l N-ter with PBs *n*, *o*, *p* and *f*. (c) Representation of protein 153l in terms of PBs. The two symbols 'Z' denote the extremities of the structure.

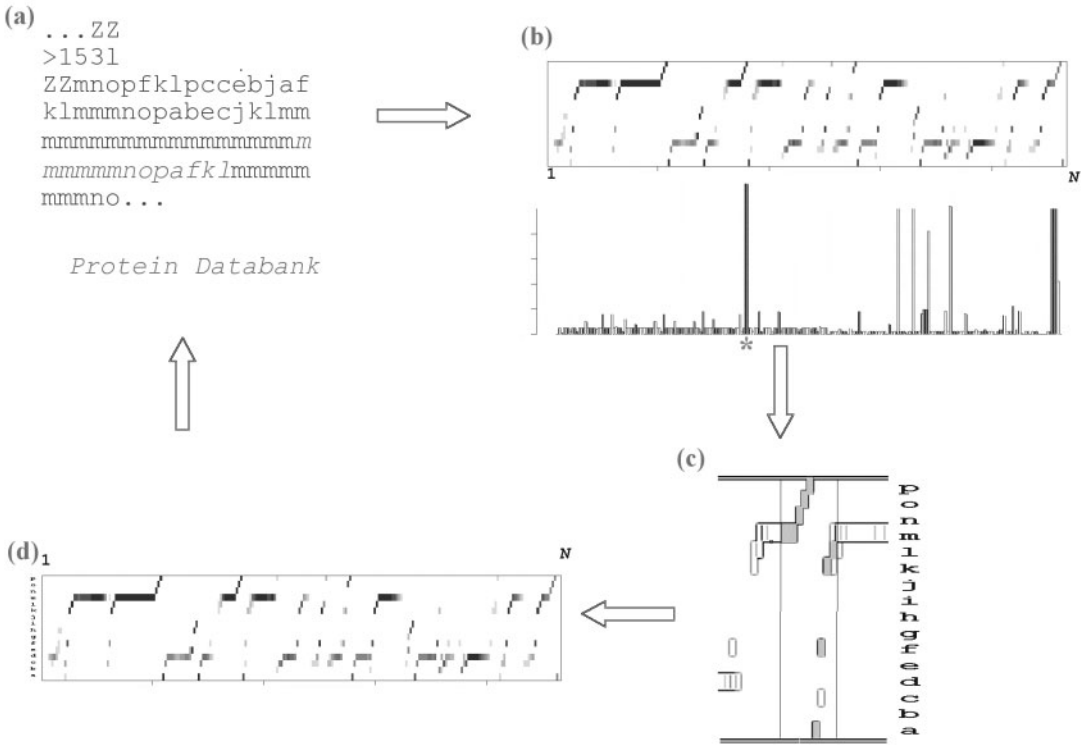

Fig. 2. Principle of the 'hybrid protein model' (HPM) training (see text).

star (cf. Figure 2b). The submatrix around this position is slightly modified to learn this fragment: the frequency of the PBs of this fragment increases (colored box), while others decrease (cf. Figure 2c). Another fragment is then presented for training (cf. Figure 2d). This training aims at the progressive improvement of the *S* score, used to cluster the protein fragments.

For a local structure $\boldsymbol{F}$ from the databank, we compute an S_i score at each position i of the hybrid protein:

$$S_i = \sum_{k=-w}^{k=+w} \ln\left[\frac{f_{i+k}(b_k)}{f_R(b_k)}\right]$$

where k denotes the position of block b_k in fragment $\boldsymbol{F}$ of length L $(= 2w+1)$. The index $k = 0$ indicates the middle

of the fragment (PB n in the example). The frequency $f_R(b_k)$ corresponds to the reference frequency of PB b_k observed in the databank.

The S_i score is the log odds score, i.e. the logarithm of the ratio of likelihoods between two hypotheses: the first is that fragment $\boldsymbol{F}$ is defined by a randomly-ordered series of PBs, and the second that it is built according to the PB distribution of the hybrid protein.

The most similar local structure prototype is determined by searching for the position i_0, the index for which S_i is maximal, i.e. $i_0 = \text{argmax}[S_i]$. The positions $i_0 - w$ to i_0+w will be modified slightly to increase the resemblance between this part of the hybrid protein and local structure $\boldsymbol{F}$. In position $i + k$, the value of the xth PB $f_{i+k}(b_x)$ is changed to $[f_{i+k}(b_x)+\alpha]/[1+\alpha]$. The value of x' of all the other PBs decreases by $[f_{i+k}(b_{x'})]/[1 + \alpha]$. The learning coefficient, α, is a user-fixed value (e.g. $\alpha = 5 \times 10^{-3}$) and decreases during the iterations.

This transformation allows us to increase the score of fragment $\boldsymbol{F}$. The training is progressive and must examine the entire local structure databank C times.

Continuity between the consecutive positions (i.e. contiguous fragment clusters) is ensured. After training, every position i of the hybrid protein characterizes a cluster of folds of length L that are structurally similar. This site maintains its continuity with the contiguous site $i - 1$, because they have in common for the score computation the distribution of $L - 1$ PBs.

Improvements for obtaining an optimal hybrid protein

Two properties characterize an optimal hybrid protein:

(i) *consistency between consecutive hybrid positions*, i.e. when a fragment $\boldsymbol{F}$, extracted from a given 3D protein structure in the first position p of the the sequence, is located in position i_0 in the hybrid protein, $\boldsymbol{F}'$ shifts by one residue in the sequence (into position $p + 1$) must be generally located in position $(i_0 + 1)$ in the hybrid protein. A 3D protein structure is thus represented by a limited number of hybrid protein regions.

(ii) *lack of redundancy within the hybrid protein*. Two regions of length L are redundant in the hybrid protein when their L consecutive PB distributions are similar.

To maximize continuity, we introduced a procedure called 'baby training,' by which long fragments are introduced in the early cycles, with the fragment size progressively reduced in the following cycles. This procedure should promote continuity in hybrid protein learning. It is called 'baby training' by analogy with the development of visual perception in infants: in the early months, they perceive crude patterns and progressively their vision becomes finer. For example, in our study, we want to build a protein hybrid able to cluster the local folds of 13 PBs ($L = 13$). In this strategy, the training starts with fragments of 23 PBs and continues on to fragments of 18, 15 and finally 13 PBs.

To minimize redundancy, we compute a confusion matrix $C(i, j)$ of dimension $N \times N$ during one cycle. A fragment $\boldsymbol{F}$ is counted in the element (i_0, j_0) of the matrix when its optimal position in the hybrid protein is i_0 (i.e. $i_0 = \text{argmax}[S_i]$) and its second best is j_0 (i.e. $j_0 = \text{argmax}_{(i \neq i_0)}[S_i]$). From this matrix, symmetrized for the analysis, we search for redundant regions, that is, those longer than a user-defined value l_0. Indeed, this matrix defines the diagonals of minimal lengths l_0, i.e. $C(i, j), C(i + 1, j + 1), \ldots, C(i + l_0 - 1, j + l_0 - 1)$ that occur more than a given n_{lim}. In our study, we set l_0 and n_{lim} at 10 and 85 respectively.

Among the paired regions, we select the longest and delete the other ones. Only one region longer than l_0 is rejected from the hybrid protein per cycle. After a certain number of cycles, the reduction of hybrid protein length is stopped. We obtain an optimal hybrid protein, conditioned by the choice of the parameters l_0 and n_{lim}.

Description of contiguous local fold clusters

The hybrid protein is a series of PB distributions from which an S score can be computed to cluster the structurally similar protein fragments of length L. From the optimal hybrid protein, we can superimpose the protein backbones of fragments located in a given position and assess the structural variability at each point by the *rmsd*, the quantity conventionally used in molecular modeling. We can also calculate the associated amino acid composition (i.e. the frequency of a given amino acid in a given position of the fragment) and assess the informativity of the sequence within the cluster.

Entropy can be computed to quantify the diversity of the PBs of the hybrid protein: $H_i = -\sum_{b=1}^{16} f_i(b).\ln[f_i(b)]$, where i denotes the position of the site and f_i the corresponding PB distribution, b indexes a given PB. The transformation of the entropy into $N_{\text{eq}} = \exp[H_i]$ allows us to assess the PB diversity in terms of 'equivalent number of PBs': N_{eq} varies between 1 (i.e. only one PB is present) and 16 (i.e. every PB occurs at the same frequency).

Each position is associated with a set of 3D protein fragments of length L with their corresponding sequences. To quantify the occurrence of each amino acid at each site, we computed the N occurrence matrices of dimensions $L \times 20$. Then we normalized this information into Z-scores to determine which amino acids were over- and under-represented (de Brevern *et al.*, 2000). Hence, positive Z-scores (respectively negative) correspond to

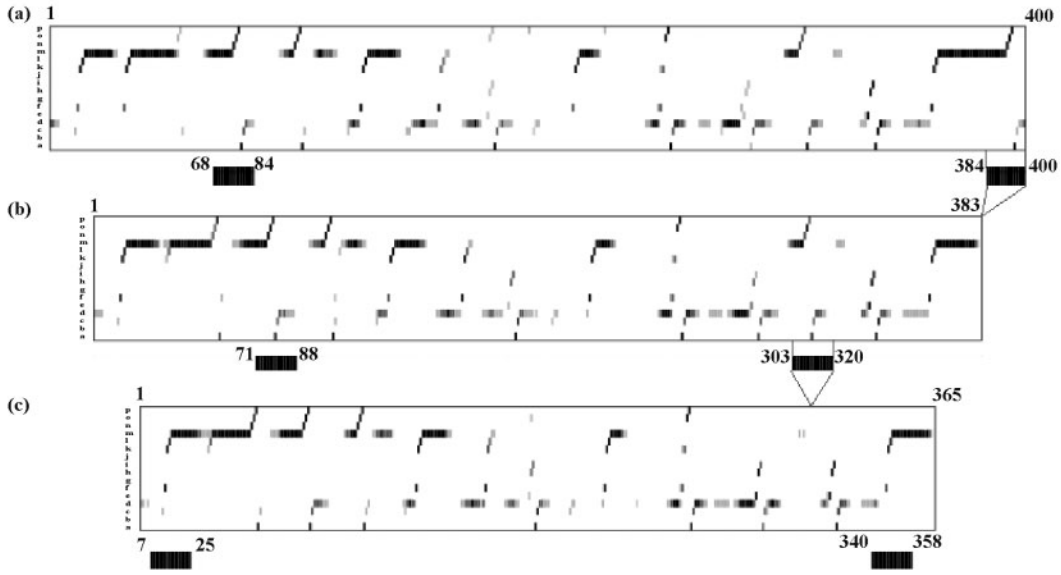

Fig. 3. Successive reductions of the hybrid protein. (a) Hybrid protein at cycle 5 shows redundancy between two regions located in positions [68:84] and [384:400]. The latter is deleted. (b) At cycle 6, the new redundant regions are located in positions [71:88] and [303:320]. (c) At cycle 7, the redundancy is between positions [7:25] and [340:358]. The black boxes under the hybrid proteins indicate the locations of the redundant regions.

over-represented (respectively, under-represented) amino acids.

Another index used to quantify the sequence informativity in a given position of the cluster is the Kullback–Leibler asymmetric divergence measure (noted *KLd*, Kullback and Leibler, 1951). With a denoting a given amino acid, it is defined as

$$KLd(\mathbf{p_i}, \mathbf{q}) = \sum_{a=1}^{20} p_i(a) \ln\left(\frac{p_i(a)}{q(a)}\right)$$

It quantifies the contrast for a given position between the amino acid frequencies observed in the cluster $\mathbf{p_i}$: $\{p_i(a)\}_{a=1,\ldots,20}$ and a reference probabilistic distribution $\mathbf{q}\{q(a)\}$, i.e. the probability of each amino acid type in the database. For a fold cluster, a *KLd* profile is built by computing this quantity for the $(L + 4)$ positions that compose the sequence windows associated with the fragments of length L.

RESULTS AND DISCUSSION

Evolution of the hybrid protein

The hybrid protein is initially defined by a series of N almost identical PB distributions: $f_i(b_x) = f_R(b_x).(1 + \epsilon_i)$, where $f_R(b_x)$ is the frequency of PB b_x in the database and ϵ_i a random value in the range $[-\tau; +\tau]$ (in our study, τ is fixed at 0.10). We then readjust $f_i(b_x)$ to obtain a total sum of 1 per site i. To avoid possible bias at the extremes, the hybrid protein is close, since the Nth site is contiguous with the first.

The initial size of the hybrid protein is fixed at 400. During the training, it shrinks as the redundancy criterion is applied. Figure 3 shows the successive reductions in length through 3 successive iterations; the black boxes under the hybrid protein indicate the redundant regions. The similarity of the PB composition is clear. After the deletion of each region, a new cycle of fragment training begins. This procedure is controlled by two parameters: the minimum size l_0 of the redundant region to be deleted, and the redundancy level $n_{\lim}$ (i.e. the minimum number of fragments located in these redundant regions). The parameters we chose—$l_0 = 10$ and $n_{\lim} = 85$—enable us to one an optimal hybrid protein with 233 positions after 30 cycles ($= C$). This reduction is substantial, i.e. 167 of 400 positions deleted.

Description of the optimal hybrid protein

Figure 4 reports the results of the training after 30 learning cycles (i.e. C-value). Figure 4a shows the composition of the PBs along the hybrid protein. The regular secondary structures (those associated with PBs m and d) are clearly detectable: eight types of α-helices distinguishable

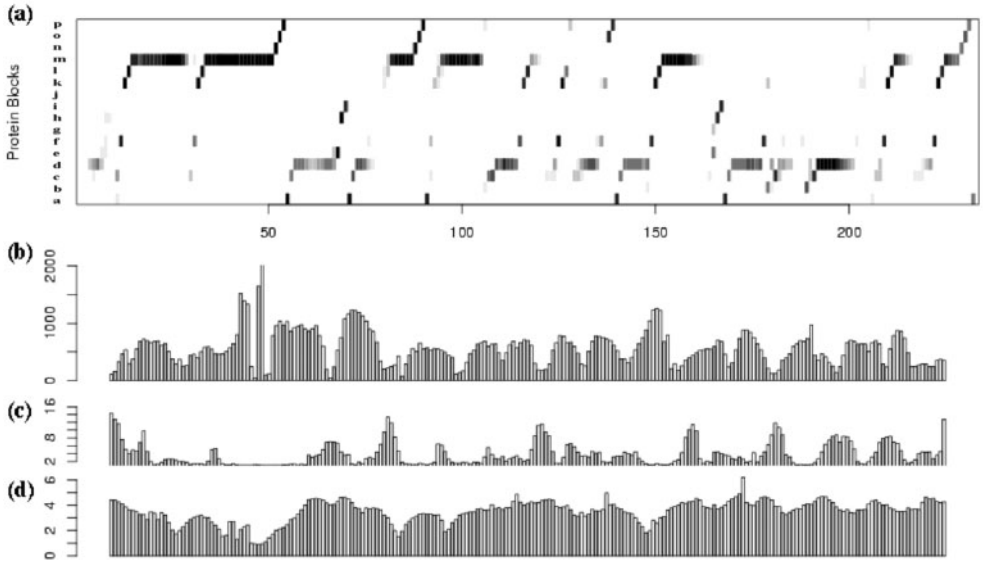

Fig. 4. Optimal hybrid protein. (a) PB distribution along the hybrid protein, with a final length of 233. (b) The number of protein fragments per site (the position 43 is associated with 4649 observations). (c) N_{eq} distribution along the hybrid protein (N_{eq} varying between 1 and 16). (d) $rmsd$ values (computed for all the fragments of length 13 Cα associated at each site).

by their sizes (4–20 PBs) and at least eight β-strands. All the transitions between regular secondary structures can be identified: α-helix to α-helix positions [34:51] and [82:105], α-helix to β-strand positions [41:67] and [100:125] β-strand to α-helix positions [57:87], [132:160] and [182:215], and a series of β-strand to β-strand between positions [57:67], [110:150] and [171:201].

We also detect different motifs located at the beginning or the end of regular secondary structures; these include *fkl* and *nop* for α-helices, and *ac* and *ehia* for β-strands. Figure 4b shows the number of fragments along the hybrid protein. The distribution is almost uniform, with the smaller sizes corresponding to coils. Figure 4(c) gives the variation of the 'equivalent number of PBs' N_{eq} index along the hybrid protein. Most positions are highly specific, with an N_{eq} value close to 1. The high N_{eq} values correspond to transition regions, such as turns between two strand positions [135:142], coils between two α-helix positions [29:34], and long coil positions [398:15], or to distorted secondary structures such as β-strand positions [177:192]. After 3D superimpositions of the protein fragments of a cluster (associated with a given position), we computed the *rmsd*. Figure 4(d) shows the variation of the *rmsd* per site. This quantity assesses the quality of the training in terms of structural variability. The average *rmsd* is 3.4 Å for the fragment 13 Cα in length. The range is [0.92 Å; 6.20 Å], which is quite good compared with other classifications (Wojick *et al.*, 1999). The lowest value corresponds to a long regular α-helix located in position 48, the highest to a variable coil in position 183. Globally, the local structures with a low *rmsd* (less than 2 Å) are α-helices or transitions between a β-sheet and an α-helix, or between two α-helices.

Examples of fragment clusters

Figure 5 shows three examples of fragment clusters. They are located in positions 48, 138 and 183 in the hybrid protein and correspond to three different *rmsds* levels (0.92, 3.43 Å and 6.20 Å, respectively). The first is a regular α-helix, the second a turn between two consecutive β-sheets and the third a long curved coil.

The occurrence matrix of PBs, i.e. the number of times a given PB is observed in a given position for all the fragments of the cluster, is displayed for each example (see Figure 5d–f). We observe that the first is defined principally by the motif $m_{10}nop$, i.e. a regular α-helix of 10 consecutive *m* PBs, followed by the often observed terminal series *nop*. The second example has a signature in terms of PBs: $d(d,\ f)(f,k)(k,n)opacd_3$, according to the most frequent PBs in every position. The motif *opac* is often observed at the beginning of a β-sheet. The third example has a variable signature, but we see that in positions 2 and 5 PBs *f* and *c* appear respectively at a frequency of more than 89%. Its first five blocks are characteristic of a coil, and the next four follow a distorted β-strand that is composed primarily of a mix of

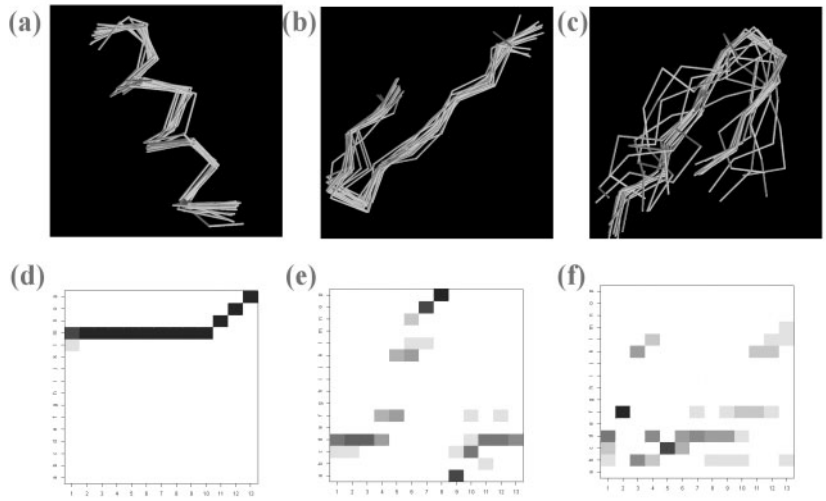

Fig. 5. Examples of local prototypes (or local fold clusters). 3D superimpositions of fragments associated with the sites : (a) # 48 (fragments associated with positions [41:55], $rmsd = 0.92$ Å), (b) # 138 (positions [131:145], $rmsd = 3.43$ Å) and (c) # 183 (positions [176:190], $rmsd = 6.20$ Å) and the PB occurrence matrices (displayed in grey levels) associated with those sites: (d) # 48, (e) # 138 and (f) # 183. The symbol # denotes the number of the central amino acid positions.

PBs—*b*, *c* and *d*. These occurrence matrices are similar to the associated hybrid protein parts, but not identical. The hybrid protein is a scoring matrix used to cluster fragments, not a PB occurrence matrix.

Figure 6 gives the occurrence matrices of the amino acids observed in each cluster, normalized into Z-scores, and the *KLd* profiles within the sequence window of 15 residues. Some amino acids are more commonly found in α-helices while others have a predisposition for β-sheets. Globally, we again find the standard propensities of amino acids in the regular secondary structures: over-representation of alanine (positions 3–14 in Figure 6a) in the α-helices, with charged residues at their N-caps and C-caps, such as lysine, arginine and glutamic acid (positions 1–6 and 12–14); similarly, glycine and asparagine in coils (position 15 for the first cluster–Figure 6a—, position 10 for the second cluster—Figure 6b—and aliphatic residues for the β-sheets—Figure 6b). The under-representation of amino acids in certain positions is also informative. For example, leucine, methionine and alanine are favored in α-helices, but not found in β-sheets (compare the Z-score matrices of Figure 6a and b). Similarly, the central region of the second motif (positions [7:10]) shows hydrophobic residues to be under-represented, with serine, threonine and proline possible. This region corresponds to a turn between two β-sheets. Figures 6d–f are the *KLd* profiles of the same clusters, i.e. the sequence informativity for every position of the window. Figures 6d and e show that some positions are highly informative ($KLd > 0.15$), mainly because of the presence of glycine or asparagine. The third cluster, on the other hand, shows substantial structural variability ($rmsd = 6.2$ Å) and low sequence informativity ($KLd < 0.15$).

Relevance of the HPM improvements

We assessed the advantages of the two procedures introduced in the training: 'baby training' and hybrid protein size reduction by redundancy deletions. To test the baby training, we eliminated it, i.e. we set a fixed size for the series of PBs to be learned by the hybrid protein ($L = 13$ in our study). The result was *a priori* surprising: much of the hybrid protein (approximately 150 sites) was not used for the training (figure not shown). In fact, only 250 sites were needed to stack the fragments of that size (13). Moreover, the hybrid protein was cut up into lengths close to L. Accordingly, while the baby training procedure is useful for extending continuity between the regions and promoting an optimal distribution of the fold clusters, some parts of the hybrid protein appear somewhat redundant. The second important point is controlling the length N of the hybrid protein by deleting the redundancy. This control is ensured by the parameters l_0 and $n_{\lim}$. The parameter l_0 should be approximately or slightly less than fragment size L (in our study $l_0 = 10$). Higher values do not reduce size significantly. Lower values cut up the hybrid protein into small pieces. The other parameter, $n_{\lim}$, i.e. the minimum number of occurrences between two redundant regions, is essential to control the size of

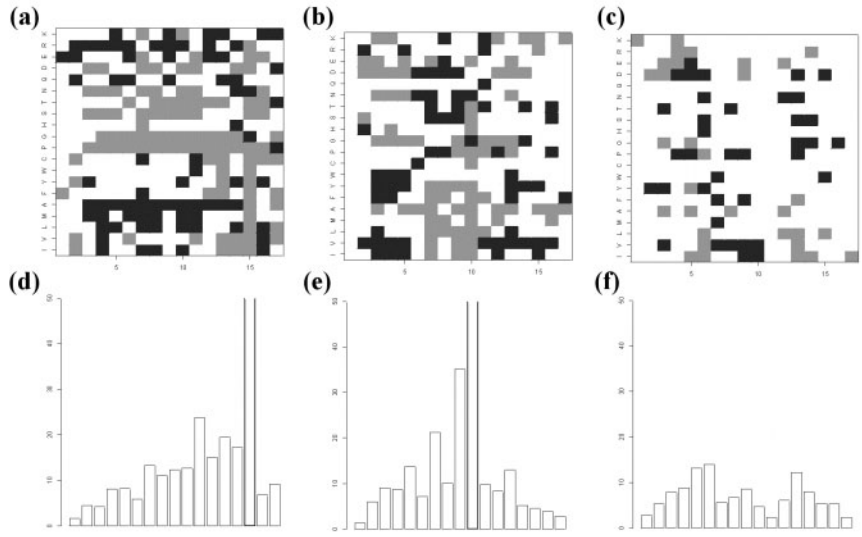

Fig. 6. Amino acid propensities for the three prototypes. (a) to (c) Amino acid occurrence matrices normalized into Z-scores associated with the prototypes in Figure 5. The elements are characterized by Z-values: grey ($Z < -1.96$), white ($-1.96 < Z < 1.96$) and black ($Z > 1.96$). (d) to (f) *KLd* profiles within the sequence window of 15 residues. The *KLd*-values are multiplied by 100.

the hybrid protein. A value of 85 was chosen to enable a repertory to be characterized with a nearly uniform fold distribution. With a lower n_{lim}-value ($= 80$), the size of the hybrid protein decreases from 233 to 175, indicating some fragmentation. Conversely, when the value is higher (=90), it increases to 325 and some redundancy remains.

An optimal hybrid protein thus attains a delicate equilibrium between deletion of the redundant regions and conservation of the fold continuity.

CONCLUSION AND PERSPECTIVES

The 'hybrid protein model' (HPM) presented here is based on the notion that 3D protein structures are composed of structural domains similar enough to be stacked. The principle of HPM is very simple: it seeks to optimize a series of PB distributions used for a clustering score, finally enabling a repertory of contiguous local protein folds to be built.

The value of HPM is that it compacts a non-redundant protein structure databank into a limited set of local folds. The 'baby training' and 'deletion of the redundancy' procedures presented in this study make it possible to build an optimal repertory with reasonable structural variability and a high level of sequence informativity.

Through this hybrid protein, we dispose of a collection of fragments able to form a protein structure whose amino acid propensities are defined. This rich collection should be very useful for the prediction of protein structures through fold recognition or for *ab initio* modeling. In a previous work, we used the example of two cytochromes to illustrate the advantages of using the hybrid protein to extract similar local folds in these proteins (de Brevern and Hazout, 2001). The success of the HPM method in fold recognition must be validated in a further work. Using the sequence informativity found in the amino acid occurrence matrices associated with different fold clusters, we should be consistently able to pick out candidates for folding simulations from the repertory.

In conclusion, a procedure of stacking local folds lets us build an optimal repertory sufficiently rich to be used in molecular modeling.

ACKNOWLEDGEMENTS

We would like to thank Pierre Tufféry for the superimposition software and Catherine Etchebest for fruitful discussion. This work was supported by a grant from the Ministry of Higher Education and Research (Ministère de l'Enseignement Supérieur et de la Recherche) and from 'Action Bioinformatique inter EPST' number 4B005F. AdB is supported by a grant from the Medical Research Foundation (Fondation pour la Recherche Médicale).

REFERENCES

Andrade,M.A., Casari,G., Sander,C. and Valencia,A. (1997) Classification of protein families and detection of determinant residues with an improved self-organizing map. *Biol. Cyber.*, **76**, 441–450.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bray,J.E., Todd,A.E., Pearl,F.M.G., Thornton,J.M. and Orengo,C.A. (2000) The CATH Dictionary of Homologous Superfamilies (DHS): a consensus approach for identifying distant structural homologues. *Protein Eng.*, **13**, 153–165.

Bonneau,R. and Baker,D. (2001) Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173–189.

Bonneau,R., Tsai,J., Ruczinski,I., Chivian,D., Rohl,C., Strauss,C.E.M. and Baker,D. (2001) Rosetta in CASP4: progress in *ab initio* protein structure prediction. *Proteins*, **45**, 119–126.

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

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

de Brevern,A.G., Camproux,A.C., Etchebest,C., Hazout,S. and Tuffery,P. (2001) Beyond the secondary structures: the structural alphabets. *Recent Adv. Prot. Eng.*, **1**, 319–331, Pandalai SG ed. Research signpost, Trivandrum, India.

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.

Ferrán,E.A., Pflugfelder,B. and Ferrara,P. (1994) Self-organized maps of human protein sequences. *Protein Sci.*, **3**, 507–521.

Fiser,A. and Sali,A. (2002) MODELLER: generation and refinement of homology models. *Meth. Enzymol.*,in press

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure. *Proteins*, **23**, 566–579.

Govindarajan,S., Recabarren,R. and Goldstein,R.A. (1999) Estimating the total number of protein folds. *Proteins*, **35**, 408–414.

Hadley,C. and Jones,D.T. (1999) A systematic comparison of protein structure classifications SCOP, CATH and FSSP. *Structure*, **7**, 1099–1112.

Hanke,J., Lehmann,G., Bork,P. and Reich,J. (1999) Associative database of protein sequences. *Bioinformatics*, **15**, 741–748.

Holm,L. and Sander,C. (1994) Parser for protein folding units. *Proteins*, **19**, 256–268.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.

Jones,S., Stewart,M., Michie,A., Swindells,M.B., Orengo,C. and Thornton,J.M. (1998) Domain assignment for protein structures using a consensus approach; characterization and analysis. *Protein Sci.*, **7**, 233–242.

Jones,S., Marin,A. and Thornton,J.M. (2000) Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, **13**, 77–82.

Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.

King,S.M. and Johnson,W.C. (1999) Assigning secondary structure from protein coordinate data. *Proteins*, **35**, 313–320.

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

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

Labesse,G., Colloc'h,N., Pothier,J. and Mornon,J.-P. (1997) P-sea: a new efficient assignment of secondary structure from Cα. *Comput. Appl. Biosci.*, **13**, 291–295.

Merritt,E.A. and Bacon,D.J. (1997) Raster3D: Photorealistic Molecular Graphics. *Methods Enzymol.*, **277**, 505–524.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chotia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 526–540.

Noguchi,T., Matsuda,H. and Akiyama,Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.*, **29**, 219–-220.

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

Pearl,F.M.G., Lee,D., Bray,J.E,, Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.

Ring,C.S., Kneller,D.G., Langridge,R. and Cohen,F.E. (1992) Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.*, **5**, 685–699.

Schneider,G. (1999) How many potentially secreted proteins are contained in a bacterial genome. *Gene*, **237**, 113–121.

Schneider,G., Sjoling,S., Wallin,E., Wrede,P., Glaser,E. and von Heijne,G. (1998) Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *Proteins*, **30**, 49–60.

Siddiqui,A.S. and Barton,G.J. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.*, **4**, 872–884.

Stahl,M., Taroni,C. and Schneider,G. (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural networks. *Protein Eng.*, **13**, 83–88.

Swindells,M.B. (1995) A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci.*, **4**, 103–112.

Taylor,W.R. (1999) Protein structural domain identification. *Protein Eng.*, **12**, 203–216.

Tuffery,P. (1995) XmMol: an X11 and motif program for macromolecular visualization and modeling. *J. Mol. Graphics*, **72**, 67–72.

Unneberg,P., Merelo,J.J., Chacon,P. and Moran,F. (2001) SOMCD: Method for evaluating protein secondary structure from UV circular dichroism spectra. *Proteins*, **42**, 460–470.

Wernisch,L., Hunting,M. and Wodak,S.J. (1999) Identification of structural domains in proteins by a graph heuristic. *Proteins*, **35**, 338–352.

Wojick,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.