Theoretical Chemistry accounts 2001; 106(1/2): 36-47

Compacting local protein folds with a "Hybrid Protein Model"

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.

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

short title: Compacting local protein folds.

key words: protein block, unsupervised classifier, fold similarity.

September 11, 2000

Abstract

The "Hybrid Protein Model" (HPM) is a fuzzy model for compacting local protein structures. It learns a non-redundant database encoded in a previously defined structural alphabet composed of 16 protein blocks (PBs) [1]. The hybrid protein is composed of a series of distributions of the probability of observing the PBs. The training is an iterative unsupervised process that for every fold to be learnt consists of looking for the most similar pattern present in the hybrid protein and modifying it slightly. Finally each position of the hybrid protein corresponds to a set of similar local structures. Superimposing those local structures yields an average root mean square of 3.14 Å. The significant amino acid characteristics related to the local structures are determined. The use of this model is illustrated by finding the most similar folds between two cytochromes P450.

Introduction

Predicting the three-dimensional structure of a protein sequence from its amino acid sequence is not an easy task without knowledge of the 3D-structure proteins that share high sequence similarity rates. Proteins with high homology rates can be treated by homology modeling based on physico-chemical scales and geometric constraints. These modeling techniques often rely on statistical methods [2, 3, 4]. When there are no proteins with high homology rates, *ab initio* modeling uses simulations to search for the most probable protein. The results of such modeling are constantly improving [5, 6], but it is often limited to small proteins. There are other approaches; they include *threading*, which involves assessing the compatibility of the target sequence and various protein structure fragments [7, 8, 9, 10]. All these modeling techniques require the initial use of a non-redundant protein database. Databases with a known rate of sequence [11, 12] and structural similarity [7, 13] are currently available.

The aim of this paper is to present a method that compacts a protein structure database into one "hybrid protein". The learning step uses a structural alphabet. Various structural alphabets have been defined, each composed of a different number of fold clusters, 6 for Fetrow *et al.* [16], 4 to 7 for Rooman *et al.* [17], 12 for Camproux *et al.* [14, 15], 13 for Bystroff and Baker [18], and 100 for Unger *et al.* [19] and Schuchhardt *et al.* [20]. We have chosen to use 16 Protein Blocks (PBs), 5 Cα in length. This alphabet approximates protein 3D-structures with reasonable accuracy, and we have already used it in a Bayesian prediction method [1]. It encodes the protein structure database to be analyzed.

Because proteins have common local structures of various length, we have tried to stack those structures locally. The process consists in building a concatenation of local structures that share common and distinct parts. In fact, the possible variations of the PB content can be expressed by a probability law for the 16 PBs. Accordingly, this stacking of the local structures results in a "hybrid protein", i.e., a series of probability laws that gives the occurrence of observations of each PB type at each position. A hybrid protein is represented by a matrix, the dimension of which is 16 x N (N denotes its length). Compared with conventional clustering and the definition of a partition into independent subsets, the hybrid protein characterizes a series of structurally dependent subsets: that is, it maintains the sequentiality of the local structures. The main purpose of this approach is to stack all the fragments of the local structure database into the hybrid protein.

After training with the database, every local structure of every protein is located in a given position of the hybrid protein.

The hybrid protein thus obtained has particularities in terms of its structures and its amino acid sequences. This approach allows similar local structures to be classified in a

given site. We have assessed the training by computing the *root mean square deviation* for every type of local structure compacted in the hybrid protein. We then computed and analyzed the amino acids associated with each local structure to evaluate the specificity of each site. The relations between the amino acid distributions of the hybrid protein and the PB distribution for each site are analyzed. As an illustration, we will describe certain parts of the hybrid protein structurally and will point out some characteristics of the amino acid distributions associated with them. Finally, we will present one application of the hybrid protein: the search for similar protein local structures in two cytochromes P450.

Materials and Methods

Database of encoded 3D protein structures

The database contains 553 non-redundant proteins with a sequence similarity of less than 25% [11, 12] (i.e., 118 915 amino acid residues). The dihedral angles ϕ and ψ describing the protein backbone have been computed. We have ignored the variations of the ω angle; they are often slight and are directly related to the other two [22].

With an unsupervised classifier that takes into account the dependence between the successive local folds along the proteins, we had previously determined a set of local prototypes, called Protein Blocks (PBs), to approximate the protein backbones locally. Table 1 describes this structural alphabet, or PB set. PBs m et d are the prototypes of the central α-helix and the central β-sheet, respectively. Their repetitive structure conformations are standard; the average numbers of repeats are 2.74 for PB d and 6.74 for PB m. PBs a through c primarily represent β-sheet N-caps and e and f, C-caps; g through j are specific to coils, k and l to α-helix N-caps, and n through p to their C-caps. This categorization is crude and provides only a partial view of the PB locations in the protein folds. In particular, the PBs considered in the category "β-sheet N-caps" (from a to c) are really different from one another; they also occur in coils. The particular utility of such an alphabet is that it enables several different N- and C-caps to be found for the two regular secondary structures; it also has 4 PBs mainly present in coils. Figure 1 shows an example of superimposed fragments for each PB [21]. Our structural alphabet thus allows a reasonable approximation of the 3D-structures of the proteins. In fact, 50% of the angles ϕ et ψ are approximated with a difference of fewer than 21^o, and less than 3% with a difference of more than 90^o, relative to reality. Globally, the average *root mean square deviation* (*rmsd*) is 0.58 Å. This alphabet has also been used to predict local protein structure with a Bayesian approach [1].

For the present study, we have encoded the complete proteins of the database into series of PBs (cf. figure 2a). Each protein structure is split into a series of overlapping fragments, each defined by 5 consecutive carbons $C\alpha_{n-2}$, $C\alpha_{n-1}$, $C\alpha_n$, $C\alpha_{n+1}$, and $C\alpha_{n+2}$. All of these in turn are described by a series of 8 angular values $\mathbf{V}$ (ψ_{n-2}, ϕ_{n-1}, ψ_{n-1}, ϕ_n, ψ_n, ϕ_{n+1}, ψ_{n+1}, ϕ_{n+2}). Hence, the attribution of a protein fragment to a protein block is based on a maximal similarity criterion. The metric is an Euclidean distance called *root mean square deviation on angular values*, computed from dihedral angles [20]. Thus the database is composed of 116 703 PBs (cf. figure 2b). The goal of the "Hybrid Protein Model" is to use the connections between the PB sequences to characterize the series of dependent local structure clusters of 10 PBs.

"Hybrid Protein Model" (HPM)

I. Definition

The hybrid protein. To compact local structure databases, we have developed a novel training approach called the "Hybrid Protein Model" (HPM). The hybrid protein is a chimeric protein composed of N sites and for which every position i is defined not by one PB, but by a probability distribution $f_i(b_x)$, with b_x denoting one of the 16 PBs (x=1, 2,..., 16 and i=1,...,N).

Figure 2 summarizes the learning steps of the HPM. The hybrid protein trains by optimally locating every local structure of the database composed of L PBs (L is a user-fixed parameter) in a region of this protein and then by modifying the distributions located in this region (cf. figure 2c to 2f). The hybrid protein stacks all the similar local structures in a given position and creates a fuzzy prototype of this subset of folds. The principal advantage of the HPM is that it conserves the overlapping between the consecutive structure prototypes.

Location of the local structure in the hybrid protein. The learning step consists in computing for a local structure $\boldsymbol{F}$ a score S_i at each position i of the hybrid protein (cf. figure 2c:

$$S_i = \sum_{k=1}^{k=L} ln[f_{i+k-1}(b_k^*)]$$

where k denotes the index associated with a PB position in $\boldsymbol{F}$ (k=1,2,...,L). Each local structure $\boldsymbol{F}$ is defined by L consecutive blocks $\{b_1^*, b_2^*, ..., b_L^*\}$ (in our study L=10 PBs and thus represents 14 Cα). The score S_i (i.e., the logarithm of the likelihood of observing the local structure $\boldsymbol{F}$ in a given site i) measures the similarity between the local structure

and a given region of the hybrid protein; this region is characterized by a subset of PB distributions (indices: $i, i+1, ..., i+L-1$)(cf. figure 2d). The most similar local structure prototype is determined by searching for the position i_0, the index for which S_i is maximal. That is, $i_0 = argmax[S_i]$ (cf. figure 2e).

Local modification of the hybrid protein. The positions i_0 to i_0+*L*-1 will be slightly modified to increase the likeness of this part of the hybrid protein to the local structure $\boldsymbol{F}$ (cf. figure 2f):

if $b = b_k^*$ (i.e., the PB at position k in the local structure), then

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b) + \alpha}{1+\alpha}$$

if $b \neq b_k^*$ (i.e., the other PBs at position k), then

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b)}{1+\alpha}$$

The symbol $\leftarrow$ denotes "calculated value replaces previous value." The learning coefficient α is equal to $\alpha_0/(1 + t/\nu)$, where α_0 is the initial rate of learning (e.g., $\alpha_0 = 0.1$ in our study), t the number of local structures of L blocks already introduced in the training, and ν the number of local structures in the database. The training is progressive and thus needs to examine the entire local structure database C times. For example, $C = 15$ in our study, and steps c-f in Figure 2 must be performed 15 times for every local structure of the database. A complete database reading is called a cycle. In the first cycle, the training coefficient α is kept constant ($\alpha = \alpha_0$) so that the hybrid protein will be substantially modified [23].

Initialization. The hybrid protein is initially defined by a series of N PB distributions $f_i(b_x)$. These are almost identical because $f_i(b_x) = f(b_x).(1 + \epsilon_i)$, where $f(b_x)$ is the frequency of the PB b_x in the database and ϵ_i is a random value in the range [-τ;+τ] (in our study, τ is fixed at 0.20). We have readjusted $f_i(b_x)$ to obtain a total sum of 1 per site i. The hybrid protein is close to this since the Nth site is continuous with the first site.

II. Interpretation step

Each site i of the hybrid protein is a complete set of local structures of length L. This site maintains its continuity with the contiguous site *i-1*. Consequently sites i and *i-1* have *L-1* distributions in common.

(i) Motifs and occurrence matrix at every position

To quantify the importance of each amino acid at each site, the amino acid occurrences are computed, by the simple method of counting for each series of L PBs associated with one site s the corresponding series of L residues in the window of range [$+L/2$:-$L/2$]. These occurrences have been normalized into a Z-score: $Z_j^i = \left(n_j^i - n_a^i\right) / \sqrt{n_a^i}$, where n_j^i is the observed number of amino acid a in window position j and n_a^i is its expected number ($n_a^i = N_i.\{q(a)\}$, with N_i and $\{q(a)\}$ denoting respectively the number of local structures at position i and the observed frequency of amino acid a in the database. Hence, positive Z-scores (respectively negative) correspond to overrepresented (respectively underrepresented) amino acids at certain positions of the window.

ii) Entropy for quantifying the diversity of the PBs along the hybrid protein

Each site of the hybrid protein is defined by a probability distribution over the 16 PBs. An entropy can be computed to quantify the diversity of the PBs at every site:

$$H_i = -\sum_{b=1}^{16} f_i(b).ln[f_i(b)]$$

where i denotes the position of the site, b a PB type and f_i the corresponding PB distribution. Thus a lower entropy value is associated with sites (or a zone of sites) dependent on a limited number of PBs.

iii) KLd profile for quantifying the amino acid distribution specificity

After building the hybrid protein, we studied the specificity of the amino acid distribution of the central residue of the local structure patterns. We collected each entire local structure (i.e., series of L protein blocks) located at each site of the hybrid protein. We studied the specificity of the residue by comparing its amino acid frequencies with those observed in the database. We used relative entropy, also known as the Kullback-Leibler asymmetric divergence measure (noted KLd) [24]:

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

where a denotes a given amino acid. This quantifies the contrast between the amino acid frequencies observed in the central residues $\mathbf{p}_i$: $\{p_i(a)\}_{a=1,...,20}$ and a reference probabilistic distribution $\mathbf{q}\{q(a)\}$. In our study, the reference distribution $\mathbf{q}\{q(a)\}$ is the probability of each amino acid type in the database. The results are assessed by a χ^2-test, since the quantity $N_i.K(\mathbf{p}_i,\mathbf{q})$ follows a χ^2 distribution, with N_i the number of local structures associated with the site i. Thus, the positions with a highly specific amino acid distribution are associated with significant values.

(iv) Similarity clusters within the hybrid protein

In this section, we want to search for similar sites according to their amino acid composition and to study the relation between this composition and the protein block type. First, the amino acid distributions observed along the hybrid protein, normalized into Z-scores (see section "Interpretation step" *(i)*), are classified with a *k-means* clustering method [25] that allows a partition to be determined into a fixed number of clusters. Next, the sites belonging to a given cluster, called a similarity cluster, are compared according to their PB composition. This study allows us to determine the level of dependence between the amino acid distributions and the PB composition of the sites along the hybrid protein.

III. Search for similar local structures between two proteins with a HPM

A hybrid protein is interesting because it compacts a structural database. Moreover, it can be used to find similar local structures in two different proteins. The first step is to convert the 3D structures of both proteins into our structural alphabet. Next the local structures encoded with the structural alphabet are searched along the hybrid protein. Thus each series of positions characterizes the 3D structures. To find the structurally similar protein zones, a dotplot is computed. The boolean matrix (or dotplot) is built according to a rule of position index identity: set 1 when the same position of the hybrid protein is observed in the two encoded proteins, else 0. The dotplot is filtered by selecting diagonals of longer than G, which is initially set high, to extract the longest local structures. Then, by progressively reducing this parameter, shorter local structures are selected in protein regions not covered by previously defined local structures. Assuming that the proteins are structurally similar, the search is limited to local structures present in a nearby protein region (i.e., for position c of the first protein the search for similar structures is carried out in a zone $[c\text{-}\theta, c\text{+}\theta]$ for the second protein and inversely). In our study, θ is fixed at 50 residues. Hence, the *rmsd* is computed to assess every pair of similar local structures that is found. This approach allows us to search for structurally similar local structures with low sequence identity.

Results

The hybrid protein

a. General observations

Figure 3 reports the results of the training after 15 learning cycles (i.e., C value). Figure 3a shows the composition of the PBs along the hybrid protein. Analysis of the hybrid

protein suggests that the regular secondary structures (those associated with PBs m and d) are clearly detectable: three types of α-helices distinguishable by their sizes (2 to 4 PBs: positions [38:41]), 7 PBs [3:9] and 10 PBs [82:91]), as well as four β-strands (positions [15:23], [51:58], [66:69] and [74:78]). Different transitions between regular secondary structures are visible: α-helix to α-helix between positions [93:96], α-helix to β-strand [9:14], β-strand to α-helix [33:37] and [99:1], and β-strand to β-strand [23:28],[58:65] and [69:71].

The specificity of the protein hybrid sites is clear: only two or three separate PBs occur frequently in each. Moreover, continuity is maintained. When a local structure is optimally located in the position i_0, then the probability is 81% that the next local structure in the protein is in position i_0+1. The local structures are almost evenly distributed with a mean of 1115 observations per site and a range of [237-10401] (cf. Figure 3b).

b. Entropy of the PB distribution

Figure 3c represents the entropy computed along the hybrid protein and shows that each site is highly specific, with a maximum value of only 2.40 and a minimum of 0.41 (40% of the positions with less than 1.0). Three categories of entropy can be defined.

The first category involves entropy less than 1.0. This cluster contains the least variable sites, which correspond to the α-helix. We notice in positions [4:13] a central α-helix followed by a specific C-cap; these local structures begin most often with 6 PBm followed by a PBn, PBo, PBp, and PBa at sites 10-13. This is noted m_6nopa. The local structures in positions [79:91] are α-helices with an N-cap $fklm_{10}$. Since the training uses local structures 10 PB in length, this motif contains the following types of local structures: $fklm_7$, klm_8, lm_9, and m_{10}. The hybrid protein enables us to obtain a well-defined β-sheet C-cap [20:25] that contains the motif d_4eh as well as such short transitions as the dfk β-sheet C-cap [58:60] and fk [35:36].

The second cluster corresponds to intermediate zones with an entropy between 1.0 and 1.5. The longest corresponds to various distorted β-sheets with N-caps in positions [48:57], with a local structure of type iac_xd_{7-x}(x=1,2,3). The two most representative sites there are an fkl α-helix N-cap in sites [1:3] and β-sheets with a d_2-type local structure at sites [77:78].

After these well-defined regions, the third cluster groups the high entropy zones (more than 1.5). Four zones have a higher variability level (entropy more than 1.5) and correspond to long coils or distorted β-strands [26:34] and [38:45], turns between two strands [62:68] or coils between two α-helices, in the presence of β-sheets [94:100].

c. Structural stability of the hybrid protein

To assess the quality of the learning in terms of structural homogeneity, we computed the *rmsd* per site by superimposing all of the complete local structures 10 PBs long (= L) at each site (Figure 3d). The average *rmsd* was 3.14 Å. Variability was higher at only 6 sites (*rmsd* more than 5 Å) and lower at 14 (*rmsd* less than 1.0 Å).

The zones with the most structural variability were primarily associated with β-sheet residues. For example, the local structures located in [19:28] correspond to two different populations: (i) short d_2 β-sheets leading to another β-sheet, and (ii) d_4 β-sheets that may lead to an α-helix. Similarly, local structures [48:57] and [52:61] are associated with heterogeneous β-sheet lengths. Site 59, corresponding to local structures [55:64], is composed of three types of β-sheets: $d_4fkopac$, $d_4fklpac$ and c_2d_2fkopa. Site 75 [71:80] contains primarily a population of $bccd_6f$ β-sheets and another one beginning with $bdcd_3$. We therefore note that the most structurally variable zones principally involve β-sheets and β-sheet N- or C-caps. Because local structures associated at the same position share some common protein blocks, it is possible to obtain fuzzy regions in the hybrid protein.

The local structures with low variability are α-helices, such as the following: at site 7 (motif $m_8$no), α-helix N-cap; at site 38 ($fklm_8$), short α-helix; at site 79 (d_4fklm_3), sharp transition between a β-sheet and an α-helix; at site 84 (klm_8), an α-helix; at site 92 (m_5cmcd_2), a central α-helix to a β-sheet.

Some β structures are also well defined. These include local structures at the following sites: structures [11:20] at site 15, which is characterized by an *opa* succession going to an N-cap β; site 27 (*ehia*) going to a N-cap β; site 53 (*acddd*), an N-cap for β-sheets; site 55, another N-cap, shifted and ended by PB f; site 57, β-sheet C-cap finished by PB f; sites 62 and 66 (*opa*), transition between two β-sheets; site 97, strongly specific *cd* at the beginning and f at the end.

Indeed, the well-defined sites are associated not only with the α-helix, but also with short transitions between two β-sheets.

Examples

Figure 4 shows three examples of superimpositions of local structures. The first, located at site 7, corresponds to local structures [3:12] and has an *rmsd* of 0.3 Å and an entropy of 0.43 (Figure 4a); the second at site 73 [69:78], with an *rmsd* of 3.3 Å and an entropy of 0.72 (Figure 4b); and the third at site 56 [52:61], with an rmsd of 6.21 Å and an entropy of 1.39 (Figure 4c) [21]. These three examples show clear differences in their structural variability. Figures 4d to 4f show the amino acid occurrence matrix, normalized in Z-scores, that is associated with the local structures in those sites. The matrix composition for the first site (figures 4a and 4d) is standard for a central α-helix with an overrepresentation of

alanine and other non-polar residues. The presence of charged residues (lysine, arginine and glutamic acid) at positions 7 and 8 is usually associated with a C-cap [26], and the overrepresentation of glycine and asparagine at position 10 with a structural breaker. The second site (Figures 4b and 4e) is characterized by the presence of aliphatic amino acids classically associated with a β-sheet. The C-cap of the β-sheet is generally seen with overrepresentations of polar residues at position 4 and of isoleucine and valine at sites 5 to 9. The last motif (Figures 4c and 4f) is structurally fuzzy, but the local structures are globally similar. The amino acid composition is less informative than the structure. Only position 8 is highly specific, with an overrepresentation of glycine and asparagine and underrepresentation of 17 other amino acids.

Amino acid specificity

Figure 5a shows the amino acid distribution as transformed into Z-scores for the local structures of the structural database. The conventional propensities of amino acids in the regular secondary structures are again seen; these include the overrepresentation of alanine (positions [2:10] and [81:90]) for the α-helices, of the charged residues (lysine, arginine and glutamic acid) at their N-caps (positions [9:10] and [87:88]), aliphatic residues for the β-sheets (positions [18:24] and [51:58]), and glycine (positions [28:32] and [39:42]) within coils. Interesting amino acid specificities are also present: for example, phenylalanine is present in one β-sheet C-cap (at positions [55:58]), but is absent from the others ([21:23], [69:70] and [76:78]).

The specificity is much more accentuated than that found in our analysis of the PBs alone. The KLd profile computed from the amino acid distributions (cf. Figure 5b) shows the most informative sites. The KLd values over the threshold of 36 (chi-square of 19 degrees of freedom and a probability = 0.05) represent mainly but not only α-helices and β-sheets; strong transitions are also present, for example, *kl* coils in positions [59:60].

Different types of amino acid distributions are revealed when they are clustered into 12 distinct groups (Figure 5c). Some clusters are uninformative; for example, cluster 1 (25 sites) does not show significant specificity as to amino acids and has many different distinct PBs. Cluster 2 (5 sites) contains many glycine and asparagine overrepresentations with, however, very different PBs (for example, PBs *d*, *j* and *m*). Cluster 3 (2 sites) has the same type of overrepresentations, but associated with greater underrepresentation of other amino acids: it clearly characterizes structure breakers in PBs *i*, *j* and *k*. Clusters 4, 5, and 6 are mainly associated with β-sheets; clusters 7 and 8 with α-helices. In cluster 4 (17 sites), aliphatic amino acids (isoleucine and valine) are overrepresented in N-caps and central β-sheets. Cluster 5 (12 sites) is composed of PB *d* and other PBs

in β-sheet C- and N-caps (from PB *a* to *f*). Valine is slightly overrepresented, and charged residues are underrepresented. Cluster 6 (1 site at position 58) is a regular β-sheet with strong overrepresentation of non-polar amino acids and underrepresentation of polar amino acids. Cluster 7 (10 sites) is an α-helix, and, more specifically, the C-cap of an α-helix; leucine, methionine and polar residues are overrepresented. Cluster 8 (8 sites) shows an overrepresentation of alanine, methionine, isoleucine, and proline, and an underrepresentation of glycine. This cluster represents only the central α-helix. Cluster 9 (15 sites), with its overrepresentation of small polar amino acids, is specific to such breaker PBs as PB *f*, *h*, *a*, *l* and *o*. Cluster 10 (1 site at position 81) is associated with PB *h*: aliphatic residues and proline are underrepresented and polar residues strongly overrepresented. In cluster 11 (3 sites) small polar amino acids are overrepresented and non-polar amino acids and proline underrepresented. In cluster 12 (1 site at position 70) glycine is underrepresented and non-polar residues overrepresented. For the latter four sites, the principal protein block observed is PB *f*. There is a difference, however, between two of these protein clusters: in cluster 11, we observe PB *f* in a *fkl* series and for cluster 12 in a *fbd* series.

Search for common local structures in two cytochromes P450

Cytochromes P450 have been well described. They show [10-30]% sequence similarity but have common structural fragments that Haseman and coworkers [27] have characterized as common secondary structures (α-helix, 3_{10}-helix, π-helix, β-sheet and β-bulge). Jean *et al.* [28] determined some Common Structural Blocks (CSBs) for different cytochromes P450 with pairwise comparisons to model a new cytochrome P450 by homology. Our research focused on two cytochromes P450: $P450_{BM3}$ (code name PDB: 2hpd[29]), a bacterial fatty acid monoxygenase that is crystallized from *bacillus megaterium*, and $P450_{terp}$ (code name PDB: 1cpt[30]), from *pseudomonas sp.*

We began by using PBs to encode the protein structures and locating each series of 10 blocks in the hybrid protein. Figure 6 shows the positions of the two protein local structures along the hybrid protein. Note that most of the local structures are contiguous. Then we extracted the common local structures from a dotplot (see Material and Methods). Table 2 reports the characteristics of the 11 local structures, labeled from **I** to **XI**, common to the two cytochromes and figure 7 shows their 3D superimpositions. Jean *et al.* found 15 CSBs, which they labeled 1 to 13; 2A-2B and 12A-12B were clearly in the same local structure but were not found directly with their method. Our approach, which uses long local structures (L=10 PBs), found 11 of these 15 CSBs. It did not detect the following CSBs: 6, 11, and 13 are too short (12, 8 and 9 residues respectively), CSB 8

is much longer (17 residues, see section "Discussion"). Local structures **I**, **IV**, **V**, **VI**,**VII** and **IX** have approximately the same length as the corresponding CSB (see Table 2).

Examination of common local structures reveals that these findings diverge from Haseman's with common secondary structures [27] for local structures **II**, **IV**, **V**, **VI**, and **VII**, and from Jean's with CSBs [28] for local structures **II**, **VII**, **X**, and **XI**.

PB similarity rates reflect the structural proximity of the local structures. Globally, they exceed 73%, except for local structure **II**, with a rate of only 58.3%. Local structure **II** is not one of the CSBs: its *rmsd* is higher. The type of local structure is nonetheless quite similar, especially for the N-cap positions. One lone motif is present, the 3_{10} helix *b*, but it is not found in $P450_{BM3}$ [27]. Local structure **IV** is composed mainly of α- and 3_{10} helices. In the secondary structure classification, however, α-helix C' does not exist in $P450_{BM3}$ and the two 3_{10} helices *b* and *c* are placed at distant positions. This difference of classification does not result in a poor *rmsd* (0.8 Å). Local structure **V** has an *rmsd* between 450_{terp} and $P450_{BM3}$ of 1.8 Å. This value is due to the absence in $P450_{BM3}$ of π-helix E', which is instead included entirely in the α-helix E. This localized distortion increases the *rmsd*, but the local structure is globally the same. The problem of the 3_{10} helix *c* recurs in local structure **VI**, but this time for $P450_{BM3}$. For $P450_{terp}$, local structures **VII** and **VI** both correspond to CSB 4; they are, however, distinct for $P450_{BM3}$. Local structure **VI** corresponds to CSB 4 ($rmsd = 1.1$ Å), while local structure **VII** expresses less structural similarity ($rmsd = 3.6$ Å). The information yielded by the PB similarity rate is very similar: 75.0% for local structure **VII**, compared with 88.8% for **VI**. It must be noted, however, that this CSB, composed of an F helix [27], has different lengths in different cytochromes: it is the only CSB manually selected by Jean *et al.* Local structure **X** includes CSB 10 and the end of CSB 9. Local structure **XI** includes CSBs 12A (Cysteine pocket [27]) and 12B (L helix [27]), found separately, but very similar with an *rmsd* of 0.7 Å.

Our results indicate that the hybrid protein is an effective method for extracting local structures to show similar local structures in two proteins. A simple dotplot with PBs alone does not allow these local structures to be easily extracted, because the presence of repetitive structures interferes with the detection. The HPM is based on fuzzy learning of the series of PBs: every position of the hybrid protein directly concerns a subset of similar local structures.

Discussion

HPM is a new method that allows a structural database to be compacted. The learning step yields a hybrid protein with very good sequentiality between the local structures (81% are contiguous). The entropy shows that the learning step results in a satisfactory structural characterization of most of the sites (i.e., every site has no more than 3 significant PBs). Because the PBs are highly specific for each site, similar local structures with high PB identity are classified in the same hybrid protein site.

We have analyzed the influence of particular parameters on the results. The first was the length of the hybrid protein: $N = 100$ was chosen to enable the local structures to be characterized correctly. With $N > 100$, the hybrid protein sites contain too few local structures, and for $N < 100$, the number of poorly approximated local structures increases dramatically. Another parameter, α_0, controls both the quality and speed of learning. An α_0-value of 0.10 allows fast but crude learning; reducing the coefficient α during learning leads to the more precise training of the local structures. A lower α_0-value can be used, but will require more cycles of learning.

The third parameter we consider involves the random values for ϵ_i in the initialization of the hybrid protein (see section "methods"): they do not modify the resultant hybrid protein. We notice only a shift in the hybrid protein, which remains highly stable. Finally, the C-value is defined by the user. In practice, C cycles were performed until, after numerous database readings, no significant modification of the hybrid protein was observed. In conclusion, these parameters have only a minor influence on the construction of the hybrid protein, both because of the high sequential dependence between protein blocks and because of the presence of various geometrically stable structures (for example, the repetitive secondary structures and their N- and C-caps).

The computation of *rmsd* highlights the stability of most of the local structures. A higher *rmsd* is due to the location of both short and long β-sheets at the same sites. Overall the local structures maintain similar patterns. The *rmsds* of the other sites are correct, despite the heterogeneity of the structural database. Another feature that emerges from the structural classification is that regular structures are not the only local folds that are well approximated.

The examples given in figure 4 represent the three principal types of structural groups we observed. The structural description of the first is excellent and includes significant information about amino acid characteristics for most of the prototype positions. The second example is also well described but only some of the amino acid positions have significant under- or overrepresentations. The third type is the least frequent and shows only a few sites of interest in terms of amino acids. This type, with its higher *rmsd*,

lacks good structural definition but nonetheless selects local structures of the same overall kind. This characterization shows that the clustering by a Hybrid Protein Model correctly categorizes local structure prototypes.

The amino acid characterization shows substantial specificity at every site. As expected, the KLd revealed the standard repetitive structures. It also, however, showed that transitions of coil zones were also highly specific. This classification indicated that one group (25 sites) had no specificity, either for amino acid sequences or PBs. The 75 other sites, therefore, are highly specific, as the specificity of the PB classes always corroborates. Because, however, the sequence is not always specific to a local structure [31], a portion of the protein sequence is uninformative for predictive purposes [32]. Moreover, we note again the importance of some amino acids in N- and C-caps [26]. Some particular distributions of coil structures also depend on the amino acid composition.

The results shown in Figure 7 and Table 2 point out the interest of HPM for finding similar local structures with a good approximation. The PB identity of the 11 local structure pairs was not 100%: instead it varied from 73.3% to 95.0% except for pair **II** with 58.3% for a 12-length residue. The amino acid identity was low (less than 30%), as expected. The differences between the results from various methods involve more variable zones, such as local structure **II**. The divergence from the secondary structure classification reflects the equivalence of some short secondary structures in very distinct sites, without taking into account the modularity of cytochromes P450 for structures such as 3_{10} helix *c*: positions [173-176] for $P450_{terp}$ compared with [109-114] for $P450_{BM3}$). Similarly, for $P450_{BM3}$, local structure **V** is composed of a short β-sheet, a π-helix and an 12-length α-helix; for 450_{terp}, the same β-sheet is present, but the π-helix is included in an 18-length α-helix. The structural alphabet and the hybrid protein allow this type of fuzzy clustering. Our method did not find CSB 8, despite its low *rmsd* of 1.2 Å. Locally, 3 CSB 8 sites for $P450_{terp}$ and $P450_{BM3}$ differed in the hybrid protein. It was thus not found with our crude filtering in the dotplot method. This point will be taken in account in future improvements to this approach.

Our approach to searching for structural similarity is easier and implies neither systematic comparison of every pairwise local structure nor the use of an optimisation algorithm. The coding and homology steps are simple and fast.

Conclusion

This approach improves our knowledge about local structure patterns. We observed that HPM provided an accurate approximation of the structure of most sites, and not simply

for repetitive structures. Because many similar local structures are found at each site, the method should be applied in a *threading* approach to simplify the search for an adequate structure. To tackle the problem of badly approximated sites, the length L of local structures and the length N of the hybrid protein might be modified. The learning step and the computation of structural stability (i.e., *rmsd*) were carried out with the same values (i.e., L=10 PBs PBs and N=100); using variable lengths might improve the structural characterization. Thus Baker and Bystroff used a structural alphabet with variable lengths [18] and improved their *ab initio* modeling [33]. The main interest of our approach is that it takes into account the sequentiality of the protein blocks and thus implicitly that of the amino acids. This can be most useful for a prediction method. As our example of the two cytochromes P450 shows, the application of HPM to homology modeling has important potential. It will be very useful for *ab initio* prediction methods and molecular modeling by homology, especially when used with our improved Bayesian method prediction.

Acknowledgements

We would like to thank Romain Gautier for the database, Pierre Tufféry for the superimposition software and Anne-Claude Camproux and Catherine Etchebest for fruitful discussion. We would also like to acknowledge the constructive comments of the referees.

References

1. de Brevern AG, Etchebest C, Hazout S (2000) Proteins, forthcoming.
2. Bowie JU, Lűthy R, Eisenberg D (1991) Science 253: 164.
3. Sali A, Blundell TL (1993) J Mol Biol 234: 779.
4. Jaroszewski L, Rychlewski L, Zhang B, Godzik A (1998) Prot Sci 7: 1431.
5. Defay T, Cohen FE (1995) Proteins 23: 431.
6. Orengo CA, Bray JE, Hubbard T, LoConte L (1999) Proteins suppl.3: 149.
7. Madej T, Gibrat J-F, Bryant SH (1995) Proteins 23: 356.
8. Reva BA, Skolnick J, Finkelstein AV (1999) Proteins 35: 353.
9. Bienkowska J, Rogers RR jr., Smith TF (1999) Proteins 37: 346.
10. Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Proteins 37: 592.

11. Hobohm U, Scharf F, Schneider R, Sander C (1992) Protein Sci 1: 409.

12. Hobohm U, Sander C (1994) Protein Sci 3: 522.

13. Holm L, Sander C (1997) Nucl Acids Res, 25: 231.

14. Camproux AC, Tuffery P, Chevrolat J-P, Boisvieux J-F, Hazout S (1999) Protein Eng 12: 1063.

15. Camproux AC, Tuffery P, Buffat L, Andre C, Boisvieux J-F, Hazout S (1999) Theor Chem Acc 101: 33.

16. Fetrow JS, Palumbo MJ, Berg G (1997) Proteins 27: 249.

17. Rooman MJ, Rodriguez J, Wodak SJ (1990) J Mol Biol 213: 327.

18. Bystroff C, Baker D (1998) J Mol Biol 281: 565.

19. Unger R, Harel D, Wherland S, Sussman JL (1989) Proteins 5: 355.

20. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P (1996) Protein Eng 9: 833.

21. Tuffery P (1995) J Mol Graph 13: 67.

22. MacArthur MW, Thornton JM (1996) J Mol Biol 264: 1180.

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

24. Kullback S, Leibler RA (1951) Ann Math Stat 22: 79.

25. Hartigan JA, Wong MA (1979) Applied Statistics 28: 100.

26. Richardson JS, Richardson DC (1988) Science 240: 1648.

27. Haseman CA, Kurumbail RG, Boddupalli, Peterson JA, Deisenhofer J (1995) Structure 2: 41.

28. Jean P, Pothier J, Dansette PM, Mansuy D, Viari A (1997) Proteins 28: 388.

29. Ravichandran KG, Boddupalli SS, Hasemann CA, Peterson JA, Deisenhofer J (1993) Science 6: 731.

30. Hasemann CA, Ravichandran KG, Peterson JA, Deisenhofer J (1994) J Mol Biol 4: 1169.

31. Sternberg MJ, Islam SA (1990) Protein Eng 4: 125.

32. Shakhanovich EI (1996) Fold Des 1: 50.

33. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Proteins suppl.3: 171.