# Hybrid Protein Model (HPM) : A Method For Building A Library Of Overlapping Local Structural Prototypes. Sensitivity Study And Improvements Of The Training

Cristina Benros, Alexandre de Brevern, Serge Hazout

# HYBRID PROTEIN MODEL (HPM) : A METHOD FOR BUILDING A LIBRARY OF OVERLAPPING LOCAL STRUCTURAL PROTOTYPES. SENSITIVITY STUDY AND IMPROVEMENTS OF THE TRAINING.

Cristina Benros, Alexandre G. de Brevern and Serge Hazout
Equipe de Bioinformatique Génomique et Moléculaire, INSERM E0346,
Université Paris 7, case 7113,
2, place Jussieu, 75251 Paris cedex 05, France.
Phone: +33 1 44 27 77 30
Fax: +33 1 43 26 38 30
E-mail: benros,debrevern,hazout@urbb.jussieu.fr
Web: condor.urbb.jussieu.fr

**Abstract. Predicting protein structure from amino acid sequence is one of the main challenges of Genomics. Various computational methods have been developed during the last decade to reach this goal. However, the problem of structure prediction remains difficult. Before facing this complex problem, our goal is to focus on the accurate analysis of protein structures at a local level. In our study, we present an approach called "Hybrid Protein Model" (HPM) which uses a training procedure similar to the one of the Self-Organizing Maps. It allows the compression of a non-redundant protein structure databank into a library of overlapping 3D structural fragments. The "Hybrid Protein Model" carries out a multiple alignment of structural fragments. We present in this study an improvement of this strategy by introducing gaps in the local structures, and a sensitivity study of the training according to the control parameters. The library obtained is composed of a finite number of structural classes, each class including fragments sharing similar local structures. These classes are representative of the structural motifs found in the protein structures from the databank. Thus, this library constitutes an efficient tool for determining structural similarities between proteins and especially for predicting the local protein structure from the amino acid sequence.**

## INTRODUCTION

During the last decade, the number of completely sequenced genomes and potential protein sequences has greatly increased. On the other hand, the number of available 3D structures has increased but to a lesser extent. At the present time, more than 20.000 protein structures are available in the Protein Databank [2]. However, this databank does not represent all protein folds [12] and the different folds are not equally represented in this databank. Hence, computational methods are interested in defining 3D structural models from the sole knowledge of sequence. There are mainly three categories : (i) *comparative modelling* when a protein with a known 3D structure shares a good sequence identity with the studied sequence [1], (ii) *threading* which permits to characterize the best compatibility between the target sequence and a 3D protein structure extracted from a non-redundant databank [23] and finally (iii) *ab initio* and *new fold* methods which mimic the protein folding using physico-chemical and statistical parameters [3].

An important way to describe and predict local structures is through the secondary structures which consist in a local structural alphabet [6] defined by three states : $\alpha$-helix, $\beta$-sheet and coil (defined as not-$\alpha$ and not-$\beta$). The prediction rate of this 3-state alphabet is now -with the use of neural networks and sequence alignment- close to 80 % [17].

More complete and accurate structural alphabets have been defined. Two different approaches have been used with (i) a high number of local prototypes to insure an excellent approximation of protein 3D structures [22, 20], or (ii) a more limited number of local prototypes that implies a less accurate 3D local approximation but which in return could be used in a prediction approach [19, 11, 4, 5].

Thus, we have defined a structural alphabet. It is derived from the dihedral angles describing the protein backbone and is obtained from an unsupervised classifier. The 16 Protein Blocks (PBs), basis element of this structural alphabet, allow a correct 3D structure approximation ($rmsd$<0.42 Å) [7]. Local prediction has been estimated by a Bayesian approach and has shown that sequence information strongly induces the local fold, but stays coarse (prediction rate of 40.7% with one PB, 75.8% with the four most probable PBs). Furthermore, we have shown that the most common series of 5 consecutive PBs share a good relationship between sequence and structure [10].

From the description of the 3D structures in terms of PBs, we have elaborated a novel clustering method called "Hybrid Protein Model" (HPM) [8], for compressing the whole 3D fragments of a non-redundant protein structure databank. This method enables the construction of a library of overlapping structural prototypes able to approximate the global protein structure by a local approach [9].

The HPM is represented by a ring of neurons, and each neuron is characterized by successions of PB distributions. The training is similar to that of Kohonen networks [21, 14], i.e. a competitive learning (also referred to as

a WTA - "Winner Takes All" - method). The specificity of the HPM is that the information associated to the different neurons is overlapping. Thus, the modification of the PB distributions associated to the winning neuron influences the neurons located in its neighborhood.

The library obtained is composed of a finite number of structural classes (corresponding to the neurons), each class including fragments with similar local structures. These classes are representative of the structural motifs of the protein structures from the databank.

In the present study, we first present a sensitivity analysis of the HPM training according to the control parameters. Then, we describe an improvement of the HPM through the introduction of gaps into the 3D protein structural fragments. Moreover, a "training in depth" strategy enables to extract the preferential transitions within the HPM.

The HPM result is equivalent to a "structural profile" obtained by the multiple local alignment of the structural fragments (i.e. PB series) and by defining the PBs frequencies along the HPM. The introduction of gaps enables to improve the structure informativity similarly to conventional sequence alignment methods. It allows firstly, to take the length variability of the regular secondary structures ($\alpha$-helix and $\beta$-sheet) into account and secondly, to describe the heterogeneity of some local structures.

## MATERIAL AND METHODS

### Protein Blocks

16 Protein Blocks (labeled from PB$a$ to PB$p$) had been defined using an unsupervised classifier close to Self-Organizing Map [14] and Hidden Markov Model [18], which takes the preferential transitions existing between the PBs into account. These PBs allow a good structural approximation of complete protein 3D structures [7]. Figure 1 shows the backbones of the 16 PBs (visualization with the VMD software [13]). The protein blocks PB$a$ to PB$f$ are associated with the $\beta$-strand. The regular central $\beta$-strand is represented by the PB$d$. The blocks prior correspond to the N-caps, the following to the C-caps. In the same way for the blocks associated with the $\alpha$-helix, the block PB$m$ corresponds to the regular central part of a right $\alpha$-helix. The PBs from $k$ to $l$ and those from $n$ to $p$ characterize essentially the N- and C-caps respectively. Finally, the PBs from $g$ to $j$ are mainly found in coils.

### Local Structure Databank

The databank used in our study is composed of 675 non-redundant protein structures (less than 30% of sequence identity, R - factor $< 0.2$, a *root mean square deviation*, rmsd $> 10$ Å) taken from the PDB-RPDB site [16].

For each protein, we have stored the series of dihedral angles and the primary sequence. Each protein backbone was transformed into a signal

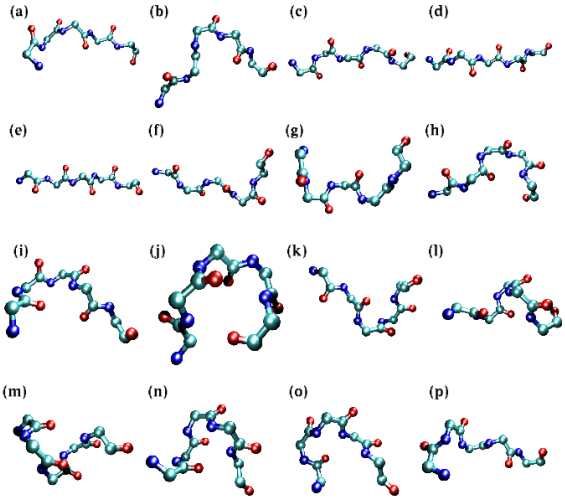


Figure 1: Backbone of the 16 Protein Blocks (PBs). PB$a$ to PB$p$ from left to right and from top to bottom.

corresponding to the succession of the dihedral angles $(\phi_i, \psi_i)$. Then, we have encoded the protein structures of the databank into series of PBs. Each protein structure was split into overlapping fragments, each defined by 5 amino acids described by 8 angular values $\mathbf{V}$ $(\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2})$.

The attribution of a protein fragment to a PB is based on a maximal similarity criterion. The metric used is an Euclidean distance called *root mean square deviation on angular values*, computed from the dihedral angles [20].

For this study, every 3D protein structure was cut into overlapping strings of $L$ PBs ($L$ fixed to 7) corresponding to local 3D structure fragments. The whole fragments constitute the "local structure databank". It contains 139 503 structural fragments.

## "Hybrid Protein Model" (HPM)

In another previous paper [8], we have developed a novel training approach called "Hybrid Protein Model" (HPM). Its goal is to compact the protein structures encoded into PBs into clusters of contiguous 3D structure fragments. The HPM is composed of $N$ sites. Each site is defined by a law of probability corresponding to the distribution of the 16 PBs. Hence, the HPM can be represented by a matrix of dimension $N$ x 16. A structural class (or neuron) represents a cluster of fragments with similar local structure, and is defined by $L$ successive probability distributions $f_i(b_n)$, with $b_n$ denoting one of the 16 PBs ($n$=1, 2, ..., 16). Two successive structural classes are overlapping since they have ($L$-1) sites in common. The last site is in continuity

with the first one. Thus, the HPM is closed and forms a ring of neurons.

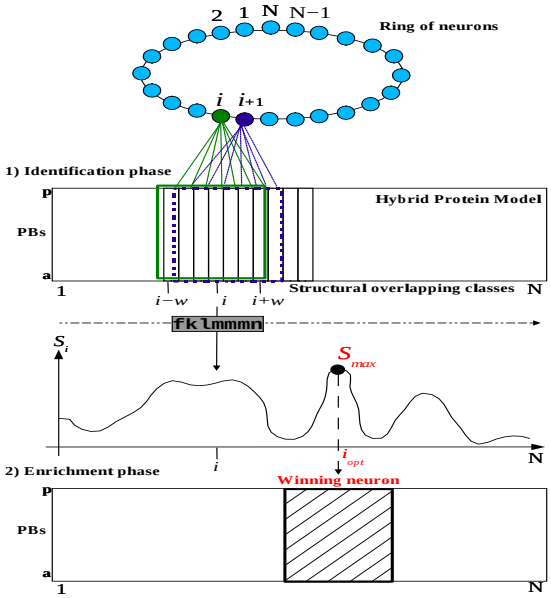


Figure 2: Training of the HPM : 1) For each structural fragment of $L$ successive PBs, taken randomly from the databank, we search for the most similar pattern present in the HPM. Consequently, a log odds score is computed along the HPM. The identification phase consists in selecting the structural class the most similar to the local structure presented, i.e. the one associated to the maximum score (or winning neuron). 2) The enrichment phase consists in slightly modifying the PBs distributions of the structural class corresponding to the site $S_{max}$, to increase the likeness between the optimal neuron and the fragment presented.

Figure 2 shows the two steps of the HPM training: (i) an identification phase, and, (ii) an enrichment phase. In the identification phase, a fragment of $L$ PBs represented by the string $F = b_{-w}, ..., b_0, ..., b_{+w}$, (with $L = 2w+1$, and $w = 3$ in our study), is taken randomly from the local structure databank. It constitutes the input signal. The fragment is presented to the Hybrid Protein Model and an adequacy score is computed to find the best fit between the fragment and a structural class of the HPM. The adequacy score $S_i$ at each site $i$ of the HPM is computed (1) :

$$S_i = \sum_{k=-w}^{k=+w} ln \left[ \frac{f_{i+k}(b_k)}{f_R(b_k)} \right] \quad (1)$$

where $k$ denotes the position of the protein block $b_k$ in the fragment $\boldsymbol{F}$ of length $L$. The index $k$=0 indicates the middle of the fragment. The frequency $f_R(b_k)$ corresponds to the reference frequency of the PB $b_k$ observed in the databank. This score corresponds to the probability of observing the fragment in a given site $i$ of the HPM according to its structural adequacy

with the prototype representing the neuron $i$. It is a log odds score, i.e., the logarithm of the ratio of likelihood between two hypotheses : the first one is that the fragment $\boldsymbol{F}$ is defined by a randomly-ordered series of PBs, and the second one that it is built according to the PB distributions of the HPM. The neuron the most similar to the local structure presented is determined by searching for the position $i_0$, the index for which $S_i$ is maximal, i.e. $i_0 = argmax[S_i]$. The training lies on a "competitive learning", i.e. the best neuron is "enriched" by the fragment PB content. Thus, in the enrichment phase, the submatrix around the optimal position (from $i_0 - w$ to $i_0 + w$) is slightly modified to learn this fragment, i.e. to increase the similarity between the winning neuron and the local structure. So, the frequency of the PBs observed in the structural fragment are increased, while the others are decreased. Hence, in position $i_0 + k$, the frequency value of the PB b, i.e. $f_{i_0+k}(b)$, is changed as follows (2), (3) :

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

$$f_{i_0+k}(b) \leftarrow \frac{f_{i_0+k}(b) + \alpha(t)}{1 + \alpha(t)} \tag{2}$$

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

$$f_{i_0+k}(b) \leftarrow \frac{f_{i_0+k}(b)}{1 + \alpha(t)} \tag{3}$$

The learning coefficient $\alpha$(t), is initially fixed at a value $\alpha_0$ and decreases in a regular way (4).

$$\alpha(t) = \frac{\alpha_0}{1 + K.t/T} \tag{4}$$

where $t$ denotes the number of fragments already presented to the HPM, $T$ the total number of fragments in the training databank, and $K$ a parameter that controls the speed of the $\alpha$ decrease. During a cycle, all the fragments of the training databank are randomly presented to the HPM. The training is progressive since several cycles are carried out until the stabilization of the PBs distribution laws. This enrichment phase allows one to improve the specificity of the different neurons. Then, the process starts again, another fragment is presented for training, and so on, until the presentation of all the fragments of the structural databank.

In Kohonen network, a procedure of diffusion is applied in the neighborhood of the winning neuron: the weights of the neurons are modified according to their relative distance to the winning neuron. On the contrary, no diffusion is performed in our training. In the Hybrid Protein Model, the principle of overlapping around the winner is used and allows a progressive diffusion by continuity.

### Library of Overlapping Local Structural Prototypes

The HPM is composed of a series of $N$ PB distributions (i.e. a ring of neurons) for which an adequacy score can be computed to cluster the fragments structurally similar. Every succession of $L$ consecutive PB distributions characterizes a local structural cluster. So it is possible to build a local prototype from each fragment cluster. At every HPM site, we define the average 3D prototype by superimposing the protein backbones of the fragments of $L$ PBs. The final prototype chosen as a representative corresponds to the fragment whose similarity with this local fold is maximal. We assess the structural variability by computing the *rmsd*.

A Shannon entropy can be calculated to quantify the PBs diversity along the HPM (5).

$$H_i = -\sum_{b=1}^{16} f_i(b).ln[f_i(b)] \tag{5}$$

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 $N_{eq}$ (6) 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). A low $N_{eq}$ value indicates a cluster of fragments structurally homogeneous according to our structural alphabet.

$$N_{eq} = exp[H_i] \tag{6}$$

Another index used to quantify the structural informativity in a given site is the Kullback-Leibler asymmetric divergence measure or relative entropy (noted *KLd*, [15]). It is defined by equation (7).

$$KLd(\mathbf{f}_i, \mathbf{f}_R) = \sum_{b=1}^{16} f_i(b)\ ln\left(\frac{f_i(b)}{f_R(b)}\right) \tag{7}$$

It quantifies the contrast for a given site between the PB frequencies observed in this site $\mathbf{f}_i$: $\{f_i(b)\}_{b=1,\ldots,16}$ and a reference probabilistic distribution $\mathbf{f}_R$: $\{f_R(b)\}$, i.e. the probability of each PB type in the databank. For a given HPM, we can assess the global PB informativity by the average $KLd(\mathbf{f}_i,\mathbf{f}_R)$ for the whole sites (8).

$$KLd_{ave} = \frac{1}{N}\sum_{i=1}^{N} KLd(\mathbf{f}_i, \mathbf{f}_R) \tag{8}$$

This quantity is used to assess the improvement of the training according to the control parameters variations.

**Sensitivity Study Relative to the Control Parameters**

The training is dependent on the HPM size ($N$), and on the learning parameters $\alpha_0$ and $K$.

(i) Optimal HPM size : in a previous paper [9], we have proposed a strategy for obtaining an optimal Hybrid Protein Model, i.e. for determining the HPM size $N$ for a given $L$ value of the fragment length (and *a fortiori* for a given 3D structure databank). Two properties have been analyzed : (a) the quality of the continuity (or sequentiality) between the consecutive HPM sites, and (b) the redundancy within the HPM. The first property of continuity specifies that when a fragment $\boldsymbol{F}$, extracted from a given 3D protein structure in position $p$ of the sequence, is located in position $i_0$ in the HPM, the fragment $\boldsymbol{F'}$ shifted by one residue in the sequence (into position $p+1$) must be in general located in position $(i_0+1)$ in the HPM. Indeed, HPM should maintain the protein backbone continuity to a maximum. This assumes that the HPM size is high.

In the opposite, the second property of redundancy specifies that any fragment must be clustered in only one site. The redundancy is highly frequent when several fragments present adequacy scores in different HPM regions close to the maximal score, i.e. the fragments can be indiscriminately classified in these different regions. Consequently, a low redundancy assumes a low $N$-value.

On this basis, we have defined two procedures. The first one, called "baby learning", enables to insure a high continuity in the training of consecutive overlapping fragments. The principle consists in learning longer fragments and progressively reducing their size ($L$ value). This procedure favors the continuity during the training. Parallel to this first procedure, we have assessed the redundancy by defining a confusion matrix according to the proximity of the scores along the HPM. From this information, we deleted some HPM redundant regions during the training.

In the present study, we have chosen a simplified strategy based on the concept of "free space occupation", which consists in starting with a long HPM (i.e. an important number of neurons) and in estimating the number of neurons not selected by the training, i.e. the number of structural classes remained empty. This assumes that the redundancy of the protein structure fragments is high, hence a limited HPM size is needed. We have assessed this size according to the training parameters.

(ii) Optimal values of the control parameters : we have carried out a study of the training parameters $\alpha_0$ and $K$. From the different trainings, we have analyzed the variations of the following quantities: the average $N_{eq}$-value for the $N$ sites ($N_{eq}.ave$), the maximal $N_{eq}$-value ($N_{eq}.max$, the minimal value of 1 is often found in the HPMs), and the average relative entropy for the $N$ sites ($KLd_{ave}$).

**Improvements for Obtaining a More Accurate Structural Library**

The first improvement insures a maximal sequentiality between the consecutive Hybrid Protein Model positions in the training of the 3D local protein structures. We have replaced the strategy of "baby learning" by a simpler procedure of forcing. The principle is as follows : a fragment is presented to the HPM, and its optimal location ($i_{max}$) is determined (corresponding to $S_{max}$). The fragment in position ($p$ - 1) previously examined is located in the HPM site $i'$. The selected neuron or site ($i_{opt}$) is modified as follows (9):

$$\begin{cases} i_{opt} = i' + 1, \text{ if } S_{i'+1} \geq \gamma S_{max} \\ i_{opt} = i_{max} \text{ otherwise.} \end{cases} \quad (9)$$

With this rule, we compare the score $S_{i'+1}$ obtained when the sequentiality is maintained, with the maximal score $S_{max}$ reduced by the factor $\gamma$ ($\gamma < 1$). According to the parameter $\gamma$, the continuity forcing is less or more efficient. As we previously said, a high continuity between the consecutive HPM positions requires an important number $N$ of neurons.

Consequently, for an Hybrid Protein Model of reduced size ($N < 50$), an elevated forcing in the training leads to an increase of the $N_{eq}$-value, i.e. a fuzzy structural library. We have assessed the training sensitivity according to the variation of the parameter $\gamma$.

The goal of the second improvement is to derive as much structural information as possible from the structure protein databank. It consists in introducing possible gaps in the structural fragments, i.e. in the PBs strings analyzed. This strategy is similar to the one used in multiple sequence alignment methods.

We can assume that the determinism observed in the architecture of a protein backbone is disturbed in a short region. So, we introduce only one possible gap of length $g$ ( $0 \leq g \leq L$-1) in the string of $L$ PBs, its location is *a fortiori* inside the chain of characters. For example, the string *fklmmno* may be transformed into *fklm- - -mno* where a gap of length 3 is introduced between two protein blocks specific of the $\alpha$-helix ($m$).

The adequacy score with gap is expressed as (10), (11), (12), (13):

(i) if the gap is introduced in position -$j$ before the central PB located in position $i$:

$$S^{-}_{i,j} = max_g \left\{ \sum_{k=-w}^{k=-j} ln \left[ \frac{f_{i+k-g}(b_k)}{f_R(b_k)} \right] + \sum_{k=-j+1}^{k+w} ln \left[ \frac{f_{i+k}(b_k)}{f_R(b_k)} \right] \right\} \quad (10)$$

(ii) if the gap is introduced in position +$j$ after the central PB located in position $i$:

$$S_{i,j}^{+} = max_g \left\{ \sum_{k=-w}^{k=+j} ln \left[ \frac{f_{i+k}(b_k)}{f_R(b_k)} \right] + \sum_{k=+j+1}^{k+w} ln \left[ \frac{f_{i+k+g}(b_k)}{f_R(b_k)} \right] \right\} \quad (11)$$

(iii) the adequacy score with gap is :

$$S_i^g = max_{j=1,w-1} \left\{ S_{i,j}^{-}, S_{i,j}^{+} \right\} \quad (12)$$

This formulate specifies that the introduction of a gap of length $g$ implies a shift before or after the central position $i$, within the structural information ($L$ PB distributions) associated with the neuron $i$. Implicitly, the neuron carries out an extension of its structural information into a larger neighborhood of neurons. This score calculation assumes that a gap has no cost whatever its length. To not favor the introduction of gaps, we penalize the score $S_i^g$.

The final rule becomes :

$$i_{opt} = argmax_i \left\{ S_i, \delta.S_i^g \right\} \quad (13)$$

In every site, we compare the score $S_i$ without gap to the maximal score $S_i^g$ with gap reduced by the coefficient $\delta$ ($\delta < 1$). The optimal location index $i_{opt}$ corresponds to the maximal score with or without gap.

**"Training in Depth"**

The remaining part of the paper is devoted to carry out a "training in depth". This strategy aims at introducing in the training only the most frequent fragments, i.e. those having a high adequacy score (i.e. $S_i > S_0$ where $S_0$ is a threshold of selection). Then, the process is reiterated by reducing the $S_0$-value, hence favoring the introduction of fragments less frequent. This approach allows us to first extract a sub-library of structural prototypes more accurate, and progressively to enrich it by the others prototypes weakly represented.

## Description of the Hybrid Protein Model

## RESULTS

Figure 3 gives the final result of the training. The parameters have been chosen after the sensitivity study (that is detailed in the following section). Hence, the optimal training parameters values chosen are : the HPM size $N$=120, the initial learning coefficient $\alpha_0$=0.20, its decrease $K$=0.5, the continuity forcing coefficient $\gamma$=0.6, the penalty coefficient for introducing gaps $\delta$=0.25 and the selection threshold value $S_0$=5.

Thus, the HPM, which represents the library of structural motifs, is composed of 120 overlapping structural classes, i.e. 120 neurons. Each class includes fragments sharing similar local structures encoded into PBs (fragments

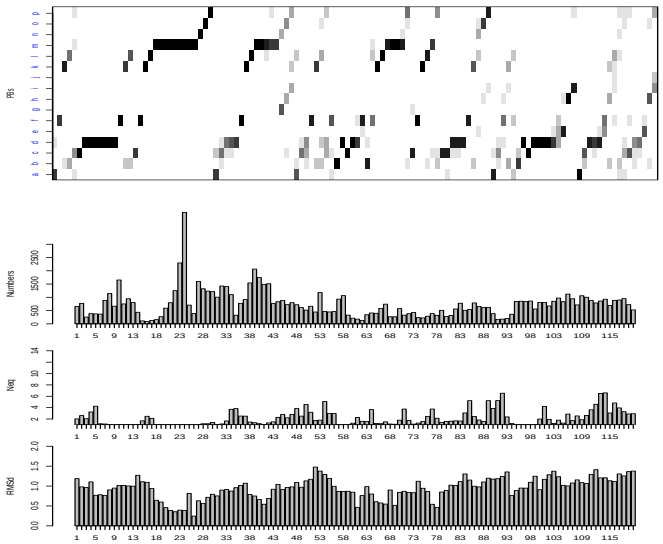


Figure 3: a) The final Hybrid Protein Model b) Distribution of the fragments c) Entropy-derived diversity index ($N_{eq}$) for quantifying the specificity along the HPM d) Average *rmsd* (Å) (from top to bottom).

of 7 successive PBs). The HPM can be defined as a probabilistic protein. It is equivalent to a "structural profile" resulting from the local alignment of the structural fragments. Here, we realized a multiple alignment of structural fragments with gaps.

Figure 3a shows the structural information of the $N$=120 neurons, i.e. the matrix of PBs distributions. The analysis of the HPM enables to locate regions of regular secondary structure and preferential transitions between them. We distinguish the presence of three $\alpha$-helices of various size (from 4 to 9 PBs) associated with series of PB *m*. The first one is located from the site 21 of the HPM to the site 29, the second one from 41 to 45, and the last one from 67 to 70. There are also seven types of $\beta$-strand (from 3 to 7 PBs long) associated with series of PB *d* and located in the HPM sites [7;13], [35;37], [58;61] (fuzzy region, break induced by PB c), [80;82], [94;100] (break induced by PB c), [107;109] and [116;117] (fuzzy region). Preferential transitions between these regular structural regions are observed. Some regions of the HPM are more fuzzy and correspond to coil regions. The gray level indicates the PB frequency in the distribution.

Figure 3b indicates the number of fragments located in the different HPM structural classes. Globally, the distribution of the fragments is uniform (826.8 fragments per site in average) except for the regular secondary structure regions. The larger $\alpha$-helix region (sites [21;29]) contains a high number of structural fragments, essentially located at the site 24 (13790 fragments). The lowest numbers of fragments are associated to coil regions, with a mini-

mum value of 86 fragments for the site 16. The analysis of the log odds score distribution shows that the fragments are well classified. These scores, which enabled during the training to assign a fragment to a neuron, are relatively high (with a maximum equal to 21.9).

Figure 3c gives the variation of the "equivalent number of PBs" ($N_{eq}$). This entropy-derived diversity index allows us to characterize the specificity of each HPM site. In average, the $N_{eq}$ value is 2.19 PBs per site. 90% of the sites have a $N_{eq}$ value lower than 4 PBs. Some sites are really specific, like for example regions of $\alpha$-helix and $\beta$-strand. 70 sites (i.e. 58.3%) have a $N_{eq}$ value lower than 2 PBs. The $N_{eq}$ values are higher for coil regions. This index allows us to extract the fuzzy regions such as the sites 46-51, 90-94 and 110-120. The maximum value is equal to 6.6 and is obtained for the site 114. These results show that the HPM sites are well determined.

Figure 3d shows that the structural approximation of the structural fragments by the prototypes is very satisfactory. Each structural class includes fragments with similar local structures. These clusters are homogeneous. The *rmsd* value for each site is comprised between 0.25 Åand 1.48 Å, with an average value equal to 0.95 Å. The lowest value is obtained for the site 26 ($\alpha$-helix region). The maximal *rmsd* value is associated to the site 52 (coil).

Figure 4 shows four structural prototypes associated to the neurons 14, 32, 64 and 104. They correspond to a $\beta$-strand with its C-cap, a turn between an $\alpha$-helix and a $\beta$-strand, a transition between a $\beta$-strand and an $\alpha$-helix, and a $\beta$-hairpin respectively.

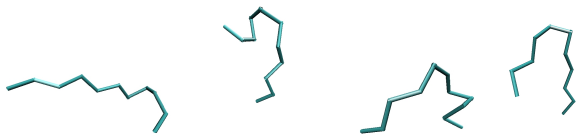

Figure 4: Average prototypes associated to the neurons 14, 32, 64 and 104 respectively (from left to right).

### Sensitivity Study Relative to the Control Parameters

(i) Learning coefficients $\alpha_0$ and $K$

Table I gives the results of the sensitivity study according to the variations of the parameters $\alpha_0$ and $K$. We point out : (i) a decrease of the average equivalent number of PBs ($N_{eq}$.ave) and of its maximal value ($N_{eq}$.max) when the learning coefficient $\alpha_0$ increases. This may be explained by the fact that a strong training in certain regions insures a best location of the fragments highly frequent. The contrast between the structural information of the neurons (i.e. PB distributions) and the structural information without learning (i.e. the $f_R(b)$ distribution) measured by the quantity $KLd_{ave}$

| $\alpha_0$ | K | $N_{eq}$.ave | $N_{eq}$.max | $KLd_{ave}$ | Seq. |
|---|---|---|---|---|---|
| 0.005 | 1 | 3.23 | 9.00 | 1.61 | 58.1 |
| | 0.5 | 3.16 | 11.41 | 1.62 | 57.1 |
| 0.025 | 1 | 3.00 | 12.48 | 1.62 | 56.7 |
| | 0.5 | 2.91 | 12.32 | 1.62 | 56.6 |
| 0.05 | 1 | 2.98 | 12.28 | 1.61 | 55.8 |
| | 0.5 | 2.97 | 10.98 | 1.66 | 58.3 |
| 0.10 | 0.5 | 2.89 | 10.32 | 1.67 | 57.6 |

TABLE 1: SENSITIVITY OF THE HPM ACCORDING TO THE DIFFERENT LEARNING COEFFICIENTS

increases weakly. The quantity "sequentiality" (Seq.) which quantifies the proportion of consecutive fragments located in consecutive neurons, is stable (around 57%). The change of the parameter $K$ from 1 to 0.5 leading to a lower decrease of the learning coefficient $\alpha_0$ does not permit a significant improvement of the training. From this study, we have chosen a higher learning coefficient $\alpha_0 = 0.20$ and a $K$-value of 1.

(ii) Forcing factor $\gamma$

By reducing the factor $\gamma$, we have obtained an increase of the continuity : around 57%, 74% and 99% for the respective $\gamma$-values of 0.8, 0.6 and 0.4. The others quantities $N_eq$.ave and $N_{eq}$.max are not largely modified.

(iii) Gap cutoff ($\delta$)

| Gap cutoff ($\delta$) | $N_{eq}$.ave | $N_{eq}$.max | $KLd_{ave}$ | Seq. |
|---|---|---|---|---|
| 1.0 | 1.98 | 5.35 | 1.83 | 67.7 |
| 0.9 | 1.98 | 6.89 | 1.87 | 73.7 |
| 0.8 | 2.07 | 6.22 | 1.82 | 73.9 |
| 0.7 | 2.13 | 5.63 | 1.77 | 74.6 |
| 0.6 | 2.09 | 6.57 | 1.81 | 75.3 |
| 0.5 | 2.18 | 6.52 | 1.74 | 75.9 |

TABLE 2: INTRODUCTION OF GAPS IN THE STRUCTURAL FRAGMENTS ($\gamma$=0.60)

Table II shows the same quantities $N_{eq}$.ave, $N_{eq}$.max, $KLd_{ave}$ and sequentiality (Seq.) according to the gap cutoff $\delta$ values. The forcing value $\gamma$ is fixed to 0.60. Whatever the $\delta$-value, the specificity measured by $N_{eq}$.ave or $N_eq$.max is significantly lower (respectively around 2.1 and 6.3) relative to those previously obtained (see Table I, around 3.0 and 11.3). Moreover the structure informativity increases. In Table II, the reduction of the gap cutoff ($\delta$) implies a decrease of the proportion of fragments with gap (the adequacy score without gap $S_i$ is generally lower than the score $S_i^g$ with gap), hence a loss of site specificity ($KLd_{ave}$). In the opposite, the quantity Seq. measuring the degree of fragment sequentiality is more elevated. Those findings indicate that when the introduction of gap in the structural fragments is favored, the

sequentiality is diminished.

**"Training in Depth"**

| $S_0$ | $N_{eq}$. ave | $N_{eq}$. max | $KLd_{ave}$ | Seq. | gap (%) | selected frag. (%) |
|---|---|---|---|---|---|---|
| 0 | 2.63 | 7.9 | 1.65 | 68.3 | 7.5 | 99.9 |
| 5 | 2.19 | 6.6 | 1.86 | 69.9 | 25.0 | 94.2 |
| 10 | 1.26 | 4.1 | 2.32 | 72.8 | 40.2 | 44.5 |
| 15 | 1.09 | 2.0 | 2.42 | 82.2 | 48.4 | 18.2 |

TABLE 3: EFFECT OF THE FRAGMENT SELECTION ($\gamma$=0.80 AND $\delta$=0.75)

In this last study, the parameters $\delta$ (gap cutoff) and $\gamma$ (forcing factor) are fixed to 0.75 and 0.80 respectively. Table III shows the variations of the quantities according to the value of the adequacy score cutoff. Unsurprisingly, a high $S_0$-value (e.g. 15) by comparison with a reference value ($S_0 = 0$) induces a strong fragment selection (e.g. 18.2%), a higher site specificity ($N_{eq}$.ave = 1.09 instead of 2.63) and a higher site informativity ($KLd_{ave}$ = 2.42 instead of 1.65). Furthermore, we point out an increase of sequentiality explained by the fact that the selected fragments have a larger field for the training, thus facilitating the sequentiality and *a fortiori* the introduction of gaps (from 7.5% to 48.4% for $S_0 = 0$ and 15 respectively).

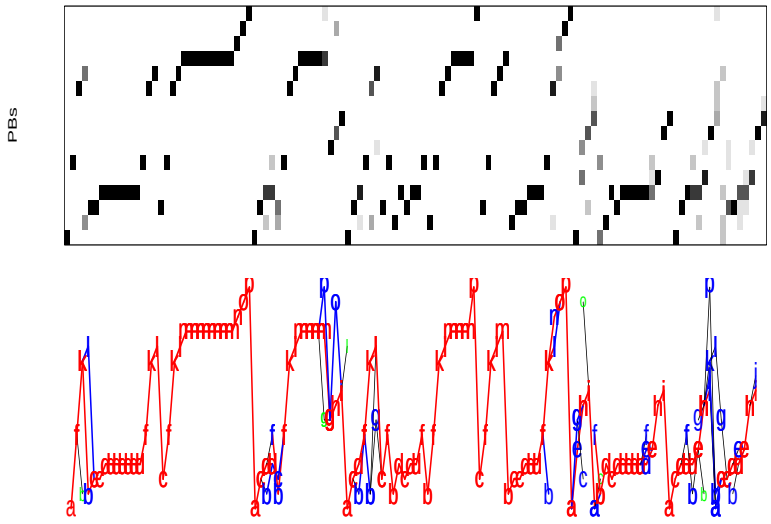


Figure 5: Hybrid Protein Model obtained after a "training in depth" procedure for $S_0$=0.10 and most frequent transitions observed along the HPM.

Figure 5a illustrates the effect of the "training in depth" procedure. It shows the HPM obtained for $S_0 = 10$. We have also drawn the most frequent transitions observed in the selected fragments (see Figure 5b). The structural prototypes of Figure 3 correspond to some of the highly frequent fragments present in this graph.

The preferential transitions are observed in the HPM regular secondary structure regions. They correspond to series of PB *m* or *d*. Different kinds of preferential transitions between these regular structural regions are also observed. As illustration, the PB series *fkl* (HPM sites [37;40]) enables the transition between a $\beta$-strand and an $\alpha$-helix, the PB series *nopa* (HPM sites [29;33]) characterizes a transition between an $\alpha$-helix and a $\beta$-strand, and the series *ehia* (HPM sites [102;105]) a transition between two $\beta$-strands.

## DISCUSSION AND CONCLUSION

The Hybrid Protein Model (HPM) presented in this paper uses the concept of self-organization as the conventional Self-Organizing Maps (SOM), but has the particularity of not using the concept of information diffusion. In our method, the input signal influences the winning neuron and its neighborhood thanks to the overlapping of the information, i.e. the consecutive neurons in the ring share a common structural information. This information sharing allows a continuity between the neurons along the ring.

The advantages of the HPM, specially for strings of characters (in particular for encoded 3D protein structures) are as follows :

(i) HPM allows one to carry out an unsupervised clustering with dependent structural classes, i.e. they share a common information. The conventional clustering methods such as k-means or SOM which tend to determine independent clusters are not well suited because the structural fragments are conditioned by their sequentiality.

(ii) The HPM is a tool composed of a set of PB distributions, allowing the calculation of an adequacy score between an observed fragment of the protein backbone and a list of structural prototypes. It is equivalent to a "profile method" used to characterize the amino acid variability along a multiple alignment of protein families.

(iii) The HPM carries out by a fast and efficient processing the compression of a local protein structure databank into 120 structural prototypes of 7 PBs. Consequently, the search of structural similarity between 3D protein structures becomes an easier task.

(iv) The introduction of gaps into the structural fragments leads to the realization of a multiple structural alignment with gaps. It allows firstly, to take the variability in length of the regular secondary structures into account and secondly, to describe the heterogeneity of some local structures. Consequently, the specificity of the HPM sites is improved.

(v) The "training in depth" procedure highlights the preferential transitions within the HPM.

The limitations of the HPM are :

(i) The determination of the optimal number of neurons has been tackled by different strategies, "baby learning" or "free space occupation". However, it is difficult to characterize an optimal size. The HPM size must represents a good compromise between representativeness of the structural motifs and low redundancy. The representativeness denotes that any fragment of the databank (i.e. an observed chain of L PBs) should find a structural prototype close to it among the N prototypes of the HPM. However, we must ensure a low redundancy to limit the number of fragments that can be clustered in different HPM structural classes.

(ii) The HPM is limited to a ring of neurons to keep the sequentiality property between consecutive fragments within proteins. However, the possible structural inputs or outputs of a given secondary structure ($\alpha$-helix or $\beta$-strand) are not unique. This implies a necessary redundancy in the HPM for describing these types of structures. A more complex network (instead of a ring) must be built to take the multiplicity of the structural pathways in the protein architectures into account.

This exploratory investigation (sensitivity study and improvements) and the analysis of the HPM performances, enable us to conclude that the Hybrid Protein Model constitutes a useful tool for compressing 3D protein structures into a library of overlapping structural prototypes. Further works will demonstrate the interest of such approach in the search of structural similarity and in structure prediction from protein sequence.

Protein local structural information is contained to an extent in local amino acid sequences. The perspectives of our work are (i) firstly to analyze, from the library of local structural prototypes, the relation between amino acid sequence and local structure, i.e. to characterize for each structural prototype the amino acid propensities and informativity, and (ii) secondly, to use the results of the local structure - sequence analysis in protein 3D structure prediction. Our aim is to propose local structural candidates for a given sequence window and then to reconstruct the global 3D structure from the predicted structural candidates. A first strategy developed uses statistical scores to assess the local compatibility between a sequence window and the structural prototypes of the library. Then, by using a dynamic programming approach, the extraction of the best local structural candidates was carried out. This first approach must now be improved.

## ACKNOWLEDGMENTS

This work was supported by a grant from the Ministère de la Recherche and from "Action Bioinformatique inter EPST" number 4B005F. AdB is supported by a grant from the Fondation de la Recherche Médicale.

## REFERENCES


[1] D. Baker and A. Sali, "Protein structure prediction and structural genomics," **Science**, vol. 294, pp. 93–96, 2001.

[2] F. Bernstein, T. Koetzle, G. Williams, E. Meyer, M. Brice, J. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The Protein Data Bank: a computer-based archival file for macromolecular structures," **J Mol Biol**, vol. 112, pp. 535–540, 1977.

[3] R. Bonneau, C. Strauss, C. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson and D. Baker, "De novo prediction of three-dimensional structures for major protein families," **J Mol Biol**, vol. 322, pp. 65–78, 2002.

[4] C. Bystroff and D. Baker, "Prediction of local structure in proteins using a library of sequence-structure motif," **J Mol Biol**, vol. 281, pp. 565–577, 1998.

[5] A. Camproux, P. Tuffery, J. Chevrolat, J. Boisvieux and S. Hazout, "Hidden Markov Model approach for identifying the modular framework of the protein backbone," **Protein Eng**, vol. 12, no. 12, pp. 1063–1073, 1999.

[6] A. de Brevern, A. Camproux, C. Etchebest, S. Hazout and P. Tuffery, "Beyond the secondary structures : the structural alphabets," **Recent Adv. In Prot. Eng.**, vol. 1, pp. 319–331, 2002.

[7] A. de Brevern, C. Etchebest and S. Hazout, "Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks," **Proteins**, vol. 41, no. 3, pp. 271–287, 2000.

[8] A. de Brevern and S. Hazout, "Compacting local protein folds with a Hybrid Protein," **Theoretical Chemistry Accounts**, vol. 106(1/2), pp. 36–47, 2001.

[9] A. de Brevern and S. Hazout, "'Hybrid Protein Model' for optimally defining 3D protein structure fragments," **Bioinformatics**, vol. 19, pp. 345–353, 2003.

[10] A. de Brevern, H. Valadie, H. S and C. Etchebest, "Extension of a local backbone description using a structural alphabet. A new approach to the sequence-structure relationship," **Protein Science**, vol. 11, pp. 2871–2886, 2002.

[11] J. Fetrow, M. Palumbo and G. Berg, "Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme," **Proteins**, vol. 27, pp. 249–271, 1997.

[12] S. Govindarajan, R. Recabarren and R. Goldstein, "Estimating the total number of protein folds," **Proteins**, vol. 35, pp. 408–414, 1999.

[13] W. Humphrey, A. Dalke and K. Schulten, "VMD - Visual Molecular Dynamics," **J Mol Graph**, vol. 14, pp. 33–38, 1996.

[14] T. Kohonen, **Self-Organizing Maps**, Berlin, Germany: Springer-Verlag, 1997.

[15] S. Kullback and R. Leibler, "On information and sufficiency," **Ann Math Stat**, vol. 22, pp. 79–86, 1951.

[16] T. Noguchi, H. Matsuda and Y. Akiyama, "PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB)," **Nucl Acids Res**, vol. 29, no. 1, pp. 219–220, 2001.

[17] G. Pollastri, D. Przybylski, B. Rost and P. Baldi, "Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles," **Proteins**, vol. 47, pp. 228–235, 2002.

[18] L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," **Proc. of the IEEE**, vol. 77, pp. 257–285, 1989.

[19] M. Rooman, J. Rodriguez and S. Wodak, "Automatic definition of reccurent

local structure motifs in proteins," **J Mol Biol**, vol. 213, pp. 327–336, 1990.

[20] J. Schuchhardt, G. Schneider, J. Reichelt, D. Schomburg and P. Wrede, "Local structural motifs of protein backbones are classified by self-organizing neural networks," **Protein Eng**, vol. 9, no. 10, pp. 833–842, 1996.

[21] K. T, "Self-organized formation of topologically correct feature maps," **Biol Cyber**, vol. 43, pp. 59–69, 1982.

[22] R. Unger, D. Harel, W. S and S. JL, "A 3D building blocks approach to analyzing and predicting structure of proteins," **Proteins**, vol. 5, pp. 355–373, 1989.

[23] D. Xu, O. Crawford, P. LoCascio and Y. Xu, "Application of PROSPECT in CASP4: characterizing protein structures with new folds," **Proteins**, vol. S5, pp. 140–148, 2001.