

“Hybrid protein model”: a new clustering approach for 3D local structures

C.Benros, S.Hazout, A.G. de Brevern

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

”Hybrid Protein Model”: a new clustering approach for 3D local structures.

C.Benros, S. Hazout and A.G. de Brevern

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

key words: fold library, protein block, unsupervised classifier, 3D protein structure, clustering.

abstract

The ”Hybrid Protein Model” (HPM) is a fuzzy model for compacting biological data series such as 3D structures into a limited number of overlapping cluster. The 3D structures of a non-redundant structural databank is encoded in a previously defined structural alphabet composed of 16 protein blocks (PBs) [3]. 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 learn 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. 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.

In a second step, to define the optimal length of the hybrid protein, we have defined an new approach based on the deletion of the redundant parts of the hybrid protein.

1 Introduction

The knowledge of the tridimensional (3D) structure of a protein remains one of the most important means of exploring its functional properties. In the context of intensive genome sequencing, the dramatic increase of the number of protein sequences (1D) identified is only partly accompanied by the resolution of the structures (3D). Hence, one faces the necessity to better understand the 1D-3D relationship between sequence and structure: This relation is the keystone of prediction methods that could provide an alternative way to overcome the gap between sequence and structure determination.

At the lower level, the 3D structure description is limited often to the succession of secondary structures (α -helix, β -sheet and coil). The α -helices and β -sheets are energetically interesting, but represent less than 50% of the protein folds. Hence, an emerging concept is the identification of a structural alphabet, i.e. of a limited number of recurrent structural elements of proteins, structural ”letters”, whose associations governed by logic rules form the words of protein structures. The definition of such an alphabet provides a new tool for better understanding protein architecture. In addition, the applications possible from such alphabet are numerous and range from simplifying the protein backbone conformation with a correct accuracy to more ambitious prediction approaches (for a review see [2]).

We have chosen to use 16 Protein Blocks (PBs) of 5 C α in length. This alphabet approximates the protein 3D-structures with reasonable accuracy, and we have already used it in a Bayesian prediction method of protein structure from the sequence [3]. Hence the protein structure description and classification are not easy tasks. The methodology ”Hybrid Protein Model” (HPM) is an attempt to tackle structural [4] or genomic data [7]. As proteins have common local structures of various lengths, we have tried to stack those structures locally. Hybrid protein model is an unsupervised classifier close to the Self-Organizing Maps (SOM [8]). In our approach, HPM consists in building a concatenation of local structures that share common parts. After training a non-redundant protein databank, every local structure of every protein is located in a given position of the hybrid protein.

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 3D databank, every local structure of every protein is located in a given position of the hybrid protein. We have evaluated the hybrid protein particularities in terms of its structures and its amino acid sequences. This approach allows similar local structures to be classified in a given site. One application of

the hybrid protein is the search for similar protein local structures in two cytochromes P450 which has given excellent results [5].

In a second step, we have improved our approach by defining an optimal hybrid protein by deleting of the redundant parts of the hybrid protein. Hence, the hybrid protein helps our understanding both in terms of its structures and its amino acid sequences.

2 Materials and method

2.1 Databank of 3D protein structures encoded into Protein Blocks

In a previous paper [3], we have established a structural alphabet for coding the 3D protein structures, i.e. a set of local prototypes, called Protein Blocks (PBs), able to approximate the protein backbone locally. The databank used in our study is composed of 717 non-redundant proteins taken from the Protein DataBank [1]. It was based on the PAPIA databank [9] selecting the chain with a resolution smaller or equal to 2 Å, an R-factor less than 0.2. Each structure selected exhibits a *rmsd* value larger than 10 Å from all the structures selected and a sequence identity smaller or equal to 30%. The whole 3D protein structures are encoded into PBs as illustrated in Figure 1. Hence, the databank is composed of 177 986 PBs.

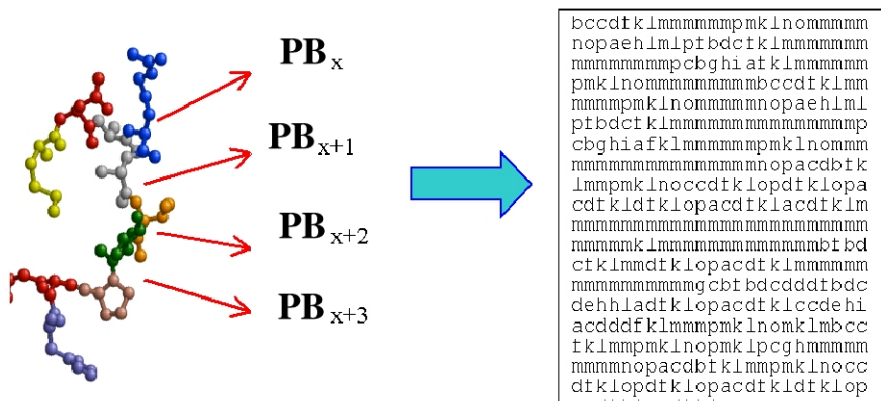


Figure 1: Coding of a structure into PBs. (left) A protein fragments translated into Protein Blocks PB_x to PB_{x+3} , (right) the whole protein is translated into Protein Blocks (see <http://condor.urbb.jussieu.fr/> for details).

2.2 "Hybrid Protein Model" (HPM)

The goal of "Hybrid Protein Model" (HPM) 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 ($n=1, 2, \dots, 16$). The principle of the training is summarized in Figure 2. Every 3D protein structure is cut into overlapping series of L PBs. In the example, the PB series is *mmmmmmnopafkl* (see Figure 2a). Each protein fragment is optimally located in a region of the hybrid protein (see Figure 2b) that is slightly modified, i.e. the PB distribution located in this region are slightly corrected (see Figure 2c).

In fact, the aim of the training is to improve progressively a score S used for clustering the protein fragments. So the hybrid protein allows the stacking and the characterization of a "prototype" for this subset of folds. For a local structure F drawn in the databank, we compute a score S_i 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 the block b_k in the fragment 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 the PB b_k , that observed in the databank.

The score S_i is the log odds score, i.e. the logarithm of the ratio of likelihoods between two hypotheses, the first one: the fragment F is defined by a series of PBs randomly ordered, and the second one : the fragment F is built according to the PB distributions of the hybrid protein.

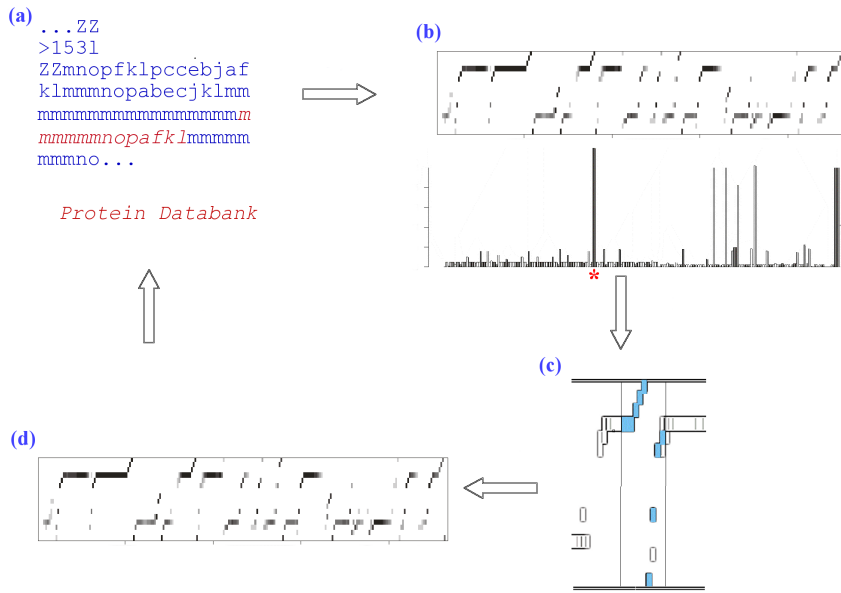


Figure 2: Principle of the "Hybrid Protein Model" (HPM) training. (a) A fragment of 13 PBs is taken randomly from the non-redundant databank, here *mmmmmmnopafkl*. (b) The fragment is presented to the hybrid protein, a score is then computed to find the best adequacy between the fragment and the hybrid protein region. The chosen position associated with the maximal score is noted by a star. (c) The sub-matrix around this position is slightly modified to learn this fragment, the frequency of the fragment PBs are increased (colored box), the others are decreased. (d) Another fragment is presented for the training (a).

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 slightly modified to increase the likeness of this part of the hybrid protein to the local structure \mathbf{F} . In position $i + k$, the value of the x^{th} PB $f_{i+k}(b_x)$ is changed into $[f_{i+k}(b_x) + \alpha]/[1 + \alpha]$. All the other PBs x' have their values decreases by $[f_{i+k}(b_{x'})]/[1 + \alpha]$. The learning coefficient α is a user-fixed value (e.g. $\alpha=5 \cdot 10^{-3}$) and may evolve during the iterations.

So this transformation allows one to increase the score of fragment \mathbf{F} . The training is progressive and thus needs to examine the entire local structure databank C times.

The continuity between the consecutive positions (i.e. contiguous fragment clusters) is insured. After training, every position i of the hybrid protein characterizes a cluster of fragments of length L structurally similar. This site maintains its continuity with the contiguous site $i-1$ since they have $L-1$ PBs distributions in common for the score computing.

2.3 Improvements for obtaining an optimal Hybrid Protein

An optimal hybrid protein can be characterized by two properties:

- (i) *a high continuity between consecutive hybrid positions*, i.e. when a fragment \mathbf{F} extracted from a given 3D protein structure at the beginning position p in the sequence is located in the position i_0 in the hybrid protein, the fragment \mathbf{F}' shifted of one residue in the sequence (in position $p + 1$) must be generally located in the position $(i_0 + 1)$ in the hybrid protein. So a 3D protein structure is represented by a limited number of hybrid protein regions.
- (ii) *low redundancy within the hybrid protein*. Two regions of length L are redundant in the hybrid protein when their L consecutive PB distributions are close, i.e. a fragment can be easily located in these two regions.

For reducing the redundancy, we compute during a cycle a confusion matrix $C(i, j)$ of dimension $N \times N$. A fragment \mathbf{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 one 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 the redundant regions of lengths more than an user-defined value l_0 . In fact, we define in the confusion matrix the whole diagonals of minimal lengths l_0 , i.e. $C(i, j), C(i + 1, j + 1), \dots, C(i + l_0 - 1, j + l_0 - 1)$ exceeding a given occurrence number n_{lim} . In our study, we have fixed l_0 and n_{lim} to 10 and 85 respectively.

Among the region pairs, we select this of maximal length and we delete one of these two candidate regions. Only one region of length more than l_0 is rejected from the hybrid protein by cycle. After a certain number of cycles, the reduction of hybrid protein length is stopped. We obtain an optimal hybrid protein, this being

3 Results

3.1 A first hybrid protein

Description of the hybrid protein. Figure 3 reports the results of the training after 15 learning cycles (i.e., C value) for a hybrid protein of length $N=100$. 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].

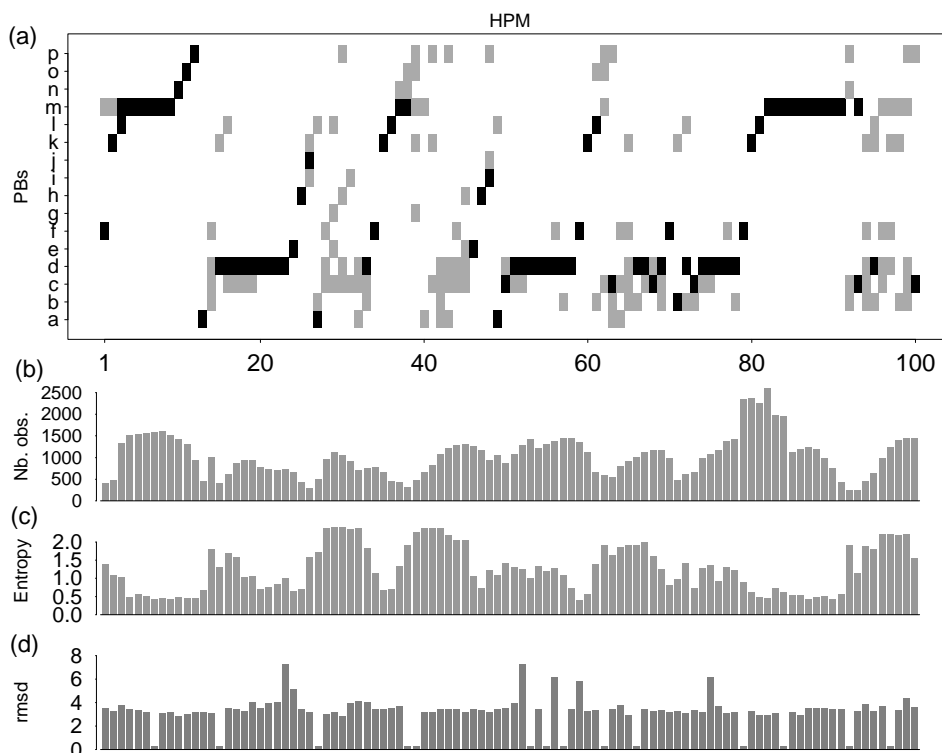


Figure 3: hybrid protein: (a) PBs distribution along the hybrid protein of length 100. (b) The number of protein fragments per site. (c) entropy distribution along the hybrid protein. (d) $rmsd$ values (computed for all the fragments of length 10 $C\alpha$ associated at each site).

Specificity of the protein hybrid sites. Only two or three separate PBs occur frequently at each position. Moreover, the continuity between consecutive positions are 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). 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).

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

3.2 Description of the optimal hybrid protein

Figure 4 reports the results of the training after 30 learning cycles (i.e., C value) of the optimal hybrid protein. Figure 4a 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: eight types of α -helices distinguishable by their sizes (4 to 20 PBs), as well as eight β -strands at least. All the transitions between regular secondary structures can be pointed out: 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].

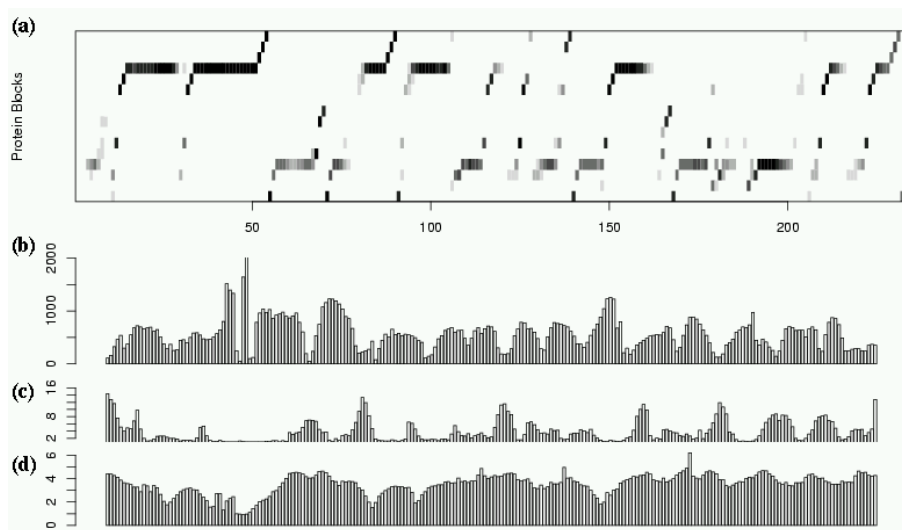


Figure 4: Optimal hybrid protein: (a) PBs distribution along the hybrid protein of final length 233. (b) The number of protein fragments per site. (c) Average number of PB, called N_{eq} , along the hybrid protein. (d) $rmsd$ values (computed for all the fragments of length 13 $C\alpha$ associated at each site).

Also we detect different motifs located at the beginning or at the end of regular secondary structures such as *flk* and *nop* for α -helices, *ac* and *ehia* for β -strands. Figure 4b shows the number of fragments along the hybrid protein. The distribution is almost uniform, the weak sizes correspond to the coils. Figure 4c gives the variation of the "equivalent number of PBs" N_{eq} index along the hybrid protein. A majority of positions is highly specific express only one PB. The sites with many different correspond to either a transition region such as turns between two strands positions [135:142], coils between two α -helices positions [29:34], and long coils positions [398:15], or distorted secondary structures such as β -strands positions [177:192]. After 3D superimpositions of the protein fragments of a cluster (associated to a given position), we have computed the $rmsd$. Figure 4d 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 of 13 $C\alpha$ length. The range is [0.92Å; 6.20Å] which is quite good compared to other classification [10]. The minimal value corresponds to a long regular α -helix located in position 48, the maximal one to a variable coil in position 183. Globally, the local structures showing low $rmsd$ (less than 2Å) are the α -helices or transitions between β -sheet and α -helix, or between two α -helices.

4 Discussion and conclusion

HPM is a method able to build a repertory of clusters of local protein folds with a fixed length. The training is carried out by using a PB series in order to obtain a continuity between the different clusters. As seen previously, a correct structural variability ($rmsd = 3$ Å) and high sequence informativity are expressed along the hybrid protein. Moreover, different protein domains are visible such as the supersecondary structures α -turn- α , β -turn- β or β - α - β .

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 have illustrated by an example of two cytochromes the advantage of using the hybrid protein for extracting similar local folds present in these proteins [5]. The success of the methodology HPM in fold recognition must be validated in a further work. Consistently, by using the sequence informativity found in the amino acid occurrences matrices associated with the different fold clusters, we should be able to pick out candidates for folding simulations from the repertory.

We have assessed the advantages of the procedures introduced the hybrid protein size reduction by deletion of the redundancy. The control of the length reduction is insured by the parameters l_0 and n_{lim} . The parameter l_0 should be close to the fragment size L or slightly lower (in our study $l_0 = 10$). A higher value is avoided since no size reduction is significantly observed. A lower one leads to a hybrid protein cut up into small pieces. The other parameter n_{lim} , i.e. the minimum number of occurrences between two redundant regions, is essential for the control of the hybrid protein size. A value of 85 was chosen to enable a repertory to be characterized with a fold distribution almost uniform. With a n_{lim} -value lower ($=80$), the size of the hybrid protein decreases from 233 to 175, hence, a fragmentation is pointed out. Conversely with a higher one ($=90$), it increases to 325. In this case, the redundancy is weakly deleted. Indeed, the optimality for the hybrid protein is a balance between the deletion of the region redundancy and the conservation of the fold continuity [6].

In conclusion, we have shown how, through a stacking procedure of local folds, we have built an optimal repertory sufficiently rich to be used in further works of molecular modeling.

Acknowledgments

This work was supported by a grant from the 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 Fondation pour la Recherche Médicale.

References

- [1] 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*, 112:535–540, 1977.
- [2] A.G. de Brevern, A. Camproux, C. Etchebest, S. Hazout, and P. Tuffery. Beyond the secondary structures : the structural alphabets. in *Recent Adv. In Prot. Eng.*, in press, 2002.
- [3] A.G. de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks. *Proteins*, 41(3):271–287, 2000.
- [4] A.G. de Brevern and S. Hazout. Hybrid protein model (HPM): a method to compact protein 3D-structures information and physicochemical properties. *IEEE - Computer Society : Proceedings of the 7th Symposium on String Processing and Information Retrieval*, 1:49–57, 2000.
- [5] A.G. de Brevern and S. Hazout. Compacting local protein folds with a hybrid protein. *Theoretical Chemistry Accounts*, 106(1/2):36–47, 2001.
- [6] A.G. de Brevern and S. Hazout. Improvement of "Hybrid Protein Model" to define an optimal repertory of contiguous 3D protein structure fragments. *Bioinformatics*, in press, 2002.
- [7] A.G. de Brevern, F. Loirat, A. Badel-Chagnon, C. Andre, P. Vincens, and H. S. Genome compartmentation by a hybrid chromosome model ($h\chi m$). application to *saccharomyces cerevisiae* subtelomeres. *Computers and Chemistry*, in press, 2002.
- [8] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, Germany, 1997.
- [9] T. Noguchi, H. Matsuda, and Y. Akiyama. Pdb-reprdb: a database of representative protein chains from the protein data bank (PDB). *Nucl Acids Res*, 29(1):219–220, 2001.
- [10] J. Wojcik, J.-P. Mornon, and J. Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289:1469–90, 1999.