Accepted for publication in J Theo Biol

Analyzing the Sequence-Structure Relationship of a Library of Local Structural Prototypes.

Cristina Benros[1], Alexandre G. de Brevern[1,2,*] and Serge Hazout[1,+]

[1]Equipe de Bioinformatique Génomique et Moléculaire (EBGM),
INSERM UMR-S726, Université Denis Diderot - Paris 7, case 7113,
2, place Jussieu, 75251 Paris, France
[2] Institut National de Transfusion Sanguine (INTS),
6, rue Alexandre Cabanel, 75739 Paris cedex 15, France

* Corresponding author:
Mailing address: Alexandre G. de Brevern, Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM UMR-S726, Université Denis Diderot - Paris 7, Institut National de Transfusion Sanguine, 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail: alexandre.debrevern@univ-paris-diderot.fr
Tel: +33(1) 44 49 30 00
Fax: +33(1) 47 34 74 31

Running title: Sequence-Structure Relationship

Keywords: protein local structure, structural alphabet, structural motif library design, sequence-structure correlation, amino acid equivalence classes.

[+] deceased

Abstract

We present a thorough analysis of the relation between amino acid sequence and local three-dimensional structure in proteins. A library of overlapping local structural prototypes was built using an unsupervised clustering approach called "Hybrid Protein Model" (HPM). The HPM carries out a multiple structural alignment of local folds from a non-redundant protein structure databank encoded into a structural alphabet composed of 16 Protein Blocks (PBs). Following previous research focusing on the HPM protocol, we have considered gaps in the local structure prototype. This methodology allow the have variable length fragments. Hence, 120 local structure prototypes were obtained. 25% of the protein fragments learnt by HPM had gaps.

An investigation of tight turns suggested that they are mainly derived from three PB series with precise locations in the HPM. The amino acid information content of the whole conformational classes was tackled by multivariate methods, *e.g.*, canonical correlation analysis. It points out the presence of seven amino acid equivalence classes showing high propensities for preferential local structures. In the same way, definition of “contrast factors” based on sequence-structure properties underline the specificity of certain structural prototypes, *e.g.*, the dependence of Gly or Asn-rich turns to a limited number of PBs, or, the opposition between Pro-rich coils to those enriched in Ser, Thr, Asn and Glu. These results are so useful to analyze the sequence – structure relationships, but could also be used to improve fragment-based method for protein structure prediction from sequence.

1. Introduction

Knowledge of protein three-dimensional (3D) structures of contributes to understand their biological functions (Baker and Sali, 2001). Predicting protein structures from amino acid sequences constitutes a major scientific challenge when both X-ray crystallography and nuclear magnetic resonance analysis are difficult to undertake. Prediction methods are based on the fact that the amino acid sequence of a protein specifies its 3D structure. Different approaches are currently used to predict global protein structures: (i) comparative modeling is carried out when the target sequence shares a good sequence similarity with proteins of known 3D structures (Sali and Blundell, 1993), (ii) threading approach is used when the template structure is hard to find due to a low sequence similarity. It searches for the best compatibility between the target sequence and known protein folds (Xu et al., 2001), (iii) *ab initio* methods are attempted for proteins without enough sequence similarity to any protein whose structure is available while *de novo* methods combine them all (Bonneau et al., 2002).

To circumvent the complex problems of global protein structure prediction, several research groups have focused on local structure prediction. To this end, they have developed fragment-based approaches (Bystroff and Baker, 1998; Haspel et al., 2003). These approaches rely on the hypothesis that the protein folding can be represented as a hierarchical process, which initiates locally (Lesk and Rose, 1981). This hierarchical concept implies that protein local structural information is largely contained in local amino acid sequences, independent on long-range interactions. Then, the global structure can be modeled by combination of local fragments with different refinements (Inbar et al., 2003).

Classical secondary structure description leaves 45% of protein structure not described, *i.e.*, coil regions. In an attempt to describe the local structural characteristics in a more accurate and comprehensive way, many research groups have designed local structural alphabets, *e.g.*, (Sander et al., 2006; Unger et al., 1989). They correspond to sets of structural

prototypes able to approximate protein 3D structures (Offmann et al., 2007).

We have developed a such structural alphabet (de Brevern, 2005; de Brevern et al., 2000); it reduces the protein 3D complexity into one-dimensional string of characters. Based on this description, the local structure descriptions was extended to longer fragments, using an unsupervised clustering method called "Hybrid Protein Model" (HPM, (de Brevern and Hazout, 2001; de Brevern and Hazout, 2003)). The HPM topology is a neural network represented by a ring of neurons. Each neuron corresponds to a conformational class of the library grouping structurally similar fragments defined by our structural alphabet. The HPM procedure builds structurally dependent consecutive classes because the successive neurons of the ring share common overlapping information. The HPM shares with Self-Organizing Maps (Kohonen, 2001) the concept of self-organization to carry out the classification of the data. But for the HPM, the information diffusion is implicitly operated thanks to the overlapping between the consecutive neurons. The HPM represents a "structural profile" (Gribskov et al., 1987), resulting from the multiple local alignment of the PB encoded structural fragments. Related approaches are summarized in (Gonzalez-Diaz et al., 2008), and examples can be found in (Gonzalez-Diaz et al., 2007; Munteanu et al., 2008).

In the present study, we have defined a new HPM by allowing gaps in the protein structural fragments during the training. This procedure yields to increase the library specificity in terms of sequence-structure relationship. In addition, we carried out a characterization of the most frequent secondary structures found in the loops, *i.e.*, the tight turns, in terms of location in the HPM and PB signatures.

In the second step, we extracted information on protein sequence – structure relationship from the library of conformational classes. Different analyses were performed by using multivariate methods such as hierarchical clustering and canonical correlation analysis (Hotelling, 1936): (i) to identify of sequence preferences for each structural class, (ii) to

determine amino acid equivalence classes, so highlighting the global sequence – structure dependence and, (iii) to extract "contrast factors" opposing certain structural prototype subsets according to their sequence – local structure dependence.

2. Materials and Methods

2.1- Non-redundant databanks of 3D protein structures

Two databanks were used in our study, *set 1* including 675 non-redundant protein structures extracted from the Protein Data Bank (Berman et al., 2000), and *set 2* ,a recent set, including 1143 non-redundant protein structures. These structures, selected from the PDB-REPRDB database (Noguchi and Akiyama, 2003), had 2 Å or better resolution. They presented less than 30% of sequence identity and a minimal root mean square deviation (*rmsd*) of 10 Å after pairwise superimposition. Each structure of the databanks was encoded into our structural alphabet. This latter is composed of a set of 16 structural prototypes called Protein Blocks (PBs, see Figure 1) able to approximate locally protein 3D structures. The 16 PBs can approximate locally the protein backbone with a high precision (*rmsd* < 0.42 Å) (de Brevern, 2005).They are labeled by letters from *a* to *p*. Each PB of 5 consecutive residues is defined by 8 dihedral angles (ϕ and ψ). Hence, protein 3D structures are encoded in a string of characters (*i.e.*, the PBs), the coding principle being based on root mean square deviation on angular values (Schuchhardt et al., 1996). Each protein 3D structure of the non-redundant databank, a string of PBs, was cut into overlapping fragments of L successive PBs. In our study, $L = 2w+1$, with $w = 3$, *i.e.*, fragments of 7 PBs. A fragment of $L+4$ amino acids is associated to each structural fragment and is defined from $-w-2$ to $+w+2$ (in our study, $L+4 = 11$).

2.2- A strategy for building a library of structural prototypes: "Hybrid Protein Model"

The "Hybrid Protein Model" (HPM) is an unsupervised clustering method. It performs the compression of the structural information by establishing a library of overlapping structural prototypes (de Brevern and Hazout, 2000). The library of structural motifs is established during a training phase. The training was carried out with three quarters of the *set 1* (105,340 fragments). The validation step was performed on the left quarter of *set 1* (34,163 fragments) and its stability was confirmed with *set 2* (277,618 fragments).

2.2.1- Training of the Hybrid Protein Model

The principle of the training is described in Figure 2. The HPM is a ring of neurons, *i.e.*, a closed linear neural network with connected extremities (Fig. 2a). The HPM can be considered as a probabilistic protein composed of N sites. Each site s is defined by a law of probability F_s(PB), corresponding to the distribution of the B PBs composing the structural alphabet (in our study, B = 16). The HPM corresponds to a matrix of dimension N x B (Fig. 2b). A structural class, *i.e.*, a neuron clusters fragments sharing similar local structure. It is defined by L successive laws of probability (in our study, L = 7). Two successive structural classes are overlapping since they share (L-1) sites. Thus, the successive neurons are dependent and share common information. The training includes two distinct steps: an identification phase and a local enrichment phase.

(i) <u>*The identification phase*</u>: Each structural fragment of L successive PBs is taken randomly in the training set. The most similar pattern present in the HPM is searched (Fig. 2c). This score is a logarithm of likelihood ratio, and, measures the adequacy of a structural fragment with a given neuron (*i.e.*, L successive laws of probability). The expression of the score $Sc(s)$ at a given HPM site s (s varying from 1 to N) is:

$$Sc(s) = \sum_{X=-w}^{X=+w} \ln\left[\frac{F_{s+X}(PB_x)}{F_R(PB_x)}\right] \qquad (1)$$

PB_x corresponds to the Protein Block located in position X in the structural fragment, X varying from -w to +w. $F_{s+X}(PB_x)$ is the frequency of PB_x in position (s+X) in the HPM. $F_R(PB_x)$ is the reference frequency of PB_x, *i.e.*, its observed frequency in the databank. The identification phase consists in selecting the position s_{opt}, the position with the maximum score $S_{\max} = \arg\max\left[Sc(s)\right]$. No other position has a better adequacy for the presented fragment.

(ii) <u>*The local enrichment phase*</u>: This phase consists in slightly modifying the PB distributions of the structural class corresponding to the site s_{opt}, in order to increase the likeness between the PBs distribution of site s_{opt} and the presented fragment (Fig. 1d). The enrichment procedure is applied to the L PB distributions for the sites from (s_{opt}-w) to (s_{opt}+w). It is carried out as follows:

$$F_{sopt+x}(PB) \leftarrow \frac{F_{sopt+X}(PB) + \alpha}{1+\alpha} \quad \text{if} \quad PB = PB_x \qquad (2)$$

$$F_{sopt+X}(PB) \leftarrow \frac{F_{sopt+X}(PB)}{1+\alpha} \quad \text{elsewhere (the 15 other PBs).}$$

at position (s_{opt}+x), x varying from -w to +w. The symbol ← specifies "is changed into". These equations allow one to increase the frequencies of the fragment protein block PB_x in the sites of the HPM located from s_{opt}-w to s_{opt}+w, and to reduce the frequencies of the other PBs (*a fortiori* not observed). Moreover, this transformation ensures to keep the frequency values within the range [0, +1] and the sum of frequencies equal to 1.

The parameter α is the training coefficient. Initially fixed at a value α_0 (*e.g.* 0.20), it decreases during the training according to the following equation:

$$\alpha = \frac{\alpha_0}{1 + t/T} \quad (3)$$

where t denotes the number of fragments already presented to the HPM and T the total number of fragments in the training databank. As the training is iterative, several cycles are necessary to obtain a stabilization of the PB distribution laws. Different parameters implicated in the quality of the training are developed in supplementary data 1.

2.2.2- Improvements of the HPM training

Gaps were introduced into strings of PB sequences describing structural fragments. This strategy leads to the definition of a multiple structural alignment with gaps. So, it permits to take into account variable length fragment instead of only fixed length (L). The rules added in the training are expressed as: (i) gaps are introduced at one position of the structural fragment, and their length can vary from 1 to 5, and (ii) a cutoff G is fixed to select only the significant gaps. The gap is considered to be significant when the optimal score of the fragment with gap is G (=4) times higher than the maximal score obtained for the fragment without gap. To check the relevance of the introduction of gaps in the structural fragments, we assessed the variations of the PB and sequence information contents between the Hybrid Protein Models built with or without gaps. To insure a continuity of the successive local prototypes, and also diminish the influence of weakly represented structural fragments, dedicated parameters have been also optimized (see supplementary data 2).

2.3- Characterization of the Hybrid Protein Model in terms of local structure

The PBs specificity at each HPM site s was quantified by an entropy-derived measure *Neq*, *i.e.*, "equivalent number of PBs". Its expression is the exponential of Shannon entropy $H(s)$ (Shannon, 1948).

$$Neq(s) = \exp[H(s)] \quad \text{with} \quad H(s) = -\sum_{x=1}^{x=B} F_s(x).\ln F_s(x) \quad (4)$$

where $F_s(x)$ denotes the frequency of the Protein Block x located in position s of the HPM (B=16 for our structural alphabet). The *Neq* value measures the average number of of PBs per site. It varies between 1 PB (*i.e.*, only one PB is observed) and 16 PBs (*i.e.*, the 16 PBs are evenly distributed). Therefore, a low *Neq* value means a highly determined site.

2.4- Global analysis of the sequence-structure relationship

2.4.1- Amino acid propensities along the HPM sites

We analyzed the distribution of the amino acids located in the central position of the fragments to point out amino acid propensities along the HPM sites. We computed an amino acid occurrence matrix of dimension N x 20 (with N = 120), which gives for each structural class the amino acid distribution of the central residue. These occurrences were then normalized into Z-scores as follows:

$$Z_s(j) = \frac{n_s(j) - f_R(j).N_s}{\sqrt{f_R(j).N_s}} \quad (5)$$

For an amino acid j in HPM position s, $n_s(j)$ denotes its observed occurrence, $f_R(j)$ its reference frequency (*i.e.*, the frequency observed in the databank), and N_S the number of fragments associated to the site s. Thus, the term $f_R(j).N_s$ corresponds to the expected occurrence of the amino acid j at the site s. To measure the sequence informativity of a HPM site s, we determined which amino acids were over- and under-represented. A threshold was fixed at 2.57 (absolute value for a risk of 1%). Thus, Z-scores greater than 2.57 (respectively lower than –2.57) correspond to amino acid over-represented (respectively under-represented). Implicitly, we assume that the number of occurrences follows a Poissonian law that can be approximated by a normal distribution.

2.4.1-Amino acid informativity along the HPM sites:

The Kullback-Leibler asymmetric divergence measure (*KLd*), or relative entropy, makes it possible to analyze the amino acid informativity of the central residue in each structural class (Kullback and Leibler, 1951). The *KLd* values were computed along the *N* sites:

$$KLd(s) = \sum_{j=1}^{j=20} f_s(j).\ln\left[\frac{f_s(j)}{f_R(j)}\right] \qquad (6)$$

$f_s(j)$ denotes the frequency of the amino acid *j* for a given HPM site *s*, and $f_R(j)$ is the reference frequency of the amino acid *j* in the databank. The relative entropy *KLd*(*s*) measures the divergence between the amino acid distributions f_s and f_R. The higher the relative entropy is, the more informative the site is, in terms of amino acid.

2.5- Extraction of sequence - structure dependencies common to prototype subsets

The purpose of canonical correlation is to extract series of pairs of correlated factors (*i.e.*, linear combination of variables) between two sets of variables. A linear combination of a set of original variables is called a canonical variable. The objective of the analysis is to find (if it exists) the relationship between distribution of amino acid and prototypes. In our study, the variable sets are those describing on the one hand, the local structure (Z-scores of the 16 PBs: $Z'_S(b)$) and on the other hand, the amino acid sequence (Z-scores of the 20 amino acids: $Z_S(a)$) associated to the *N* structural classes of the HPM. The canonical correlation analysis aims at finding the maximal correlation between a linear combination of the first set of variables and a linear combination of the second set of variables. The canonical scores are the values of the two canonical variables for a given HPM site *s*.

$$F_k(s) = \sum_{a=1}^{a=20} \alpha_k(a).Z_s(a) \quad \text{and} \quad G_k(s) = \sum_{b=1}^{b=16} \beta_k(b).Z'_s(b) \qquad (7)$$

where $F_k(s)$ and $G_k(s)$ correspond to the k-th pair of correlated linear combinations in the site s for the amino acids and the PB series, respectively.

In the canonical correlation analysis, the coefficients $\alpha_k(a)$ and $\beta_k(b)$ are determined by setting a maximal correlation between F_k and G_k, and an independence with the factors previously defined (a rank less than k). The correlation between $F_k(s)$ and $G_k(s)$ is the k-th canonical correlation, *i.e.*,$R_k = \mathrm{corr}(F_k, G_k)$. The percent of explained variance is given by the value of the canonical correlation squared, R_k^2.

A canonical factor loading is the correlation of a canonical variable with an original variable, *i.e.*, $\mathrm{corr}(F_k(s), Z_s(a))$ or $\mathrm{corr}(G_k(s), Z'_s(b))$. We computed the canonical factor loadings in order to interpret the meaning of the canonical variables relative to the original variables. For each structural classes, the loadings specify the significance of each PB and each amino acid kind. They are significant if they are greater than an empirical cutoff fixed at 0.3. We computed also the canonical communality coefficient which corresponds to the sum of the squared canonical factor loadings for a given variable. It measures how much of a given original variable's variance is reproducible from the canonical variables. The maximum number of canonical correlations between two sets of variables is limited to the number of variables for the smallest set, *i.e.,*16 here. R package was used (Ihaka and Gentleman, 1996).

3. Results

3.1- Library of overlapping structural motifs

The library characterized by the final HPM is composed of 120 overlapping structural classes. Each class includes fragments sharing similar local structures encoded into PBs (fragments of 7 successive PBs corresponding to 11 amino acids). These classes are representative of the protein local folds observed in the databank. 94.2% of the fragments

from the training set were involved in the HPM training, *i.e.*, 99,214 structural fragments. The remaining was discarded due to their low log odds scores (see supplementary data 2). They represent the less frequent series of PBs.

Analysis of the final HPM enables to locate regions of regular secondary structure and their preferential transitions. Figure 3a shows the matrix of PB distributions with gray levels proportional to PB frequencies. The Hybrid Protein Model described here allows the characterization of structural prototypes representative of the local protein fragments observed in the databank. Each structural class of the library is composed of structurally similar fragments. So each class can be represented by the protein fragment structure, defined as structural prototype, the closest to the average local fold. Figure 4 (and supplementary data 3) shows one third of the structural prototypes of the library (Humphrey et al., 1996). The overlap insures a structural continuity rate between the successive structural classes, it reaches 70%. Moreover, the *rmsd* values computed along the HPM for each prototype indicate a high structural stability. The average *rmsd* value of 1.94 Å is rather small for fragments of 11 amino acids. The clusters obtained are globally homogeneous since the *rmsd* values vary within the range [0.38Å; 3.31Å]. The lowest value is obtained for the site #26 associated to a long α-helix and the maximal value for the site #52 corresponding to a region weakly structured.

The analysis of the HPM according to the three states secondary structure assigned by STRIDE (Frishman and Argos, 1995), shows that the PBs *m* and *d* approximate with accuracy α-helices and β-strands respectively, and that the variability of the coil state is globally taken into account by the other PBs. Three types of α-helix characterized by series of PB *m* and variable in length (from 4 to 9 PBs) are distinguishable. They are located in the regions [21-29], [41-45], [67-70] and are labeled α1, α2 and α3, respectively. Different types of β-strand (from 3 to 7 PBs long) characterized by series of PB *d* are distinguishable with a strong PB

signature. Some of the β-strands are well delimited, *i.e.*, those located in the regions [7-13], [35-37], [94-100], while others are located in fuzzier regions, *e.g.*, [58-61], [80-82], [107-109] and [116-117]. The different β-strands are labeled from β1 to β7 along the HPM. The introduction of PB *c* in series of PB *d* leads to distorted β-strands. Preferential transitions between these regular structural regions are observed. As an illustration, the PB series *fkl* (region [37-40]) defines a transition between a β-strand and an α-helix, the PB series *nopa* (sites [29-33]) characterizes a loop linking an α-helix to a β-strand, and the series *ehia* (sites [102-105]) a transition between two β-strands. Some regions of the HPM are fuzzier, *e.g.*, [47-56], [89-92], [111-120]. They correspond to what is generally considered as coil regions.

3.2- Introducing gaps improves the structural prototype library

Figure 3b shows the number of fragments associated to the different HPM structural classes. The structural fragments are evenly distributed along the HPM except for regular secondary structure regions (black bars in Fig. 3b). The α-helix region extending between sites 21 and 29 contains a high number of structural fragments mainly centered at site #24 (13,790 fragments). The lowest numbers of fragments are associated to coil regions, with site the least populated (86 fragments) for the site #16. Moreover, the values of the adequacy scores, measuring the divergence between the probability of observing a given fragment in a HPM region, and, the probability of observing this fragment in the databank, are quite significant. They thus testify of the quality of the training, and *a fortiori* certify the correct representativeness of our prototype library.

To analyze the relevance of introducing gaps in some structural fragments, we compared this HPM with an HPM for which the training was carried out without introduction of gaps (103,542 fragments, *i.e.*, 98.3% of set 1 was used).

The entropy-derived diversity measure *Neq* (Etchebest et al., 2005; Hazout, 2007), *i.e.*, the

equivalent number of PBs, permits to characterize the structural specificity of each HPM site. Figure 3c shows the improvement of this specificity along the HPM sites. The average *Neq* value for the HPM without gaps is equal to 3.06 compared to 2.19 for the HPM with gaps. The maximal value is strikingly decreased, *i.e.*, 6.6 for the HPM with gaps compared to a value twice higher for the no gap HPM. In some HPM coil regions, the *Neq* gain is high such as the zones [44-54] and [79-89], for which the mean *Neq* value drops from 5.0 to 2.9 and from 4.9 to 2.5, respectively. In the final HPM with gaps, 90% (respectively 58%) of the sites have a *Neq* value less than 4 (respectively 2) equivalent PBs. These sites highly specific are mainly located in regions of α-helix and β-strand. The *Neq* values increase for the less determined HPM regions, this is the case of certain coil regions, *e.g.*, the maximum value, 6.6, is observed at the site #114.

The introduction of gaps concerned 25% of the fragments used for the training (*i.e.*, 24,837 fragments, white bars in Fig. 3b). Thus, a quarter of the structural fragments used the possibility to adapt their length. As an illustration, Figure 5 shows fragments with and without gaps associated to the same site #27. The α-helix region ranging between sites 21 to 29 is essentially characterized at the C-terminal extremity by the PB series *nopa*. Introducing gaps enables to point out another type of α-helix C-cap encoded by less frequent PB series *pcc*. In this site, 13.9% of the structural fragments have gaps. Interestingly, a combination of these two different C-caps can lead to a longer loop, *e.g.*, *nopcc*. Hence, long loops may be combinations of short PB series. The proportions of fragments with gaps of length 1, 2, 3, 4 and 5 are 17.7%, 15.3%, 16.6%, 23.7% and 26.7%, respectively. The presence of a gap in a fragment can be explained by (i) a short repetitive secondary structures (see Fig. 5), (ii) an extremity of a regular structure less frequent, as previously described, or (iii) a high variable structure. We can distinguish along the HPM regions with less or more gaps (white bars in Fig. 3b). The site #24, for instance, has only 0.3% of fragments with gaps because they are

included in a long regular α-helix (*i.e.*, PB series with more than 7 PB *m*), whereas the site #26 have 95.8% of fragments with gaps, as most of them being α-helices showing a large variation in length but characterized by the same C-cap.

3.3. Characterization of specific structural motifs: turns and prototypes

3.3.1- Eight major HPM regions corresponding to turns are highlighted

Turns are known to induce a polypeptide chain reversal. We considered that a fragment of $L = 7$ PBs associated to if this latter was centered in the fragment. We focused our attention on structural prototypes that can be associated to γ-turns, β-turns and α-turns (Bornot and de Brevern, 2006; Fuchs et al., 2007; Hutchinson and Thornton, 1996; Pavone et al., 1996; Rose et al., 1985). Table I provides the three main locations of the motifs of interest in the HPM. The different turn types are grouped in 8 main zones of the HPM. They are labeled from T1 to T8 with sub-regions specified by letters from *a* to *g* (see Table I and caption). They correspond to 50 sites of the HPM. These varying in length HPM zones are well defined, with a mean *Neq* value varying within the range [1.2; 5.3]. These turn zones from T1 to T8 correspond to the HPM sites [29-33], [38-41], [45-55], [63-67], [83-91], [101-108], [110-114] and [117-5] (as HPM is closed), respectively. They characterize with gradual specificity transitions between regular secondary structures (α-helix and β-strand). T3 corresponds to the largest turn zone. It appears from this analysis that only two HPM regions linking regular secondary structures are not associated to turns, *i.e.*, [14-20] and [71-77]. These two regions are poorly populated.

3.3.2- Most of the different turn types have accurate locations in the Hybrid Protein Model

This analysis shows clearly that the turns are globally well delineated in the HPM. For each type of turn, more than 50% are found in only three HPM zones. As an illustration, 49.8% of

the β type I turns, 66.3% of the β type II turns, 95.5% of the α II-LU turns are identified when considering only their three most populated HPM zones. Only the inverse γ-turns and the β type VIII turns have not precise locations and constitute a miscellaneous category compared to the β type IV turns. These results are interesting. Indeed, the HPM was not constructed to discriminate the different turn types, but to learn a complete protein structure databank. Nevertheless, the HPM enables to analyze turns in their environment. Furthermore, different turn types present similar PB signatures, and hence are found in the same HPM regions. Turns share a few number of distinct signatures (see supplementary data 4), with mainly three common patterns (*nopa*, *ehia* and *fklm*). Most of the other motifs mainly derived from these patterns by deletion, insertion or substitution of PBs. In the same way, HPM underlines a possible formation of multiple turns involving β- and α turns.

3.4.- Global analysis of the sequence - structure relationship

We analyzed the amino acid informativity of the central residue for each HPM structural class thanks to the Kullback-Leibler asymmetric divergence measure *KLd* (Kullback and Leibler, 1951). The higher this relative entropy is, the more informative is the site in terms of amino acids. The sequence informativity is high along the HPM with *KLd* values varying within the range [0.03; 1.23]. HPM regions with high sequence information content were highlighted, *e.g.*, [11-22] corresponding to the C-terminal extremity of a β-strand and a transition towards an α-helix (the *KLd* sum over this region is equal to 2.71). The five highest *KLd* values are associated with Glycine and Asparagine rich prototypes (#32, #48, #87, #104 and #120; see Fig. 6). The average *KLd* value for these five prototypes is equal to 0.95. They are all located in HPM turn regions. This result is related to the propensity of Gly and Asn for belonging to reverse turns (Fuchs and Alix, 2005). Other particularly informative sites were pointed out, *e.g.*, #64, #91 and #112. They are also characterized by Gly and Asn but to a

lesser extent and are located in turn regions.

Moreover, we assessed the variation of the *KLd* values for the HPM built with or without introduction of gaps. This analysis shows a light increase of the sequence informativity along the HPM when gaps are present. The sum of *KLd* values along the no gap HPM sites is equal to 19.16 versus 21.46 for the gap HPM.

Hierarchical clusterings have been done on the amino acid distribution and in parallel on the distribution of PBs. Concerning the amino acids, seven clusters have been identified. The two first groups isolated amino acids that play particular roles, namely Glycine and Proline. The third and fifth groups were composed principally of polar amino acids whereas the fourth, sixth and seventh groups include mainly hydrophobic amino acids. The second analysis corresponds to a hierarchical clustering of the 120 structural prototypes according to their amino acid central residue distributions. Four clusters have been identified (see supplementary data 5). The first one corresponds to 37 structural classes, β-strand local structures, *i.e*. located in HPM regions from β1 to β7. The second cluster gathers 38 structural classes in majority associated with HPM α helical regions, *i.e*., from α1 to α3, but also to their edges, *i.e*., T1. The third cluster is composed of 40 structural classes primarily associated with HPM loop and turn areas. Finally, the fourth cluster isolates 5 structural classes primarily characterized by the presence of Glycine and Asparagine. They are the most informative ones with regard to their sequence information content, as highlighted previously by the *KLd* analysis.

3.5- Extraction of sequence - structure dependence common to prototype subsets

Figure 7 summarizes the results of the four first canonical correlations.

3.5.1. The first canonical correlation highlights five G- and N-rich structural classes

The first canonical correlation coefficient R_1 is equal to 0.97, which corresponds to 94% of explained variance (= R_1^2). This canonical correlation between sequence and local structure points out 5 particular classes with a well defined PBs and amino acids composition, *i.e.*, #32, #104, #120, #87 and #48 (Fig. 7_{F1}); they have loading values far away from all the other sites. These classes located in turn regions have been detailed and discussed above. The analysis of the canonical factor loadings (threshold fixed at 0.3) confirms their structural characteristics with the presence of PBs *i*, *j* or *p*, and to some extent by the absence of *m* and *d* (*i.e.*, regular secondary structures). In terms of amino acid characteristics, the major role played by Glycine and Asparagine in these structural prototypes is confirmed. These two amino acids are opposed to nearly all the others, which are strongly under-represented when they are under-represented. As expected, this first contrast factor emphasizes structural classes that present a strong sequence - structure determinism.

3.5.2. Opposition between α-helices and β-strands is pointed out by the second canonical correlation

The second canonical correlation highlights an opposition between α-helices to β-strands, creating on the Figure 7_{F2} a gradient from one to another. The canonical correlation coefficient R_2 is equal to 0.94, *i.e.*, 89% of explained variance. This gradient opposes the classes characterizing the α-helices, *e.g.*, α1, α2, α3, and their edges, to those characterizing the β-strands, *e.g.*, β1, β5, β6, and their edges. With regard to the canonical factor loadings, the significant PBs are *m*, *n* and *l* what are opposed to *d* and *c.* The amino acid preferences pointed out to the equivalence classes previously associated to the local structures are found: the amino acid equivalence classes IV and V favored in α-helices opposed to the equivalence classes VI, VII preferred in β-strands and to the class II, present in the edges.

3.5.3. The third canonical correlation stresses a differentiation between β-strands and loops

The third canonical correlation opposites primarily loops, *e.g.*, T2, T3, T5, T6, T7 to β-strands and their transition towards α-helices, *e.g.* β1, β3, T1, T5. The canonical correlation coefficient R_3 is equal to 0.93, *i.e.*, 86% of explained variance. The significant PBs that emerges from the factor loadings analysis are *f*, *k* and *h*, opposed to *d* (Fig. 7 $_{F3}$). Concerning the amino acids, we observe the equivalence classes II and III opposed to parts of IV, VI and VII, a result in agreement with their preferential local structures. The PBs (*d*, *f*, *h* and *k*) pointed up with the third canonical correlation characterize transitions between β-strands and α-helices. Theses *transitions* are also found in agreement with the literature, as the canonical correlation opposite amino acid characteristic of loops (*e.g.*, by Pro, Asp, Ser, Asn, Thr) and amino acid characteristic of β-strands (*e.g.*, Ile, Val, Phe, Tyr, Met) and α-helices (*e.g.*, Leu).

3.5.4. The fourth canonical correlation reveals a contrast between two categories of loops

The fourth canonical correlation more surprisingly puts in opposition two different categories of loops. The percentage of explained variance, however, strongly decreased (R^2_4=68%, *i.e.*, a fall of around 20% relative to the previous R^2 contribution). The significant PBs according to their canonical factor loadings are *k* opposite to *f* and *o*. The amino acids analysis exhibits a category of loops characterized by Pro (amino acid equivalence class II) in contrast to loops characterized by His, Asn, Ser, Asp and Thr (amino acid equivalence class III, primarily) (Fig. 7 $_{F4}$). It is worthwhile to note that successive structural classes appear in different extremities of the gradient, *e.g.*, #19, #39, #53, #65, #84 opposing to #18, #38, #54, #64, #83. Interestingly, the structural class #38, which was associated with #39 and #53 at the same extremity of the gradient in the third canonical correlation, is now opposite to them. This point underlines specific amino acid contents in two contiguous HPM sites. Hence, this

fourth canonical correlation enables to discriminate between different types of loops. Other analyses are presented in supplementary data 6.

4. Discussion

The Hybrid Protein Model enables us to carry out the compression of a whole protein structure databank by a fast and efficient processing. The library obtained is composed of 120 contiguous prototypes with a high structural stability and representative of the local structural motifs encountered in the protein structures from the databank. Hence, it constitutes a useful tool for accurately describing 3D structures in terms of long local structural prototypes. These results are interesting as the characterization of local structures is a complex problem (Karchin et al., 2003; Kolodny et al., 2002; Micheletti et al., 2000). The use of overlapping fragments encoded into a structural alphabet enables to circumvent the difficulty of finding a length for describing all protein local structures. Indeed, despite the use of a fixed fragment length (7 PBs) for the training, the HPM approach, based on the concept of "information sharing", allows the definition of longer structural regions.

The HPM training involves different parameters that directly influence the final prototype library. The sensitivity study of the training according to these control parameters has already been carried out to ensure the building of an optimal prototype library (Benros et al., 2003). A global structural stability of the library is observed in the different trials of HPM building: the significant structural regions related to the regular secondary structures and to their transitions are systematically found.

Compared to our previous study, we have added different improvements in the HPM strategy (de Brevern and Hazout, 2001). For example, the introduction of gaps within the local structural fragments is an interesting tool. It permits to take account of the regular secondary structure length variability and enlighten the heterogeneity of some local structures,

e.g., regular structure edges (see Fig. 5). Interestingly, our results suggest that long loops are combinations of small ones like as suggested in previous studies and strengthen the idea developed by (Ring et al., 1992) who described compound loops as combinations of simple loops. In addition, it appears clearly that gaps contribute to improve the HPM site specificity in terms of PB signatures. The HPM is equivalent to a "structural profile" (Gribskov et al., 1987), resulting from multiple alignments of local structural fragments with gaps. The HPM can also be compared to a local Hidden Markov Model (Rabiner, 1989), with the advantage of not requiring *a priori* parametric distribution laws.

Among the different structural motifs of the library, tight turns are of great interest because they occur frequently in protein structures (Chou, 2000). We highlight eight zones for these turns precisely located along the HPM. This latter enables the analysis of turns in their environment since the fragments clustered are 7 PBs in length, *i.e.*, 11 amino acids. It could be of great help for the precise analysis of turns as their assignment is strongly dependant of the assignment of repetitive structures (Bornot and de Brevern, 2006). Specific PB series characterized turns and three main signatures have been identified; they show highly specific transitions between the successive PBs. Moreover, the effect of environment can lead a same population of turns to be encoded by different series of PBs, like the β type I turn associated with the series *fklm* and *mnop*. So, they are distributed in different HPM regions. Furthermore, a precise PB series associated to a particular turn can be found in different HPM regions, like the series *ehia* associated to the β type II turns and encountered in HPM regions T3b, T6b and T8b. Conversely, different turn types can be encoded by the same PB series and associated to the same HPM regions, like the series *ehia* for the β type II and inverse I turns localized in T6b. The flanking zones of the turn thus play a major role in the determination of its location in the HPM. Furthermore, our analysis suggests that longer turns are generated by a limited number of PB motifs. These motifs are found in the Structural Words we have previously

described (de Brevern et al., 2002a).

Analysis reveals that the amino acid informativity is high along the HPM. Beyond the informativity associated with the regular secondary structures, the HPM highlights informative structural classes corresponding to loops. As an illustration, the sequence information content is particularly high for the HPM regions associated to turns.

Through the building of our structural prototype library, we have identified seven amino acid equivalence classes, and, linked these equivalence classes with their preferential local structures. As expected, the specific roles played by Glycine, which can increase local flexibility in structures, and Proline, which forms kinks, are pointed out, since these two amino acids are clearly isolated from the others. The strong preferences of Alanine and Leucine for α-helices and of Isoleucine and Valine for β-strands appear clearly. The HPM exhibits also significant preferences for Glycine and Asparagine in turn zones. The results obtained are in accordance with classical preferences (de Brevern et al., 2000; Offmann et al., 2007). Moreover, the HPM provides additional information notably on the association of the different equivalence classes. With these seven amino acid equivalence classes identified through the HPM, the complex relation existing between the set of 20 amino acids and the structure can be simplified. This approach is slightly different from the classical ones used (Etchebest et al., 2007; Murphy et al., 2000; Wang and Wang, 1999).

The canonical correlation analysis gives several interesting properties concerning the prototypes sequence - structure specificity. For example, the structural prototypes localized in HPM turn areas and characterized primarily by Glycine have a higher sequence – structure correlation than the regular secondary structures. This observation emphasizes the sequence – structure determinism existing in loops. Furthermore, it is worthwhile to note that loops can be separated only with the fourth correlation. It underlines the difficulty of differentiating them and the interest of HPM approach as discriminatory criteria. A major result concerns the

identification of two categories of loop prototypes distinguishable by their structure and sequence information content. These results will be of great interest for protein local structure prediction. However, it appears from the analysis of the canonical communality coefficient that the amino acid variability is only partially explained. Other factors are probably required, such as long range dependency.

5. Conclusion

Protein structures can be seen as a combination of small local structures yielding a more detailed description than classical secondary structures. A complete set of prototypes defines ‘‘a structural alphabet’’ that approximates accurately protein structures (Benros et al., 2007; Karplus et al., 2003; Offmann et al., 2007; Sander et al., 2006; Tyagi et al., 2007). We have proposed such a structural alphabet which is composed of 16 average protein fragments of 5 residues in length called Protein Blocks (PBs) (de Brevern, 2005). This alphabet was used both to describe 3D protein backbones but also to perform local structure prediction (de Brevern et al., 2000; de Brevern et al., 2007; de Brevern et al., 2004; Etchebest et al., 2005). Moreover, PBs have proven their efficiency both in description and prediction of longer fragments (de Brevern et al., 2002a) and loop conformations (Fourrier et al., 2004), to predict protein class(de Brevern et al., 2005a), to compare protein structures (Tyagi et al., 2008; Tyagi et al., 2006a; Tyagi et al., 2006b), to detect magnesium-binding sites (Dudev and Lim, 2007), to build protein structures (Dong et al., 2007) and transmembrane proteins (de Brevern et al., 2005b).

We extended this description to longer fragments thanks to HPM (de Brevern and Hazout, 2000; de Brevern and Hazout, 2001). New developments were proposed (de Brevern and Hazout, 2003) as a specific one to treat genomic data (de Brevern, 2002; de Brevern et al., 2002b). It was recently used to predict protein local structure from sequence (Benros et al.,

2006). Here, we evaluated the influence of introductions of gaps during the training process of HPM library. Thanks to the specific research done on the different kinds of turns; we have proved the efficiency of HPM to take into account characteristic local protein structures in different local neighborhood. Thus, this original procedure yields new interesting features about structural motifs and their sequence signature. Canonic correlations show also that HPM training capture classical sequence – structure relationship, even only the structure encoded as PB strings are used.

Local protein structure is one of the most successful and approaches to generate structural model based (Du et al., 2003; Pei and Grishin, 2004). Compared to similar fragment-based approaches, our library will have the advantage of presenting overlapping prototypes, which will be of great interest while reconstructing the global 3D structure.

Acknowledgments

This paper is dedicated to the memory of Pr. Serge Hazout. This work was supported by grants from the Ministère de la Recherche, from the French Institute for Health and Medical Care (INSERM), from Université Paris Diderot –Paris 7 and from "Action Bioinformatique inter EPST" 2001-2002 number 4B005F and 2003-2004. Cristina Benros had a grant from the Ministère de la Recherche. Alexandre G. de Brevern was supported by a grant from the Fondation de la Recherche Médicale. The second author wishes to express his appreciation to Aurélie Bornot and Catherine Etchebest for reading parts of the manuscript and for their valuable suggestions.

Captions

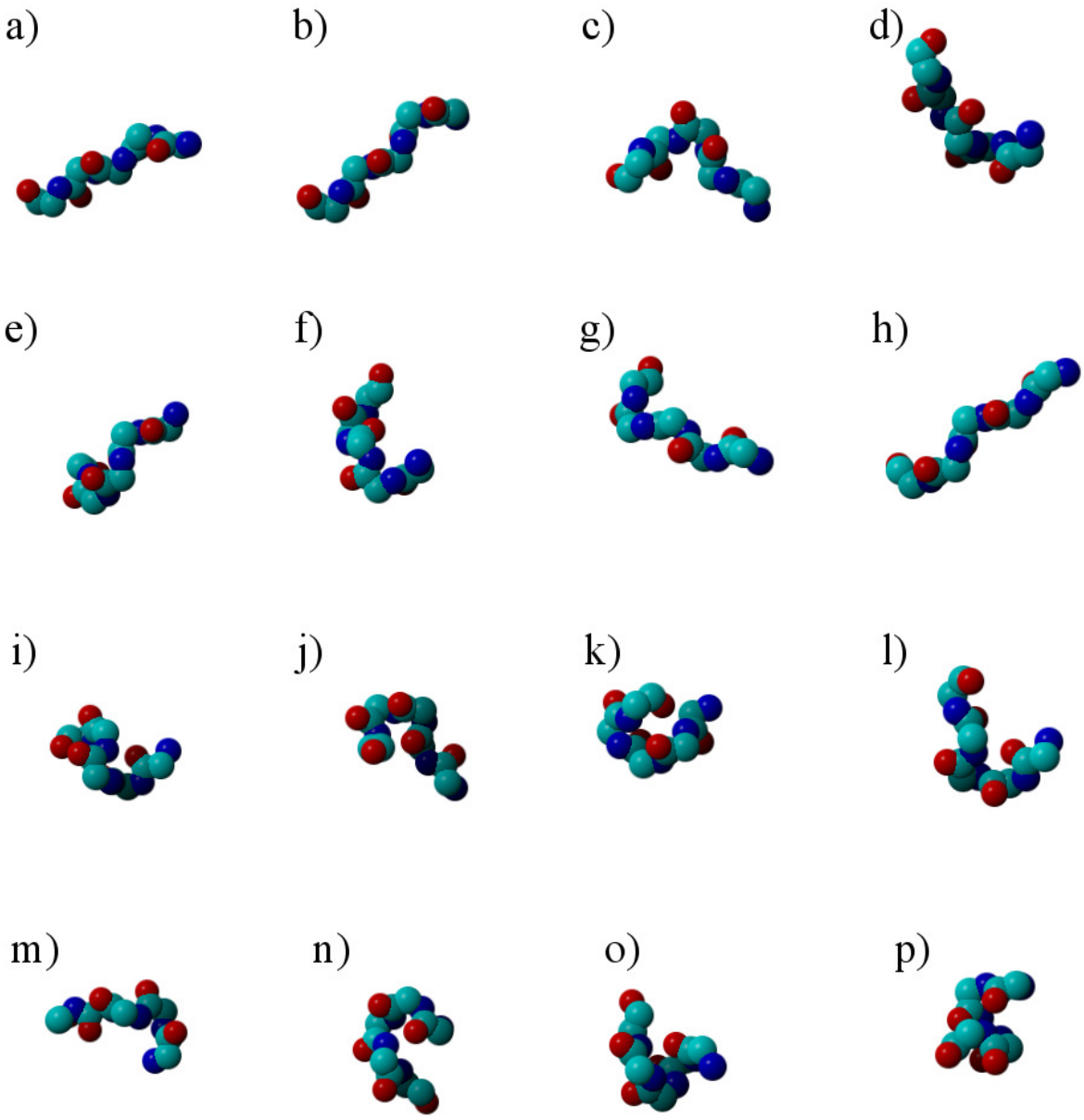

Figure 1. Protein Blocks: From left to right and top to bottom, YASARA (http://www.yasara.org/) images of the 16 Protein Blocks of the structural alphabet. Each prototype is five residues in length and corresponds to eight dihedral angles (ϕ,ψ). The PBs *m* and *d* can be roughly described as prototypes for the central α-helix and the central β-strand, respectively. For each PB, the N-cap extremity is on the left and the C-cap on the right.

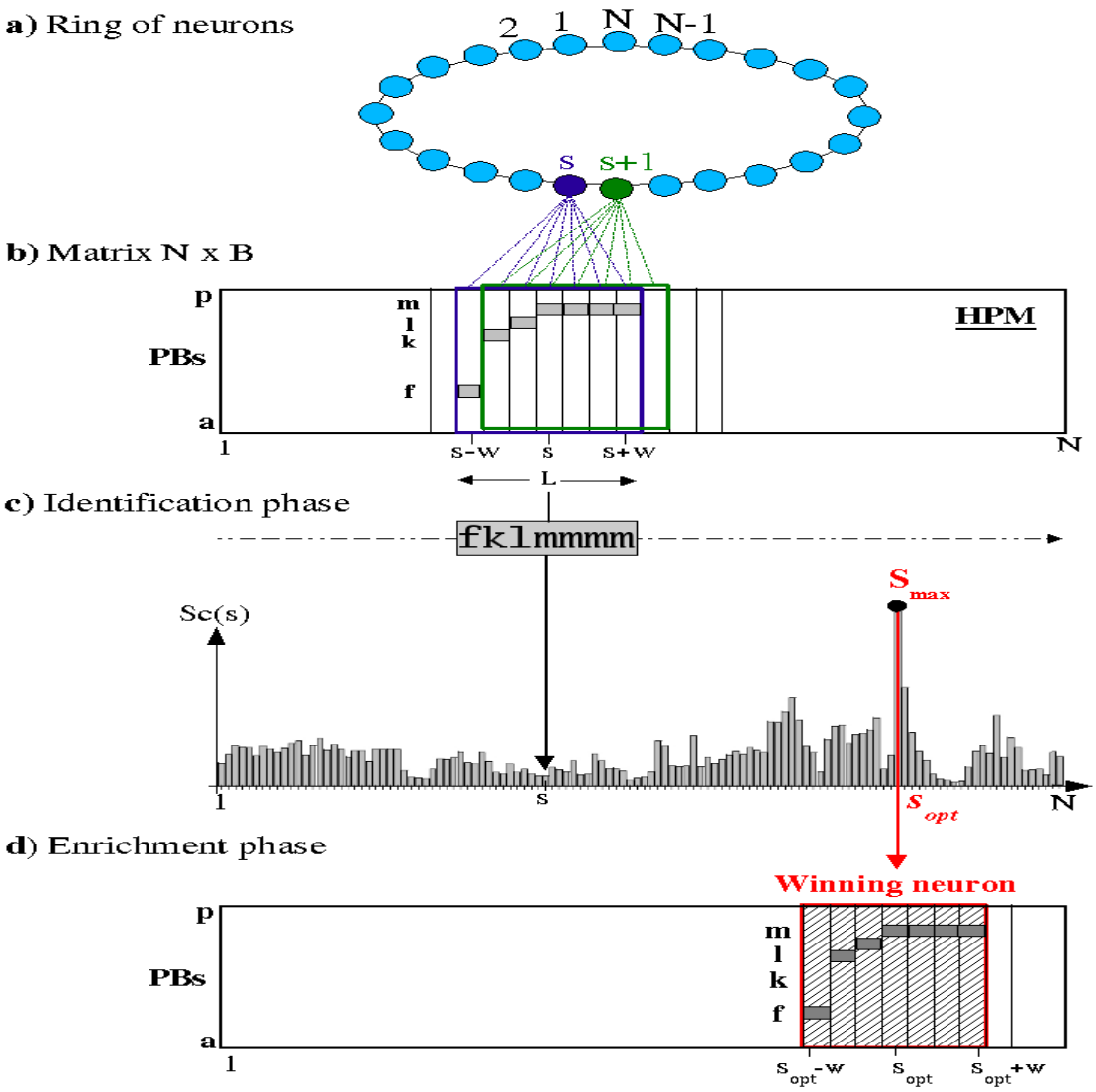

Figure 2 Training of the Hybrid Protein Model (HPM): (a) The HPM is a closed linear neural network and can be represented by a ring of *N* neurons, *i.e.*, structural classes (in our study, *N* = 120). (b) Each neuron is associated to *L* distributions of the 16 PBs composing our structural alphabet (in our study, *L* = 7). Hence, the HPM can be represented by a matrix of dimension *N* x 16. Two successive structural classes are overlapping since they share (*L*-1) sites. (c) For each structural fragment of *L* successive PBs taken randomly from the databank (for example, *fklmmm*), we search for the most similar pattern present in the HPM. Consequently, a log odds score *Sc*(*s*) is computed along the HPM. The identification phase consists in selecting the structural class s_{opt} the most similar to the local structure presented, *i.e.*, the winning neuron associated to the maximum score S_{max}. (d) The enrichment phase consists in slightly modifying the PB distributions of the structural class corresponding to the site s_{opt}, to increase the likeness between the winning neuron and the presented fragment.

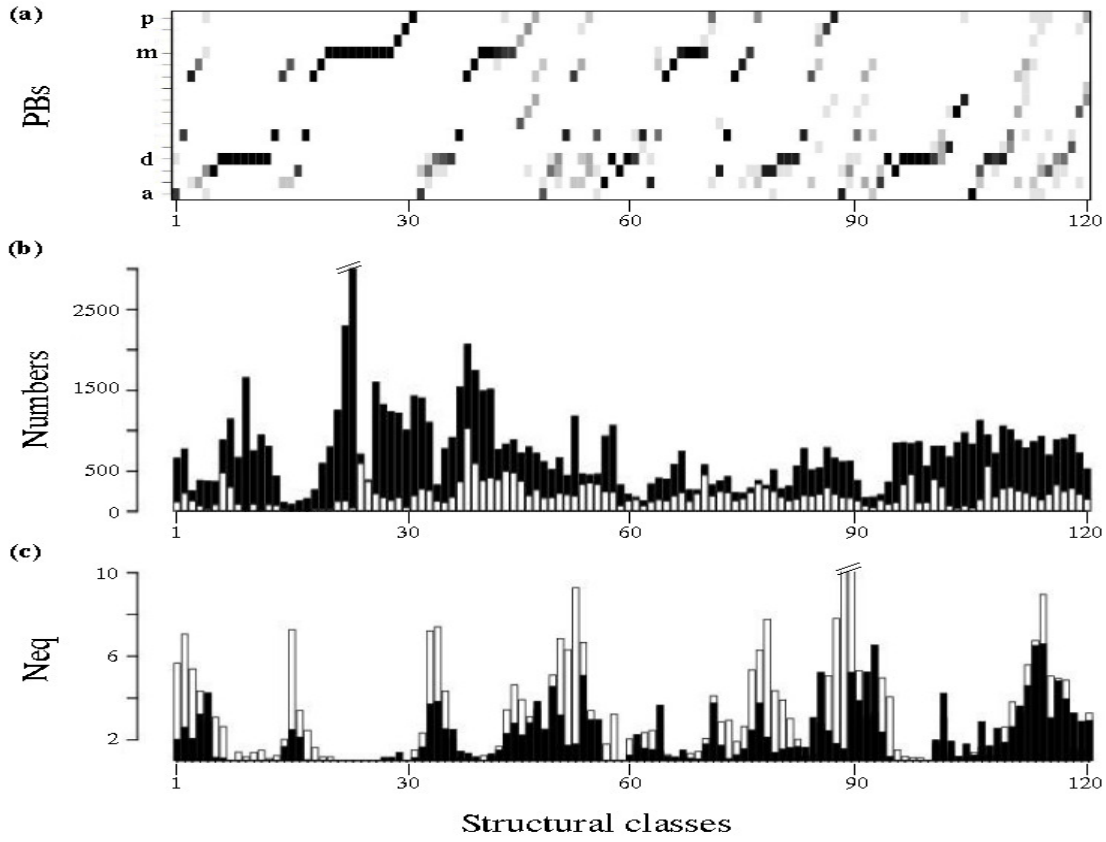

Figure 3. (a) Final Hybrid Protein Model, *i.e.*, matrix of PBs distributions. The gray level indicates the PB frequency, which varies from 0 (*i.e.*, white) to 1 (*i.e.*, back). (b) Distribution of the structural fragments: total number of fragments (in black) and number of fragments with gaps at each HPM site (in white). The number of fragments in the site 24 is 13,790. The peak has been truncated (c) Specificity along the HPM quantified by the Neq value (equivalent number of PBs) for fragments with gaps (in black) and without gaps (in white). The Neq values for the sites 88 and 89 are 10.3 and 12.4, respectively.

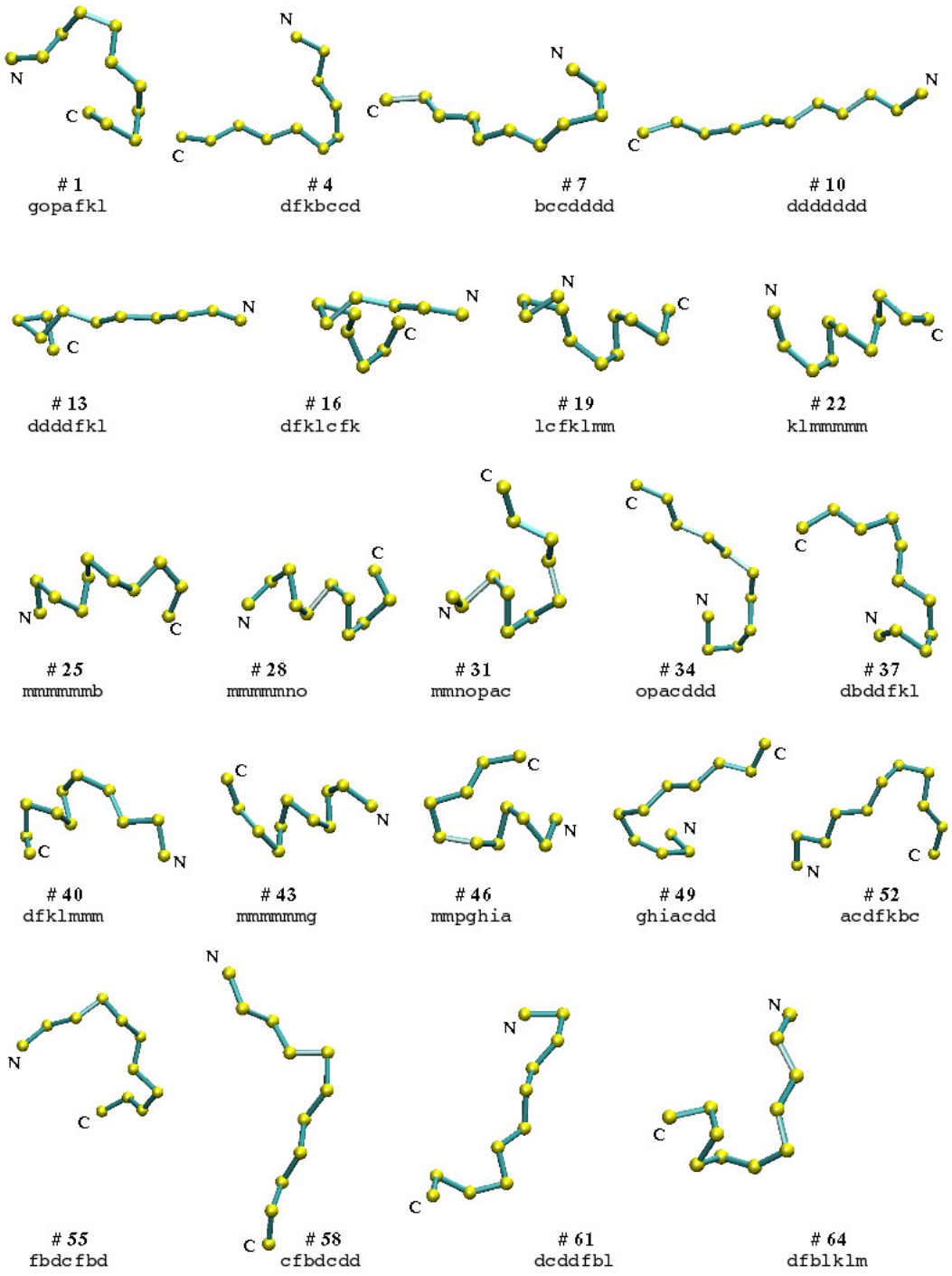

Figure 4. Structural prototypes of the HPM, corresponding to the sites 1 to 64 by steps of 3 (see supplementary data 3 for the last ones). Each prototype of 11 amino acids corresponds to a series of 7 PBs. The PB chains of the corresponding structural classes are indicated as their N- and C-terminal extremities. It is possible to see the overlapping between consecutive prototypes, *e.g.*, the 8 N-terminal residues of the prototype 4 overlaps the 8 C-terminal residues of prototype 1, and so on.

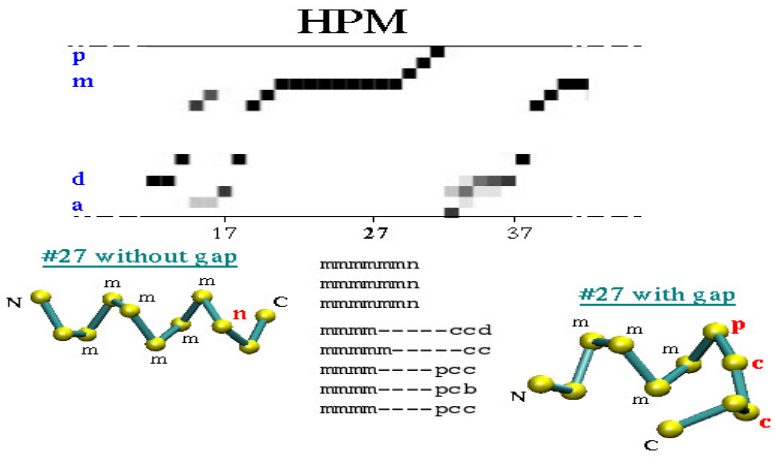

Figure 5. Examples of structural fragments with and without gaps associated to the HPM site 27. The introduction of gaps in some structural fragments enables to point out two different α-helix C-terminal extremities.

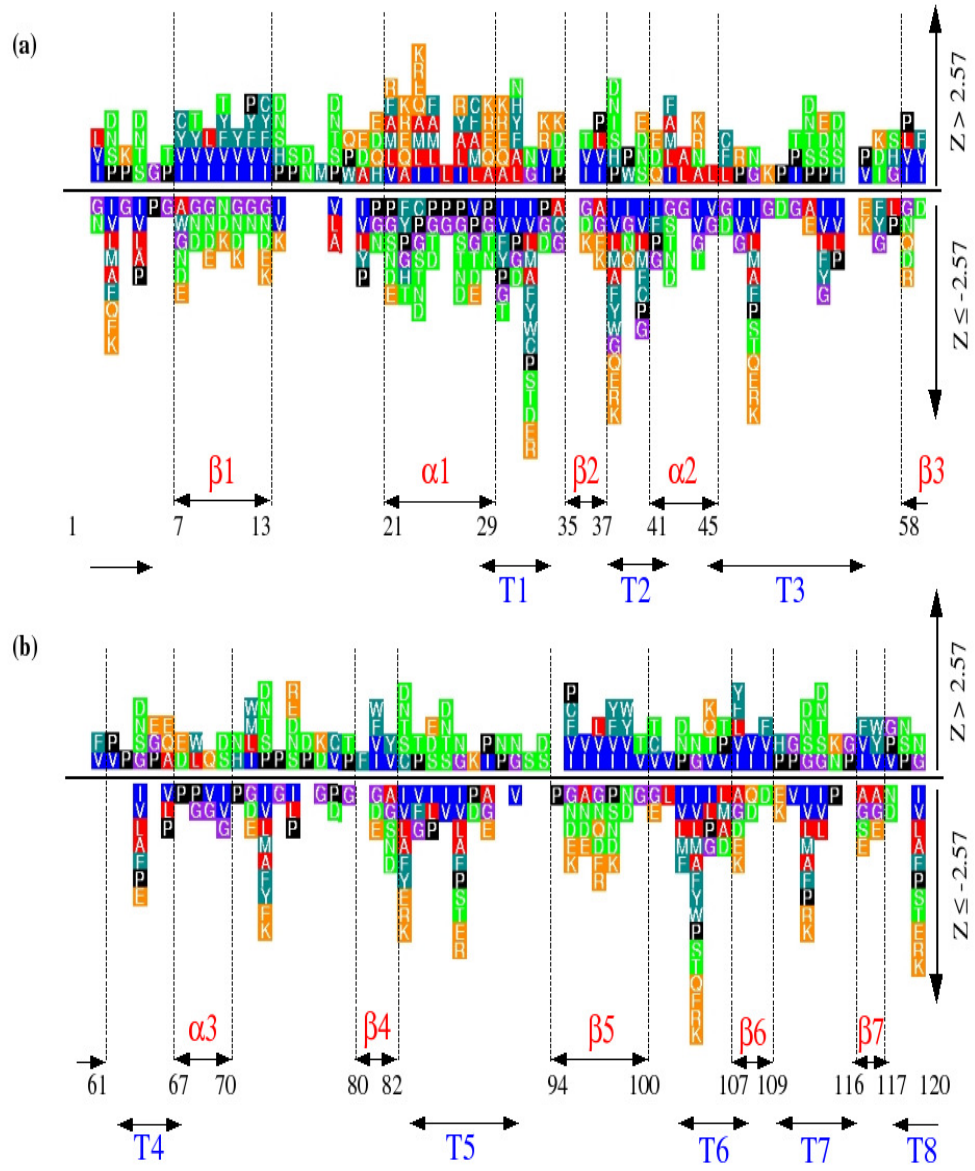

Figure 6. Over-represented (respectively under-represented) amino acids are displayed above the line (respectively under the line) along the HPM (Z-scores > 2.57 and Z-scores ≤ -2.57, respectively). (a) sites 1 - 60 and (b) sites 61 - 120. A color is associated to each amino acid group previously identified by the hierarchical clustering (7 groups; see paragraph 3.4 and supplementary data 5). We have located the different regions of regular secondary structures (α-helices and β-strands) as well as the regions corresponding to turns.

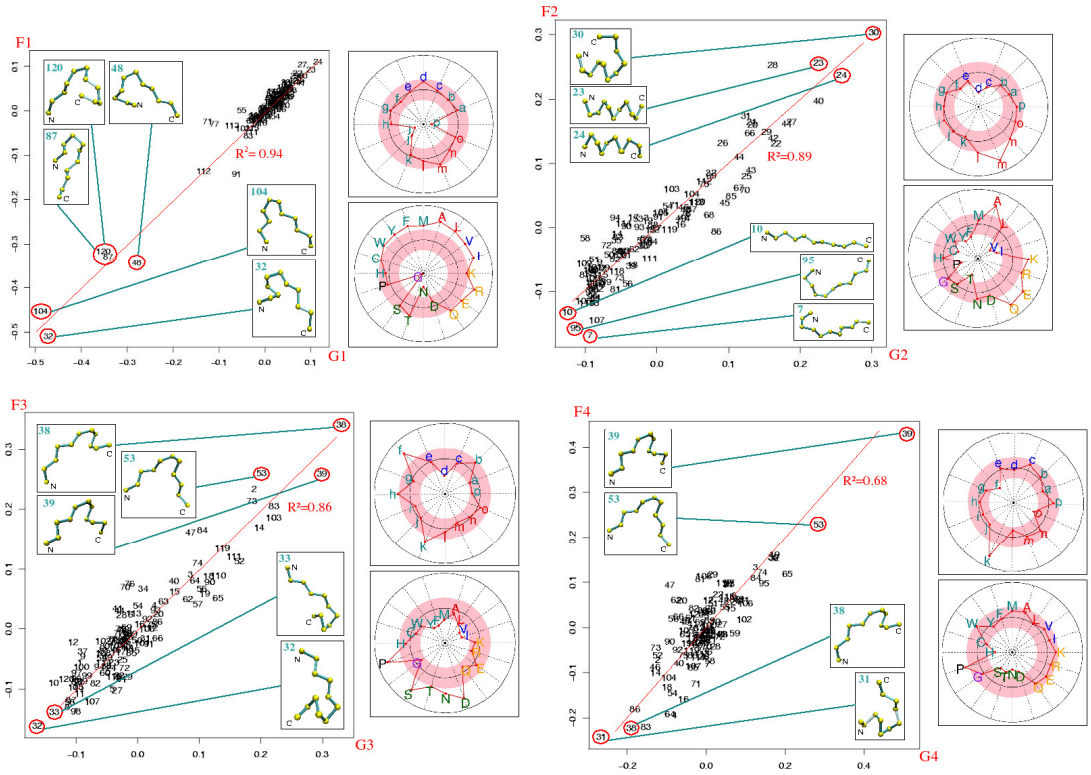

Figure 7. Canonical correlation analysis between the sequence and the structure data associated to the HPM sites. Only the results of the four first canonical correlations are displayed. For each one, we have represented the canonical correlation plot with particular local structures pointed out. The squared canonical correlation, which corresponds to the percent of explained variance, is printed on the regression line. The values of the canonical factor loadings associated to the PBs and to the amino acids are also displayed with a circular representation. Thresholds are fixed at +0.3 for the over-representations and -0.3 for the under-representations, delimiting a non-significant region (in pink). Hence, variables here and there from this region are opposed. The amino acid colors are assigned according to the 7 groups previously identified by the hierarchical clustering. The PBs are colored according to their local structures. (a) The first canonical factor points out 5 structural prototypes, i.e. #32, #48, #87, #104 and #120, corresponding to transitions between regular secondary structures and characterized by Glycine and Asparagine. (b), (c) and (d) The three other canonical factors correspond to gradients.

Table I

Analysis of the γ-turns, β-turns and α-turns in the HPM. We focused on the three major locations of the turn of interest in the HPM and the corresponding turn proportions are shown as well as the Neq value of the HPM region. Moreover, for each turn type, it is specified its main PB signatures and the number found in the databank. The PBs displayed in bold are highly significant (Z-scores > 4.4). Eight different HPM turn regions have been identified and each one is characterized by sub-regions. They were labelled as follows : T1a=[29-32] and T1b=[29-33]; T2=[38-41]; T3a=[45-49], T3b=[46-49], T3c=[46-50], T3d=[47-50], T3e=[49-53] and T3f=[52-55]; T4=[63-67]; T5a=[83-87], T5b=[84-88], T5c=[84-92], T5d=[85-88], T5e=[85-89], T5f=[86-87;89-91] and T5g=[88-91]; T6a=[101-105], T6b=[102-105], T6c=[103-105] and T6d=[105-108]; T7a=[110-114], T7b=[110-115], T7c=[111-113], T7d=[111-114] and T7e=[114-115;117-1]; T8a=[117-1], T8b=[118-1], T8c=[118-2], T8d=[119-1], T8e=[119-2] and T8f=[119-120;3-5]. Six sub-regions show gaps, namely T3e, T5c, T5f, T7b, T7e and T8f. The turns have been extracted from the 1143 protein chains of set 2.

References

Baker, D., and Sali, A., 2001. Protein structure prediction and structural genomics. Science 294, 93-6.

Benros, C., de Brevern, A.G., and Hazout, S., Hybrid Protein Model (HPM) : A Method For Building A Library Of Overlapping Local Structural Prototypes. Sensitivity Study And Improvements Of The Training, IEEE Workshop on Neural Networks for Signal Processing 2003, pp. 53-72.

Benros, C., de Brevern, A.G., Etchebest, C., and Hazout, S., 2006. Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62, 865-80.

Benros, C., Martin, J., Tyagi, M., and de Brevern, A.G., Description of the local protein structure. I. Classical approaches, in: de Brevern, A. G., (Ed.), Recent Adv. In Structural Bioinformatics, Vol. 1. Research signpost 2007, pp. 1-33.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235-42.

Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., and Baker, D., 2002. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322, 65-78.

Bornot, A., and de Brevern, A.G., 2006. Protein beta-turn assignments. Bioinformation 1, 153-5.

Bystroff, C., and Baker, D., 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 281, 565-77.

Chou, K.C., 2000. Prediction of tight turns and their types in proteins. Anal Biochem 286, 1-16.

de Brevern, A.G., 2002. Compartimentation chromosomique. Biofutur 225, 20-22.

de Brevern, A.G., 2005. New assessment of Protein Blocks. In Silico Biology 5, 283-289.

de Brevern, A.G., and Hazout, S., 2000. Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties. IEEE - Computer Society S1, 49-54.

de Brevern, A.G., and Hazout, S., 2001. Compacting local protein folds with a "hybrid protein model". Theo Chem Acc 106, 36-47.

de Brevern, A.G., and Hazout, S., 2003. 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19, 345-53.

de Brevern, A.G., Etchebest, C., and Hazout, S., 2000. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41, 271-87.

de Brevern, A.G., Benros, C., and Hazout, S., Structural Alphabet: From a Local Point of View to a Global Description of Protein 3D Structures, in: Yan, P. V., (Ed.), Bioinformatics: New Research Nova Publishers 2005a, pp. 127-169.

de Brevern, A.G., Valadie, H., Hazout, S., and Etchebest, C., 2002a. Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. Protein Sci 11, 2871-86.

de Brevern, A.G., Etchebest, C., Benros, C., and Hazout, S., 2007. "Pinning strategy": a novel approach for predicting the backbone structure in terms of Protein Blocks from sequence. J Biosciences 32, 51-70.

de Brevern, A.G., Loirat, F., Badel-Chagnon, A., Andre, C., Vincens, P., and Hazout, S., 2002b. Genome compartimentation by a hybrid chromosome model (HXM). Application to Saccharomyces cerevisae subtelomeres. Comput Chem 26, 437-45.

de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S., and Etchebest, C., 2004. Local backbone structure prediction of proteins. In Silico Biol 4, 381-6.

de Brevern, A.G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C., and Etchebest, C., 2005b. A structural model of a seven-transmembrane helix receptor: The Duffy antigen/receptor for chemokine (DARC). Biochim Biophys Acta 1724, 288-306.

Dong, Q.W., Wang, X.L., and Lin, L., 2007. Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med 37, 1610-6.

Du, P., Andrec, M., and Levy, R.M., 2003. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. Protein Eng 16, 407-14.

Dudev, M., and Lim, C., 2007. Discovering structural motifs using a structural alphabet: Application to magnesium-binding sites. BMC Bioinformatics 8, 106.

Etchebest, C., Benros, C., Hazout, S., and de Brevern, A.G., 2005. A structural alphabet for local protein structures: Improved prediction methods. Proteins 59, 810-827.

Etchebest, C., Benros, C., Bornot, A., Camproux, A.C., and de Brevern, A.G., 2007. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36, 1059-69.

Fourrier, L., Benros, C., and de Brevern, A.G., 2004. Use of a structural alphabet for analysis of short loops connecting repetitive structures. BMC Bioinformatics 5, 58.

Frishman, D., and Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23, 566-79.

Fuchs, P., Etchebest, C., and de Brevern, A.G., Turns prediction, in: de Brevern, A. G., (Ed.), Recent Adv. In Structural Bioinformatics, Vol. 1. Research signpost, Trivandrum 2007, pp. 37-56.

Fuchs, P.F., and Alix, A.J., 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59, 828-39.

Gonzalez-Diaz, H., Vilar, S., Santana, L., and Uriarte, E., 2007. Medicinal chemistry and bioinformatics--current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7, 1015-29.

Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F.M., and Uriarte, E., 2008. Proteomics, networks and connectivity indices. Proteomics 8, 750-78.

Gribskov, M., McLachlan, A.D., and Eisenberg, D., 1987. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84, 4355-8.

Haspel, N., Tsai, C.J., Wolfson, H., and Nussinov, R., 2003. Reducing the computational complexity of protein folding via fragment folding and assembly. Protein Sci 12, 1177-87.

Hazout, S., Entropy-derived measures for assessing the accuracy of N-state prediction algorithms., in: de Brevern, A. G., (Ed.), In Recent Advances in Structural Bioinformatics. , Research signpost, Trivandrum, India 2007, pp. pp. 395-417.

Hotelling, H., 1936. between two sets of variates. Biometrika 28, 321-377.

Humphrey, W., Dalke, A., and Schulten, K., 1996. VMD: visual molecular dynamics. J Mol Graph 14, 33-8, 27-8.

Hutchinson, E.G., and Thornton, J.M., 1996. PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci 5, 212-20.

Ihaka, R., and Gentleman, R., 1996. R: a language for data analysis and graphics. J Comput Graph Stat 5, 299-314.

Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H.J., 2003. Protein structure prediction via combinatorial assembly of sub-structural units. Bioinformatics 19 Suppl 1, i158-68.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K., 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51, 504-14.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., and

Hughey, R., 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 53 Suppl 6, 491-6.
Kohonen, T., 2001. Self-Organizing Maps (3rd edition). Springer.
Kolodny, R., Koehl, P., Guibas, L., and Levitt, M., 2002. Small libraries of protein fragments model native protein structures accurately. J Mol Biol 323, 297-307.
Kullback, S., and Leibler, R.A., 1951. On information and sufficiency. Ann Math Stat 22, 79-86.
Lesk, A.M., and Rose, G.D., 1981. Folding units in globular proteins. Proc Natl Acad Sci U S A 78, 4304-8.
Micheletti, C., Seno, F., and Maritan, A., 2000. Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 40, 662-74.
Munteanu, C.R., Gonzalez-Diaz, H., Borges, F., and de Magalhaes, A.L., 2008. Natural/random protein classification models based on star network topological indices. J Theor Biol, in press.
Murphy, L.R., Wallqvist, A., and Levy, R.M., 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13, 149-52.
Noguchi, T., and Akiyama, Y., 2003. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 31, 492-3.
Offmann, B., Tyagi, M., and de Brevern, A.G., 2007. Local Protein Structures. Current Bioinformatics 3, 165-202.
Pavone, V., Gaeta, G., Lombardi, A., Nastri, F., Maglio, O., Isernia, C., and Saviano, M., 1996. Discovering protein secondary structures: classification and description of isolated alpha-turns. Biopolymers 38, 705-21.
Pei, J., and Grishin, N.V., 2004. Combining evolutionary and structural information for local protein structure prediction. Proteins 56, 782-94.
Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected application in speech recognition. Proceedings of the IEEE 77, 257-286.
Ring, C.S., Kneller, D.G., Langridge, R., and Cohen, F.E., 1992. Taxonomy and conformational analysis of loops in proteins. J Mol Biol 224, 685-99.
Rose, G.D., Gierasch, L.M., and Smith, J.A., 1985. Turns in peptides and proteins. Adv Protein Chem 37, 1-109.
Sali, A., and Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.
Sander, O., Sommer, I., and Lengauer, T., 2006. Local protein structure prediction using discriminative models. BMC Bioinformatics 7, 14.
Schuchhardt, J., Schneider, G., Reichelt, J., Schomburg, D., and Wrede, P., 1996. Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng 9, 833-42.
Shannon, C., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379-423.
Tyagi, M., Benros, C., Martin, J., and de Brevern, A.G., Description of the local protein structure. II. Novel approaches, in: de Brevern, A. G., (Ed.), Recent Adv. In Structural Bioinformatics, Vol. 1. Research signpost, Trivandrum 2007, pp. 34-47.
Tyagi, M., de Brevern, A.G., Srinivasan, N., and Offmann, B., 2008. Protein structure mining using a structural alphabet. Proteins 71, 920-37.
Tyagi, M., Gowri, V.S., Srinivasan, N., de Brevern, A.G., and Offmann, B., 2006a. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 65, 32-9.
Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G., and

Offmann, B., 2006b. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 34, W119-23.

Unger, R., Harel, D., Wherland, S., and Sussman, J.L., 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5, 355-73.

Wang, J., and Wang, W., 1999. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 6, 1033-8.

Xu, D., Crawford, O.H., LoCascio, P.F., and Xu, Y., 2001. Application of PROSPECT in CASP4: characterizing protein structures with new folds. Proteins Suppl 5, 140-8.