New assessment of a structural alphabet

Alexandre G. de Brevern *[1]

[1]Equipe de Bioinformatique Génomique et Moléculaire (EBGM),
INSERM U 726, Université Denis DIDEROT - Paris 7, case 7113,
2, place Jussieu, 75251 Paris, France

* Corresponding author:
mailing address: Dr. de Brevern A.G., Equipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U 726, Université Denis DIDEROT - Paris 7, case 7113, 2, place Jussieu, 75251 Paris, France
E-mail : debrevern@ebgm.jussieu.fr
Tel: (33) 1 44 27 77 31
Fax: (33) 1 43 26 38 30

Running title: Protein Blocks

key words: secondary structure, structure-sequence relationship, *ab initio*.

Summary: A statistical analysis of the Protein Databank (PDB) structures had led us to define a set of small 3D structural prototypes called Protein Blocks (PBs). This structural alphabet includes 16 PBs, each one defined by the (Φ, Ψ) dihedral angles of 5 consecutive residues. Here, we analyze the effect of the enlargement of the PDB on the PBs' definition. The results highlight the quality of the 3D approximation ensured by the PBs. These last could be of great interest in *ab initio* modeling.

Introduction. Protein folds are often described as a succession of secondary structures. Nonetheless, this description does not allow the protein structures to be precisely described at the 3D level, because (i) it omits the relative orientation of connecting regions, (ii) irregularities are found both in helical and extended structures [1, 2] and (iii) the coil state, which represents 50% of all residues, corresponds to a large set of distinct local protein structures [3].

These observations have led to a new view of 3D protein structures. They are now thought to be composed of a combination of small local structures or fragments defining "a structural alphabet" [4] Different teams have described these local protein structures according to different criteria (*e.g.* [5]).

Our structural alphabet is composed of 16 average protein fragments, 5 residues in length, which we call Protein Blocks. These PBs have been used both to describe 3D protein backbones and to predict local structures [6 - 8]. They have been used to analyze long protein fragments [9 - 12] and to predict short loops [13]. Karchin and co-workers have compared the features of this alphabet with those of 8 other structural alphabets. Their results show clearly that our PB alphabet is highly informative, with the best predictive ability of those tested [14]. Here, we present a new evaluation of the PB features with an updated databank. This analysis focuses on the distribution of PB's frequencies, their main transitions, the relationship

between PBs and secondary structures and the evaluation of geometrical features of PBs with different criteria.

Datasets: This study considers four sets of proteins used in recent work [7, 9]. We preferentially used the *PAPIA* set, from the PDB-REPRDB database [15] composed of 717 protein chains and 180,854 residues. The set contains proteins with no more than 30% pairwise sequence identity, X-ray crystallographic resolutions better than 2.0 Å, and an R-factor less than 0.2. Each selected structure has a *rmsd* (root mean square deviation, average Euclidean distance between superimposed C_α) value greater than 10 Å between every representative chain. An updated dataset [8] is defined from the PDB-REPRDB database [15] with the same criteria as *PAPIA*. It comprises 1407 protein chains and 293,507 residues.

Protein coding: The protein structures are encoded as sequences of ϕ - ψ dihedral angles. They are cut into consecutive overlapping fragments, each M (= 5) amino acids in length. A fragment is defined by a signal of 2(M-1) dihedral angular values (= 8). The fragment signal is compared with each PB signal with the *rmsda* (root mean square deviation on angular values, the Euclidean distance of dihedral angles) measure. The lowest *rmsda* value for the 2(M-1) angles determines the PB assignment [6]. A PB is assigned to each position along the sequence except for the two residues on each end.

PB's frequencies: The frequencies of the PBs remain stable between all the different databanks used (see column 2 of Table 1). Most of the PBs have a frequency variation less than 0.2% (compared to our previous study [6]). The greatest difference is observed for PB *c*, which decreases by 0.5% (from 8.63% to 8.12%). The PB *j* remains the less frequent PB (0.83%). The central part of repetitive structures, *i.e.* PB *m* for the α-helix and PB *d* for the β-

strand represent 49.1% of all the PBs. Coarsely, the C and N-caps of PB *m* (PBs *k*, *l*, *n*, *o* and *p*) represent 19.1% of the databank, the C and N-caps of PB *d* (PBs *a*, *b*, *c*, *e* and *f*) 25.5% and the PBs entirely coils (PBs from *g* to *j*) 6.3%. The number of PB repeats, *anr*, (see column 3 of Table 1), of PB *m* increases slightly from 6.74 to a value of 7.00, *i.e.* the helices are longer than in the previous work. This fact is observed in all the new databanks.

Transitions : The transitions between PBs are similar to the previous ones (cf. Table 1, col. 4 to 6) both in order and in percentage. Three PBs, *i.e.*, PBs *a*, *m* and *j*, differ from the previous study (see Table 1 of [6]). For PBs *a* and *m*, only a small inversion of proportion for the third and fourth main transitions is observed. For PB *a*, its third main transition is now PB *d* (9.4%, previously 7.9%) instead of PB *b* that has a transition rate of 9.3% (previously 8.0%). For PB *m*, PB *k* (now 11.3% and previously 9.3%) has switched with PB *b* (8.1% and 9.7% respectively). As regards, no clear preferential transitions were favoured for PB *j* in the previous study [6]. The same conclusion is found again with some inversions.

With only the three most frequent transitions per PBs, 89.3% of all the transitions between the PBs of the databank are taken into account. This value does not take account of the repetition of PBs upon themselves, if they are considered the final value increases to 94.3%. This fact implies a high dependency between successive PBs, as shown in our previous work based on the analysis of series of 5 PBs [9].

Thus our structural alphabet is highly conditioned by the presence of a limited number of transitions between the PBs, *i.e.*, the number of rarely observed transitions (transition rate < 0.5%, cf. Table 1, col.8) is high. Each PB has on average 7 transitions rarely observed and half them has a transition rate less than 0.1%. Moreover, 18 of them are never seen in the whole PDB. The rate of rarely observed PBs ranges between 2 (PB *j*) to 12 (PB *f*).

Secondary structures: The columns 9 to 14 sum up the secondary structure assignments done by PSEA [16] and STRIDE [17]. In our previous work [6], the assignment was done with a consensus method [18] based on other assignment methods. Comparisons between all those studies show the difficulties to define rules describing the secondary structures [13]. STRIDE and PSEA assignments differ sometimes by more than 20% (PB *l*, 40.4% vs. 61.0%). STRIDE is the closest to the previous consensus approach for most of the PBs (except for the PBs *k*, *l* and *m*). Globally, STRIDE assigns more α-helix state and PSEA more β-strand. They are roughly equivalent for the PBs with a high content in coil state. This result is in adequacy with (i) the observed differences between distinct secondary structure assignments [13, 18] and (ii) the distinction between secondary structure assignment and structural alphabet assignment [4].

Structural approximations: The PBs have been defined using *rmsda* measures (see Table 2, col. 2 to 5), *i.e.*, a Euclidean distance on dihedral angular values (ϕ and ψ angles). Compared to our previous study [6], the mean *rmsda* remains at 30° with a standard deviation of 20°. The median *rmsda* equals only 26°. For 11 of the PBs, the median *rmsda* value is slightly smaller than the mean value. For PB *m* and PB *n* the median *rmsda* values drop to 7.6° and 15.0°.

The computation of *rmsda* values is less classical than *rmsd* values. Hence to estimate its discriminative power, we have calculated the difference between the smallest *rmsda* value which gives the assignment, and the second smallest *rmsda* value. They correspond to the two minimal Euclidean distances on dihedral angular values. This difference is high (mean value of 29.5°, *i.e.*, the mean value is of 30° for the minimal *rmsda* and 59.5° for the second one, cf. Table 2, col.5). The most discriminative PBs are those associated with PB *m*, they show a low *rmsda* and the greatest difference with the second *rmsda*. The PBs associated with PB *d*

constitutes the second category, i.e., low *rmsda* values and fair difference with the second *rmsda*. The least determined ones are PB *g* and PB *j* with high mean *rmsda* values of 50.6° and 49.0° and low differences values of 14.9° and 19.6° respectively.

The *rmsd* (average Euclidean distance between superimposed C_α) computation gives a mean value of 0.41 Å with a standard deviation of 0.25 Å and a median value of 0.34 Å (see Table 2, col. 6 to 8). In the previous work [6], this *rmsd*-value was higher (0.58 Å) because it was computed as all the fragments associated with a PB against all. Here, the *rmsd* value is the mean value obtained by superimposing all the fragments associated with a PB and the average PB 3D prototype. Hence, it corresponds to the approximation of protein structures. The median *rmsd* values are lower for all the PBs. For instance, PB *m* median *rmsd* value equals 0.21Å, PB *d*, 0.36Å and PB *j*, 0.76Å. Only PB *g* has a *mean* rmsd value slightly lower than its *median* rmsd value.

The *mda120* value was used by Bystroff and Baker [19] and is adapted here for fragments five-residue long (see Table 2, col. 9). The *mda120* criterion consists in the comparison of the 8 angles of the true protein fragment with the 8 angles of the assigned PB, i.e. PB with the lowest *rmsda* value. If no angle deviates of more than 120°, the 5 residues are considered as correctly approximated. The *mda120* value is the proportion of residues considered -at least one time- as correctly approximated. As expected, most of the fragments are good according to the *mda120* criterion. Analysis of the fragments considered as "not correctly approximated" by the *mda120* criterion shows that they are always associated to high *rmsd* and high *rmsda* values. The best PBs, using this criterion, are the repetitive PBs *m* and *d*. As always, the poorest PB is PB *j* with a *mda120* rate of only 92.5%. In fact, this last corresponds to the most variable PB.

C_α distance : The distribution of C_α - C_α distances have been analyzed for every PBs

and between every position *i* (*i* ranging from *i*-2 to *i*+2). A statistical analysis of all these measures has been done using a Principal Component Analysis (PCA [20]). This last shows that the most significant parameter is the distance between the N and C ends (see Table 2, col. 10). The other distances are less important, their significance is directly related to the number of residues, *i.e.*, a distance between 4 residues is more informative than a distance between 3 residues and so on. These values categorize clearly a gradient between the full extended β-strand (PB *d* with a C_1-C_5 distance of 12.5 Å) to the C-cap of α-helix (PB *n*, distance of 6.5 Å). This last is slightly more compact than the helix, PB *m* (distance of 6.6 Å).

Conclusion. In this study, we verified that the PB definitions remain valid after the size of the databank more than tripled (from 86,628 [6] to 293,507 residues). We have highlighted different points. First, the distribution of the PB frequencies remains equivalent in all the non-redundant databank. Second, the transitions between all the PBs remain also highly constant. Third, the comparison with classical secondary structure assignments is not trivial as the assignment methods differ and the correspondence with PBs is not simply direct. For instance, the PB *d* has the geometrical feature of β-strand, but is assigned by PSEA [16] and by STRIDE [17] to coil state with a rate equal to 19.6% and to 29.0%, respectively. Finally, the analysis of the geometrical properties of PBs shows that they approximate well every part of the protein structures. Moreover, the use of *rmsda* distance allows a good discrimination. The comparisons of the *rmsd* and *rmsda* values of PBs show that *rmsda* is more sensitive to small structural variations than *rmsd*. The *rmsda* is more discriminative to split up the N- or C-caps of repetitive structures from the core of the repetitive structures. For instance, PBs *m* and *n* have a discriminative *rmsda* value for a low *rmsd* value. Furthermore, we analyze for the first time the PB with *mda120* criterion, finding consistent correspondence with the results found with *rmsda* criterion.

PBs may be used to analyze particular function-related structural motifs. For instance, the relation of PBs to the canonical structures in immunoglobulins could be very interesting [21]. As the prediction of the PBs from sequence has been recently improved [8], PBs could be of great interest in an *ab initio* modeling. Indeed, Pei and Grishin have recently emphasized the interest of prediction of blocks for predicting local protein structures [22]. In fact, the use of dihedral angles –basis of the PB definitions- has already shown its interest in exploring the folding process [23], especially outside the repetitive structures [24]. This kind of approaches lies on the fact that proteins are constructed from a small catalog of recognizable parts that fits together in a limited number of ways [25]. The major limitation of these approaches is the assembly of the different parts to build a complete and reliable protein structure.

Acknowledgments

I would like to thank Cristina Benros, Laurent Fourrier, Massimo Tortue, Patrick Fuchs, Delphine Flatters, Anne-Claude Camproux, Serge Hazout and Catherine Etchebest for fruitful discussions. This work was supported by grants from French Institute for Health and Medical Research (INSERM), from the Ministère de l'Enseignement Supérieur et de la Recherche and from "Action Bioinformatique inter EPST" 2003-2004.

References

[1] Bansal, M. and Kumar Velavan, R. (2000). HELANAL - A program to characterize helix geometry in proteins. J Biomol Struct. Dyn. **17**, 811-819.

[2] Chan, A.W., Hutchinson, E.G., Harris, D. and Thornton, J.M. (1993) Identification, classification, and analysis of β-bulges in proteins. Protein Sci. **2**, 1574-1590.

[3] Michalsky, E., Goede, A. and Preissner, R. (2003). Loops In Proteins (LIP)--a comprehensive loop database for homology modeling. Protein Eng. **16**, 979-985.

[4] de Brevern, A.G., Camproux, A.C., Hazout, S., Etchebest, C. and Tuffery, P. (2001). Beyond the secondary structures : the structural alphabets. In Recent Adv In Prot Eng., Sangadai SG ed. Research signpost, Trivandrum,India, pp. 319-331.

[5] Camproux, A.C., Gautier, R. and Tufféry, P. (2004). A Hidden Markov Model derived structural alphabet for proteins. J Mol Biol. **339**, 591-605.

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

[7] de Brevern, A.G., Benros, C., Gautier, R., Valadié, H., Hazout, S. and Etchebest, C. (2004). Local backbone structure prediction of proteins. In silico Biology. **4**, 31.

[8] Etchebest, C., Benros, C., Hazout, S. and de Brevern. A.G. (2005). A structural alphabet for local protein structures: improved prediction methods. Proteins. *accepted*.

[9] de Brevern, A.G., Valadié, H., Hazout, S. and Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. Protein Sci. **11**, 2871-2886.

[10] de Brevern, A.G. and Hazout, S. (2001). Compacting local protein folds by a "Hybrid Protein Model". Theor Chem Acc. **106(1/2)**, 36-47.

[11] de Brevern, A.G. and Hazout, S. (2003). Improvement of "Hybrid Protein Model" to define an optimal repertory of contiguous 3D protein structure fragments.

Bioinformatics. **19**, 345-353.

[12] Benros, C., de Brevern, A.G. and Hazout S. (2003). Hybrid Protein Model (HPM): A method for building a library of overlapping local structural prototypes. sensitivity study and improvements of the training. IEEE Int Work. NNSP 2003, **1**, 53-70.

[13] 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.

[14] 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**, 504-514.

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

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

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

[18] Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. and Mornon, J.-P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng. **6**, 377-382.

[19] Bystroff, C. and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motif. J Mol Biol. **281**, 565-577.

[20] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) Multivariate Analysis. Academic Press.

[21] Al-Lazikani B., Lesk A.M. and Chothia C. (1997). Standard conformations for the canonical structures of immunoglobulins J. Mol. Biol. **273**, 927-948.

[22] Pei, J. and Grishin, N.V. (2004). Combining evolutionary and structural information for local protein structure prediction. Proteins. **56**, 782-794.

[23] Jurkowski, W., Brylinski, M., Konieczny, L., Wiiniowski, Z. and Roterman I. (2004). Conformational subspace in simulation of early-stage protein folding. Proteins. **55**, 115-127.

[24] Kuang, R., Leslie, C.S. and Yang, A.S. (2004). Protein backbone angle prediction with machine learning approaches. Bioinformatics. **20**, 1612-1621.

[25]. Fitzkee, N.C., Fleming, P.J., Gong, H., Panasik, N., Street, T.O. and Rose G.D. (2005) Are proteins made from a limited parts list? Trends in Biochemical Sciences. **30**, 73-80.

Captions

Figure 1: *Protein Blocks*. From left to right and top to bottom the 16 Protein Blocks (labeled from *a* to *p*) are shown. For each PB, the N-cap is on the left and the C-cap is on the right.

Table 1: *Protein Blocks characteristics*. For each protein block (PB; labeled from *a* to *p*) is given: (i) the occurrence frequency (*freq*), (ii) the average number of repeats (*anr*), *i.e.*, the average number of times a PB repeats upon itself, (iii) the three main PB transition proportions to other PBs (*major transitions*), and the sum of these transitions (sum), only the transition frequencies more than 5% are noted, (iv) the transitions rarely observed, *i.e.* associated to a frequency < 0.5% (in bold their frequencies are < 0.1%) and, (v) the repartition in secondary structures of the central residue (α-helix, coil and β-strand) assigned by PSEA [16] and STRIDE [17] (in bold are highlighted the frequencies > 50%).

Table 2: *Protein Blocks characteristics*. For each protein block (PB; labeled from *a* to *p*) is given: (i) the root mean square deviation on angular value (*rmsda*) with its mean (*mean*), standard deviation (*s.d.*) and median (*median*) values, and the difference between the smallest *rmsda* and the second smallest *rmsda* (*dif.*), (ii) the root mean square deviation (*rmsd*) with its mean (*mean*), standard deviation (*s.d.*) and median (*median*) values, (iii) the mda120 value (mda_{120}) and (iv) the distance between the first Cα and the last Cα (dC_1-C_5). In bold is given the minimal value between mean and median values. It enables to highlight, in most of the cases, the overestimation of the mean values due to some extreme values.