Brief Communication in *Computational Biology and Chemistry* (2009)

Protein short loop prediction in terms of a structural alphabet.

Manoj Tyagi[1,2,3], Aurélie Bornot[2,4], Bernard Offmann[1,5,6]
& Alexandre G. de Brevern [2,4,*]

[1] Laboratoire de Biochimie et Génétique Moléculaire, Université de La Réunion,
BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09,
La Réunion, France.
[2] INSERM UMR-S 726, Equipe de Bioinformatique Génomique et Moléculaire (EBGM),
Université Paris Diderot - Paris 7, case 7113,
2, place Jussieu, 75251 Paris, France.
[3] Current Address: Computational Biology Branch, National Center for Biotechnology
Information (NCBI), National Library of Medicine (NLM),
8600 Rockville Pike, Bethesda, MD 20894.
[4] INSERM UMR-S 665, DSIMB, Institut National de Transfusion Sanguine (INTS),
6, rue Alexandre Cabanel, 75739 Paris cedex 15, France.
[5] INSERM UMR-S 665, DSIMB, Université de La Réunion, BP 7151, 15 avenue René
Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France.
[6] PEACCEL, Technopole, Saint-Denis de la Réunion, France.

Short title: protein short loops prediction

* Corresponding author:
Mailing address: Dr. de Brevern A.G., INSERM UMR-S 665, DSIMB, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail : alexandre.debrevern@univ-paris-diderot.fr
Tel: (33) 1 44 49 30 38
Fax: (33) 1 47 34 47 31

Abstract

Loops connect regular secondary structures. In many instances, they are known to play crucial biological roles. To bypass the limitation of secondary structure description, we previously defined a structural alphabet composed of 16 structural prototypes, called Protein Blocks (PBs). It leads to an accurate description of every region of 3D protein backbones and have been used in local structure prediction.

In the present study, we used our structural alphabet to predict the loops connecting two repetitive structures. Thus, we showed the interest to take into account the flanking regions, leading to prediction rate improvement up to 19.8%, but we also underline the sensitivity of such an approach. This research can be used to propose different structures for the loops and to probe and sample their flexibility. It is a useful tool for *ab initio* loop prediction and leads to insights to flexible docking approach.

Key-words: local protein structure, protein loops, secondary structure, protein function, bioinformatics, biophysics.

Introduction

The knowledge of the three-dimensional (3D) structures of proteins contributes to understand their biological functions. Protein 3D structures are often described as a succession of repetitive secondary structures (mainly α-helices and β-sheets (Pauling and Corey, 1951; Pauling et al., 1951)). This mono-dimensional description helps to simplify coarsely this 3D information.

Nevertheless, the secondary structures description focuses only on two kinds of regular local structures that compose only a part of protein backbone. The remaining residues are only assigned if they can be associated with some particular structures such as the β-turns (Bornot and de Brevern, 2006; Wang et al., 2006). It has led some scientific teams to develop local protein structure libraries that (i) are able to approximate all (or almost all) the local protein structures and (ii) do not make use of the classical secondary structures assignments. These libraries lead to the categorization of 3D structures without any *a priori* into small prototypes that are specific to local folds found in proteins. The complete set of *local structure prototypes* defines a *structural alphabet* (Karchin et al., 2003; Offmann et al., 2007). The number of libraries or structural alphabet is important, *e.g.* (Martin et al., 2008; Sander et al., 2006; Tung et al., 2007).

Our structural alphabet is composed of 16 mean protein fragments of 5 residues in length, called Protein Blocks (PBs, see Figure 1) (de Brevern et al., 2000). They have been used both to describe the 3D protein backbones (de Brevern, 2005) and to perform a local structure prediction (de Brevern et al., 2000; de Brevern et al., 2004; Dong et al., 2008; Etchebest et al., 2005; Zimmermann and Hansmann, 2008). PBs have proven their efficiency in the description and the prediction of long fragments (Benros et al., 2009; Benros et al., 2006; Bornot et al., 2009; de Brevern and Hazout, 2003; de Brevern et al., 2002; de Brevern

et al., 2007), to compare protein structures (Tyagi et al., 2008; Tyagi et al., 2006a; Tyagi et al., 2006b), to build globular (Dong et al., 2007) and transmembrane protein structures (de Brevern, 2009; de Brevern et al., 2009; de Brevern et al., 2005), to define a reduced amino acid alphabet dedicated to mutation design (Etchebest et al., 2007), to design peptides (Thomas et al., 2006), to define binding site signatures (Dudev and Lim, 2007) or to analyze protein contacts (Faure et al., 2009). The features of this alphabet have been compared with those of 8 other structural alphabets showing clearly that our PB alphabet is highly informative, with the best predictive ability of those tested (Karchin et al., 2003).

In this paper, we focus on the prediction of short loops from sequence. Loop prediction is one of the major drawbacks of homology modeling methods even if loops play crucial biological role. Loop prediction is frequently performed after the positioning of repetitive secondary structures. Current protocols are based on the sampling of the conformational space of the loop fragment and the scoring of the corresponding sampled conformations, but do not perform well for loops more than 10 residues (Zhu et al., 2006).

Prediction using a structural alphabet is an efficient way to predict loop structures. Indeed, as the PBs are overlapping and more precise than secondary structure, a small modification of PB series has a fewer consequence in comparison to a change observed in secondary structure assignment. Moreover, PBs are successful to align structural homologues even in difficult case with very divergent loops (Tyagi et al., 2008; Tyagi et al., 2006a). Following a previous research (Fourrier et al., 2004), we described short loops that connect two repetitive structures using our structural alphabet, and performed local loop structure predictions from the amino acid sequence. In this study we assessed the importance of *a priori* knowledge of the flanking regions and of sequence identity of the protein databank on the loop prediction.

Methods

Data sets. 10 non-redundant protein databanks have been used in this study. They are based on the PISCES database (Wang and Dunbrack, 2005) and represents between 162,830 (namely DB0) and 1,572,412 residues (namely DB8, see supplementary data I). They are available at http://www.dsimb.inserm.fr/~debrevern/DOWN/DB/new. The sets are defined as containing no more than *x*% pairwise sequence identity with *x* ranging from 20 to 90%. The selected chains have X-ray crystallographic resolutions less than 1.6 Å with an R-factor less than 0.25 or less than 2.5 Å with an R-factor less than 1.0. Each chain was carefully examined with geometric criteria to avoid bias from zones with missing density (Tyagi et al., 2009).

Protein Blocks & short loop description. Protein Blocks (PBs) correspond to a set of 16 local prototypes, labeled from *a* to *p* (cf. Figure 1), of 5 residues length based on Φ, Ψ dihedral angles description. They were obtained by an unsupervised classifier similar to Kohonen Maps (Kohonen, 1982) and Hidden Markov Models (Rabiner, 1989). The PBs *m* and *d* can be roughly described as prototypes for central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent β-strand N-caps and PBs *e* and *f*, C-caps; PBs *g* through *j* are specific to coils, PBs *k* and *l* to α-helix N-caps, and PBs *n* through *p* to C-caps. This structural alphabet allows a reasonable approximation of local protein 3D structures (de Brevern et al., 2000) with a root mean square deviation (*rmsd*) now evaluated at 0.42 Å (de Brevern, 2005). PBs (de Brevern, 2005) have been assigned using in-house software (available at http://www.dsimb.inserm.fr/DOWN/LECT/), it follows similar rules to assignment done by PBE web server (http://bioinformatics.univ-reunion.fr/PBE/) (Tyagi et al., 2006b). We have defined the short loops as PB series of length 2 to 6. They represent the most frequent loop lengths. These series must be composed of non-repetitive PBs, *i.e.* all the PBs except PBs *d* and *m*. They must have flanking regions composed of series of PBs *mm* and

/ or *dd* (Fourrier et al., 2004).

Prediction. In a strategy of structure prediction from sequence (de Brevern et al., 2000; de Brevern et al., 2004; Etchebest et al., 2005), we must compute for a given sequence window $S_{aa} = \{aa_{-w}, \ldots, aa_0, \ldots, aa_{+w}\}$, the probability of observing a given protein block PB_x, *i.e.* $P(PB_x \mid S_{aa})$. For this purpose, each PB is associated with an occurrence matrix of dimension l x 20 centered upon the PB, with $l = 2\ w + 1$ (in the study, $w = 7$). Using the Bayes theorem to compute this *a posteriori* probability $P(PB_x \mid S_{aa})$ from the *a priori* probability $P(S_{aa} \mid PB_x)$ deduced from the occurrence matrix allows us to define the odds score R_x :

$$R_x = \prod_{j=-w}^{j=+w} \frac{P\left(aa_j = i \middle| PB_x\right)}{P\left(aa_j = i \middle| DB\right)}$$

The highest score R_x corresponds to the most probable PB (de Brevern et al., 2000; de Brevern et al., 2002; de Brevern et al., 2007; de Brevern et al., 2004; Etchebest et al., 2005; Fourrier et al., 2004). The Q_{16} value computed is the total number of true predicted PBs over the total number of predicted PBs. We also computed a Q_{14} value, specific for loops, *i.e.* the PB *m* and *d* are not taken into account in the accuracy rate computation (Fourrier et al., 2004).

Thus different predictions have been done: (i) one general prediction done for comparison purpose, it gives the Q_{16} and Q_{14} values, (ii) a specific prediction of only short loops, it gives a Q_{14} short values and (iii) specific prediction of only short loops α-α, α-β, β-α and β-β, it gives a Q_{14} short values for each kinds of short loops.

Results & Discussion

Prediction of structural alphabets could be strongly dependant of the selection of the

training and validation set (Etchebest et al., 2005). This feature is amplified in the case when the letters of the structural alphabets have low frequency, *e.g.* π-helix (Wang et al., 2006), type VI β-turns (Fuchs and Alix, 2005) or PB *b* (de Brevern, 2005). Specific learning and validation were performed (i) for all the PBs (for comparison purpose), (ii) a specific research focusing on all the protein loops, and finally (iii) for each kinds of loops, *i.e.* α-α, α-β, β-α and β-β. Protein databanks have been split randomly into 2/3 of the proteins used for the training step and 1/3 for the validation step. We have done large number of independent simulations for optimizing the prediction rates, *i.e.* for each simulation the proteins used for the training and the validation sets are different.

Global prediction rates. Figure 1 shows 1000 independent simulations performed with the databank DB0 (887 protein chains, see supplementary material I). This Figure shows the importance of sampling. The difference between average and best prediction values is 1% for the prediction rate Q_{16} and increases with the diminution of the databank, *e.g.* this difference equals 8% for the prediction of short loops connecting α-helix to β-strand. The best prediction rates for each of the PBs have been summarized in Table 1. These results highlight the interest to perform specific prediction for each kind of short loops; a specific learning with the *a priori* knowledge of the flanking regions improves the prediction rate. Nonetheless, this improvement is not uniform for all the short loop kinds.

From a correct Q_{16} value of 35.8%, a prediction specific to loops, *i.e.* Q_{14}, reaches 37.5% (+1.7%), and to 43.2% (+5.7%), if only the short loops are used. An increase ranging from 0.7% to 19.0% was observed for the 4 types of short loops.

We thus confirm our previous results, but also underline the importance of shuffling the proteins between validation and learning set (as we have also shown in (Etchebest et al., 2005)). It allows to improve all the prediction rates in comparison to (Fourrier et al., 2004),

thus Q_{16} goes from 35.2 to 35.8% and the corresponding Q_{14} from 36.0 to 37.4% highlighting the quality of prediction, *i.e.* this improvement is not due only to PBs *m* and *d*, the repetitive PBs. In the same way, Q_{14} for the short loops also increases from Q_{14} = 42.3 to 43.2%. For the 4 types of short loops, the difference between average and best prediction rates bypasses 5% in all the cases.

This global improvement is also associated with an increase of the over – training, *i.e.* a prediction largely better with the learning set than with the validation set. For the prediction of all residues, the over – training rate is less than 1% on average; it increases to 4% for the short loops and increases again for the short loop types, this value can reach 20% and more. Nevertheless, a strong linear correlation exists ($r > 0.95$) between the prediction rate and the over-training. The better is the prediction rate, the lower is the over – training. Thus interestingly, the selected predictions are associated with the lower over – training rates.

These values represent only a compatibility between (2D') representation, *i.e.* true PBs and predicted PBs. So, we also assessed the quality of local structure approximation using C_α root mean square deviation (*rmsd*) measure. We superimposed all the true PBs to the local protein structure, even the good PB, in order to compute the expected *rmsd* random values and we found they are always between 1.95 and 2.00 Å. The percentage of predicted PBs better than random value corresponds to 64% and 66% for the loops and short loops prediction respectively. These values increase with the specialisation of short loops with prediction rates equal to 68% for β-α, 74% for β-β, 76% for α-α and 82% for α-β.

PBs prediction rates. Improvements of prediction rates for the short loops are due to different PBs. For the *all* short loops, only two PBs clearly lost prediction rate, *i.e.* PBs *j* and *l*. The most frequent PBs have often very good prediction rate. For the α-α short loops, only two PBs present a decreased prediction rate (PBs *a* and *h*), for α-β, 5 (PBs *f*, *h*, *i*, *k* and *l*), for

β-α 5 (PBs *a*, *c*, *h*, *i* and *j*) and 3 for β-β (PBs *h*, *n* and *o*). As most of these PBs have very low frequencies (mainly less than 5% of the PBs in their kind of loops), these losses of their prediction rate is not significant. We can note that PBs *h* never obtains a prediction gain with this approach, at the opposite, the prediction of PBs *b* and *p*, which are related to some β-strand N capping regions is improved in each case.

Influence of protein databanks. All the prediction rates presented bellow belongs to the smallest databank (DB0). Interestingly these prediction rates do not increase with the growth of the protein databank, *i.e*. databanks with more redundancy. The best prediction rates remain always near identical, thus the similarity sequence rate does not improve this prediction approach based on a single sequence. Nonetheless, the over-training rate decreases greatly with the size of the databank, *i.e*. for (80% of sequence identity, resolution better than 2.5 Å and R-factor less than 1.00), the biggest one, the over-training rate is less than 2% thanks to the redundancy of the protein databank.

In the same way, it must be noticed that it does not change the kind of amino occurrence matrices, *i.e*. no new amino acid specificities are observed or disappeared (data not shown). This fact underlines that Bayesian prediction is an interesting tool to predict and analyze the local protein structures, we expect to use in future more sophisticated methods less dependant on the size of databank. One advantage of such an approach is that it enables us to compute the most significant series of PBs and from this information propose alternative 3D candidate structures as we have shown previously (Fourrier et al., 2004).

Conclusion.

Prediction of short loops taking into account of the flanking regions clearly improves

the quality of prediction, with only limited effect on β to α short loops, but near 20% of gain for α to α short loops. The use of numerous non-redundant databanks show: (i) global prediction rates remain equivalent for all the protein databank used, with no influence of the sequence identity or quality of crystallographic resolution and (ii) the importance of over – training, this last diminished greatly with the increase of the protein databank size.

Though sometimes the prediction rates of each PB could vary a little, most of the time it is less than 5%. We can notice that four PBs are associated with more variation of their prediction rates, interestingly their frequencies in the most redundant databanks are decreased in regards to the lower redundant databanks. Thus, a slight statistical bias is observed.

In this field, prediction of loops using a structural alphabet is an interesting tool. Instead of describing entire loops; our approach predicts each position in the loops locally. To take into account the flanking regions of the loops greatly improve the prediction rates. As nearly all protein fragments can be can be observed in the Protein DataBank (Balamurugan et al., 2005; Balamurugan et al., 2006; Du et al., 2003), this approach can be used to propose many different structures for the loops and to probe and sample their flexibility, and, is a useful tool in *ab initio* loop prediction.

Acknowledgements

This work was supported by grants from the Ministère de la Recherche, Université Paris Diderot – Paris 7, Université de Saint-Denis de la Réunion and the French Institute for Health and Medical Care (INSERM). MT had a grant from the Conseil Régional de Réunion and AB has a grant from Ministère de la Recherche.

Figures

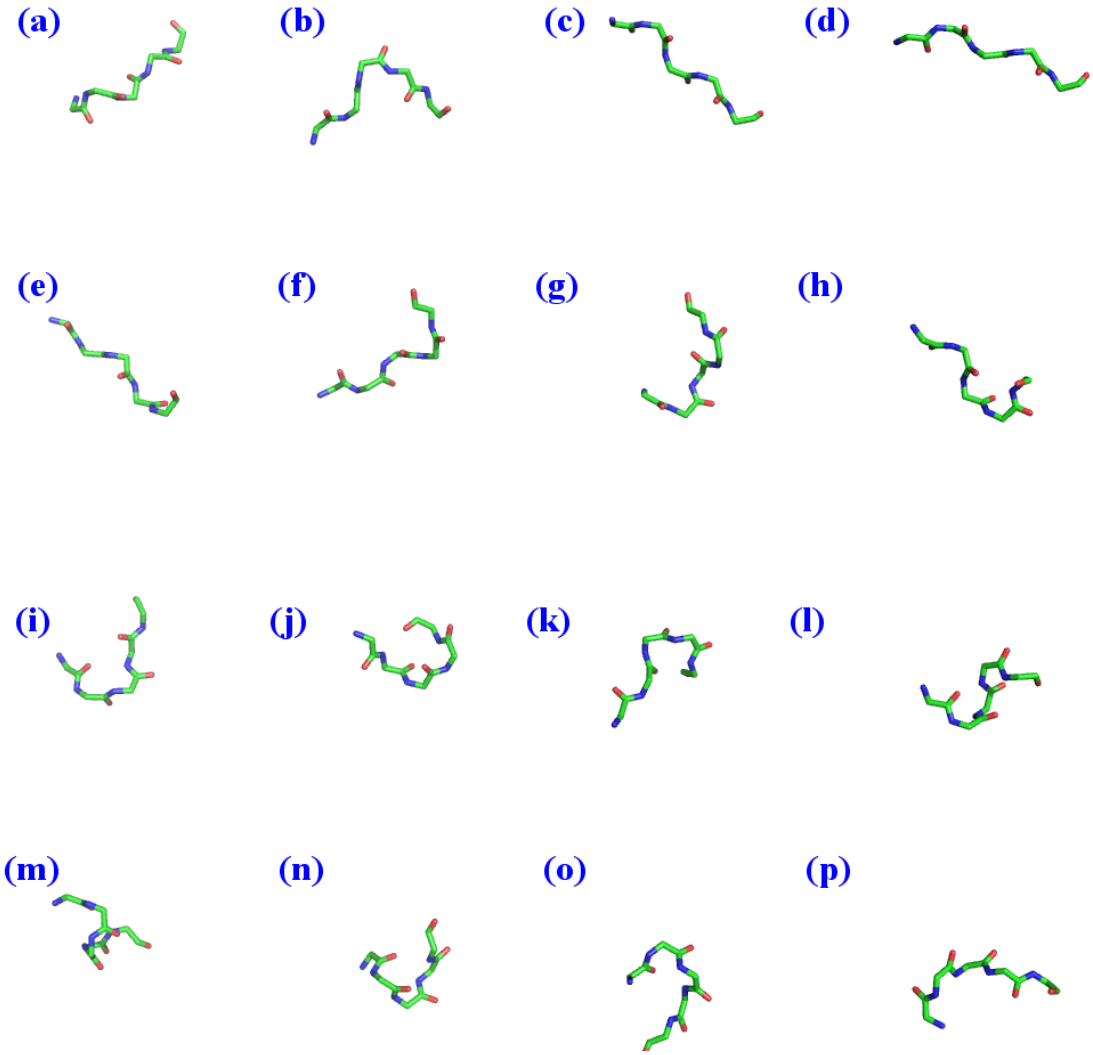

Figure 1 - The 16 Protein Blocks. From left to right and top to bottom, PyMol (DeLano, 2002) images of the 16 Protein Blocks of the structural alphabet (de Brevern, 2005; de Brevern et al., 2000). Each prototype is five residues in length and corresponds to eight dihedral angles (φ,ψ). The PBs *m* and *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.

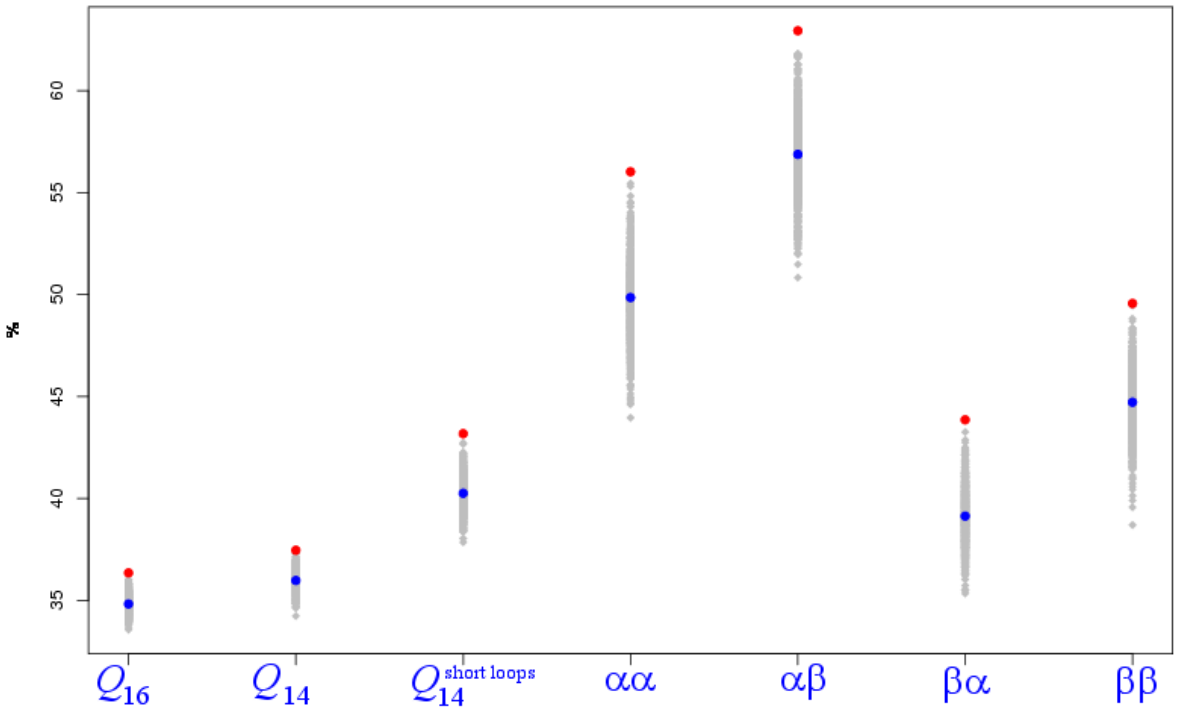

Figure 2 - Distribution of prediction rates. Are given the prediction rate for the classical prediction (Q_{16}), a prediction dedicated to the loops (Q_{14}), to the short loops ($Q_{14}^{\text{short loops}}$), to the short loops connecting α-helix to α-helix ($\alpha\alpha$), α-helix to β-strand ($\alpha\beta$), β-strand to α-helix ($\beta\alpha$) and β-strand to α-helix ($\beta\beta$). The grey dots correspond to the 1000 independent simulation prediction rates, the blue dots to the average prediction value of the prediction rates and the red ones to the best prediction rates.

	complete db.		Loops		Short loops		αα		αβ		βα		ββ	
PBs	pred.	freq.	pred.	freq.	pred.	freq.	pred.	freq.	pred.	freq.	pred.	freq.	pred.	freq.
a	60.28	*3.87*	62.22	*8.48*	68.11	*9.07*	63.10	*7.50*	80.51	*16.29*	27.78	*1.39*	72.00	*12.33*
b	14.03	*4.02*	11.06	*6.48*	17.92	*4.78*	21.74	*2.05*	23.08	*3.26*	33.66	*7.78*	20.41	*5.37*
c	27.75	*7.64*	26.13	*10.12*	32.79	*7.31*	35.82	*11.96*	46.51	*10.78*	28.12	*2.46*	40.71	*6.19*
d	28.50	*16.37*	24.69	*4.78*										
e	39.98	*2.18*	37.39	*4.90*	47.27	*4.71*					55.71	*5.39*	50.00	*10.74*
f	29.65	*6.44*	36.40	*13.60*	40.94	*15.60*	70.44	*14.20*	0.00	*1.00*	48.16	*33.41*	54.72	*13.92*
g	25.15	*1.11*	28.28	*2.39*	27.89	*2.52*	35.71	*3.75*	41.67	*4.01*				
h	44.39	*2.08*	43.82	*5.17*	50.18	*4.81*	33.33	*1.34*	33.33	*2.26*	46.88	*4.93*	48.55	*9.48*
i	36.22	*1.64*	37.41	*3.99*	45.88	*3.32*			37.04	*2.26*	22.22	*2.08*	49.19	*6.79*
j	56.98	*0.77*	44.98	*1.50*	29.17	*1.23*					16.67	*2.31*	37.04	*1.48*
k	38.66	*5.33*	42.05	*12.55*	37.15	*15.85*	64.81	*19.29*	20.00	*2.09*	54.75	*30.79*	52.91	*11.29*
l	38.85	*5.40*	24.48	*4.63*	16.32	*3.25*	22.22	*0.80*	12.50	*1.34*	19.64	*4.31*	18.18	*3.62*
m	38.19	*34.36*	17.26	*2.02*										
n	51.23	*2.28*	54.61	*4.90*	59.35	*6.87*	65.29	*10.80*	76.92	*17.38*			26.83	*2.25*
o	49.05	*2.95*	52.58	*6.66*	55.07	*8.95*	63.64	*11.79*	74.16	*17.46*			53.66	*6.74*
p	33.39	*3.56*	35.76	*7.85*	43.17	*10.27*	57.04	*12.68*	64.84	*21.39*			53.50	*8.60*
Q_{16} / Q_{14}	35.79		37.45		43.17		56.02		62.94		43.85		49.56	

Table 1 - The different PBs' predictions. Are given the prediction rates by PBs (*pred*) with the corresponding frequencies (*freq*) for a classical prediction (*complete db*), only for the loops (*Loops*), for the short loops (*Short Loops*) and dedicated for every kind of short loops (α-helix to α-helix, α-helix to β-strand, β-strand to α-helix and β-strand to β-strand). The last line corresponds to the global prediction rate Q_{16} for prediction of all residues and Q_{14} only for short loops. Results are based on DB0.

References

Balamurugan, B., Samaya Mohan, K., Ramesh, J., Roshan, M.N., Sumathi, K., Sekar, K., 2005. SSEP-2.0: Secondary Structural Elements of Proteins. Acta Crystallogr D Biol Crystallogr 61, 634-636.

Balamurugan, B., Roshan, M.N., Michael, D., Ambaree, M., Divya, S., Keerthana, H., Seemanthini, M., Sekar, K., 2006. SMS: sequence, motif and structure--a database on the structural rigidity of peptide fragments in non-redundant proteins. In Silico Biol 6, 229-235.

Benros, C., de Brevern, A.G., Hazout, S., 2009. Analyzing the sequence-structure relationship of a library of local structural prototypes. J Theor Biol 256, 215-226.

Benros, C., de Brevern, A.G., Etchebest, C., Hazout, S., 2006. Assessing a novel approach for predicting local 3D protein structures from sequence. Proteins 62, 865-880.

Bornot, A., de Brevern, A.G., 2006. Protein beta-turn assignments. Bioinformation 1, 153-155.

Bornot, A., Etchebest, C., de Brevern, A.G., 2009. A new prediction strategy for long local protein structures using an original description. Proteins.

de Brevern, A.G., 2005. New assessment of a structural alphabet. In Silico Biol 5, 283-289.

de Brevern, A.G., 2009. New opportunities to fight against infectious diseases and to identify pertinent drug targets with novel methodologies. Infect Disord Drug Targets 9, 246-247.

de Brevern, A.G., Hazout, S., 2003. 'Hybrid protein model' for optimally defining 3D protein structure fragments. Bioinformatics 19, 345-353.

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

de Brevern, A.G., Valadie, H., Hazout, S., 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.

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

de Brevern, A.G., Autin, L., Colin, Y., Bertrand, O., Etchebest, C., 2009. In silico studies on DARC. Infect Disord Drug Targets 9, 289-303.

de Brevern, A.G., Benros, C., Gautier, R., Valadie, H., Hazout, S., Etchebest, C., 2004. Local backbone structure prediction of proteins. In Silico Biol 4, 381-386.

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

DeLano, W.L.T., 2002. The PyMOL Molecular Graphics System DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.

Dong, Q., Wang, X., Lin, L., 2008. Prediction of protein local structures and folding fragments based on building-block library. Proteins 72, 353-366.

Dong, Q.W., Wang, X.L., Lin, L., 2007. Methods for optimizing the structure alphabet sequences of proteins. Comput Biol Med 37, 1610-1616.

Du, P., Andrec, M., Levy, R.M., 2003. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. Protein Eng 16, 407-414.

Dudev, M., 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., 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., de Brevern, A.G., 2007. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 36, 1059-1069.

Faure, G., Bornot, A., de Brevern, A.G., 2009. Analysis of protein contacts into Protein Units. Biochimie 91, 876-887.

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

Fuchs, P.F., Alix, A.J., 2005. High accuracy prediction of beta-turns and their types using propensities and multiple alignments. Proteins 59, 828-839.

Karchin, R., Cline, M., Mandel-Gutfreund, Y., Karplus, K., 2003. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51, 504-514.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern 43, 59-69.

Martin, J., de Brevern, A.G., Camproux, A.C., 2008. In silico local structure approach: a case study on outer membrane proteins. Proteins 71, 92-109.

Offmann, B., Tyagi, M., de Brevern, A.G., 2007. Local Protein Structures. Current Bioinformatics 3, 165-202.

Pauling, L., Corey, R.B., 1951. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci U S A 37, 251-256.

Pauling, L., Corey, R.B., Branson, H.R., 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 37, 205-211.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected application in speech recognition. Proceedings of the IEEE 77, 257-286.

Sander, O., Sommer, I., Lengauer, T., 2006. Local protein structure prediction using discriminative models. BMC Bioinformatics 7, 14.

Thomas, A., Deshayes, S., Decaffmeyer, M., Van Eyck, M.H., Charloteaux, B., Brasseur, R., 2006. Prediction of peptide structure: how far are we? Proteins 65, 889-897.

Tung, C.H., Huang, J.W., Yang, J.M., 2007. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biol 8, R31.

Tyagi, M., de Brevern, A.G., Srinivasan, N., Offmann, B., 2008. Protein structure mining using a structural alphabet. Proteins 71, 920-937.

Tyagi, M., Bornot, A., Offmann, B., de Brevern, A.G., 2009. Analysis of loop boundaries using different local structure assignment methods. Protein Sci, in press.

Tyagi, M., Gowri, V.S., Srinivasan, N., de Brevern, A.G., Offmann, B., 2006a. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 65, 32-39.

Tyagi, M., Sharma, P., Swamy, C.S., Cadet, F., Srinivasan, N., de Brevern, A.G., Offmann, B., 2006b. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 34, W119-123.

Wang, G., Dunbrack, R.L., Jr., 2005. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 33, W94-98.

Wang, Y., Xue, Z.D., Shi, X.H., Xu, J., 2006. Prediction of pi-turns in proteins using PSI-BLAST profiles and secondary structure information. Biochem Biophys Res Commun 347, 574-580.

Zhu, K., Pincus, D.L., Zhao, S., Friesner, R.A., 2006. Long loop prediction using the protein local optimization program. Proteins 65, 438-452.

Zimmermann, O., Hansmann, U.H., 2008. LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. J Chem Inf Model 48, 1903-1908.